519.72:534.78 K97s
NORGES TEKNISKE HØGSKOLE

SEGMENTATION AND LABELLING OF SPEECH

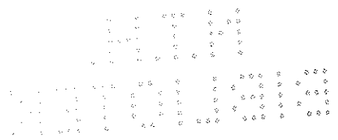
A DISSERTATION SUBMITTED TO
THE DEPARTMENT OF TELECOMMUNICATIONS OF
THE NORWEGIAN INSTITUTE OF TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DOCTORAL DEGREE — "DOKTOR INGENIØR"

by

Knut Kvale

1993

NTH-TRYKK
1993



Abstract

During the last decades, significant research efforts have been aimed at developing speech technology products such as speech input and output systems. In order to train and evaluate these systems huge speech databases have been compiled in laboratories all over the world. However, neither the recording protocols nor the annotation conventions used have been standardised, making assessment of speech technology products across laboratories and languages difficult. The aim of this thesis work is to contribute towards a standardisation of segmentation and labelling of multi-lingual speech corpora.

Segmentation is here defined as the process of dividing the speech pressure waveform into directly succeeding discrete parts. These segments are labelled with phoneme symbols. Continuous speech from five different languages; English, Danish, Swedish, Italian, and Norwegian, have been studied with respect to segmentation and labelling.

Due to coarticulation effects, exact segmentation of speech as defined above is theoretically impossible, but the segmentation and labelling provides a link between the speech waveform and the phonological labels which is nevertheless essential for both speech research and for the development of speech technology. Thus, this thesis takes a pragmatic approach to the segmentation and labelling of speech and suggests methods to make the annotation processes accurate and reliable enough for practical use.

There exists no single "correct" segmentation strategy. However, based on a comparison of the manual segmentation of the languages mentioned, a set of multi-lingual segmentation conventions can be standardised. This thesis suggests a set of segmentation conventions for Norwegian, and many of these conventions may as well be applied for other languages. Using the same conventions, two manual segmentations of the same Norwegian speech recording resulted in 96.5% coincidence of the segment boundaries within a ±20ms deviation threshold.

Manual segmentation is time-consuming and prone to random human errors and inconsistencies. We therefore automised the segmentation process. Our automatic segmentation algorithm consists of two independent parts; the acoustic segmentation, entirely based on signal processing, and the phonemic segmentation, based on hidden Markov modelling of the phonemes within each language. The inputs to the automatic segmentation algorithm were the speech waveform, the corresponding label string, and the endpoints of the utterance. The goal for the development of the automatic segmentation algorithm was to produce segment boundaries with accuracy comparable to the performance of the human labellers for each language.

The acoustic segmentation algorithm is shown to be robust against slightly different recording conditions. When the algorithm was forced to calculate twice as many acoustic segments as the number of phonemes, the acoustic segment boundaries coincided well with the corresponding manual segmentation and most of the acoustic segments could be given a phonetic interpretation.

Preface

This work is intended to satisfy the thesis requirements for the *Doktor Ingeniør* degree at the Norwegian Institute of Technology.

The topics addressed are **manual and automatic annotation of multi-lingual speech databases.** Since manual annotation of speech is both laborious and tedious the ultimate goal is to develop an algorithm which automatically annotates multi-lingual speech corpora. However, in order to train and evaluate the automatic procedure, manually annotated speech corpora were needed as reference sets.

In *chapter 1* the terms segmentation and labelling of speech are defined and the importance of segmented and labelled speech corpora is argued for. Fundamental and more pragmatical problems regarding speech annotation are discussed.

In *chapter 2*, entitled **speech production and perception,** relevant phonetics for this thesis work is summarised.

Chapter 3, entitled **speech annotation units**, discusses the terms phonology and phonetics. Different levels of annotation are described and several subword units are compared with respect to a set of requirements.

Chapter 4 addresses the problems of **manual segmentation and labelling of speech.** Annotation conventions used for similar speech material in different languages are compared. The annotation conventions developed in this thesis work for Norwegian are described in detail.

Chapter 5 reviews mathematical **modelling of speech** relevant for this thesis and sets the framework for the automatic segmentation.

In *chapter 6*, **automatic segmentation of speech,** several methods for automatic segmentation of speech are discussed and a taxonomy of the methods is given. Then our two-step algorithm for automatic multi-lingual speech segmentation is described in more detail. Finally, the problem of how to assess the automatic segmentation performance is discussed.

In *chapter 7* several simulation **experiments** with our automatic segmentation algorithm are evaluated both qualitatively and quantitatively. Some of the experimental setups used for automatic segmentation are also shown used on a simple phoneme recogniser.

Chapter 8, the **conclusion,** summarizes the main results and contributions of this thesis and presents some possible directions for further work.

Since people with different educational, cultural and language backgrounds work in the area of speech research and technology, a **Glossary** which interprets most of the terms and abbreviations used in this thesis is provided.

Four appendices are included. *Appendix A* lists the equivalents of IPA and SAMPA symbols and the Norwegian, English, Swedish, Danish, and Italian phoneme inventories. *Appendix B* summarizes the analyses details of the Norwegian recording described in chapter 4. *Appendix C* discusses examples of Swedish, Danish, and English manual annotation and provides analyses of the speech material employed in this thesis. Finally, *Appendix D* summarizes detailed results of the experiments described in chapter 7.

Since people with different educational, cultural and language backgrounds work in the area of speech research and technology, a **Glossary** which interprets most of the terms and abbreviations used in this thesis is provided.

Four appendices are included. *Appendix A* lists the equivalents of IPA and SAMPA symbols and the Norwegian, English, Swedish, Danish, and Italian phoneme inventories. *Appendix B* summarizes the analyses details of the Norwegian recording described in chapter 4. *Appendix C* discusses examples of Swedish, Danish, and English manual annotation and provides analyses of the speech material employed in this thesis. Finally, *Appendix D* summarizes detailed results of the experiments described in chapter 7.

Contents

Chapter 1

INTRODUCTION

During the last decades, vigorous research efforts have been aimed at developing speech technology products such as speech input and output systems. In order to train and evaluate these systems many speech databases have been recorded and annotated. However, neither the recording protocols nor the annotation conventions used have been standardised, making assessment of speech technology products across laboratories and languages difficult.
This thesis work is intended to contribute towards a standardisation of segmentation and labelling of multi-lingual speech corpora.

This introductory chapter addresses first the terms segmentation, labelling and annotation of speech as applied in this thesis. Then the rationales for carrying out speech annotation are elaborated in more detail. Problems connected with segmentation and labelling of speech are discussed and finally automatic segmentation of speech is defined and argued for.

1.1 WHAT IS SEGMENTATION AND LABELLING?

In this thesis the word *segmentation* is interpreted as the process of dividing something continuous into discrete, non-overlapping entities; i.e. the process of deciding boundaries. By *labelling* is meant a classification of a given segment as defined above, and *annotation* is used as a cover term for segmentation and labelling.

Segmentation and labelling of speech provides a symbolic interface, a bridge, between measurable acoustic parameters and abstract phonological categories. This interface is essential for speech research and technology, as elaborated in section 1.2. Figure 1.1 displays a short section, 1.15 sec., of a speech waveform which is manually segmented and labelled with phonemic SAMPA [Wells'92] symbols according to the definitions given in chapter 4.

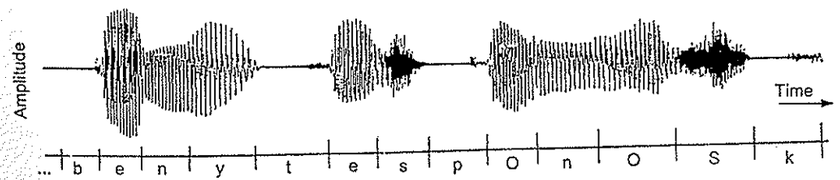


Figure 1.1. *The speech pressure waveform of the Norwegian utterance "benyttes på norsk" ('is used in Norwegian') read aloud by a male speaker. It is manually segmented and labelled into the phoneme string / b e n y t e s p O n O S k /.*

In practical life we often have to *segment* something big and unwieldy into smaller and more manageable parts, e.g. dividing up a carcass into appropriate dinner sizes. When doing the segmentation, we may use our experience and intuition, a cordon bleu cookery book, or we may have some vague and general rules as guidelines.

Labelling can be compared to the process of grouping objects and events possessing many of the same properties into one group and giving them a label or name, such as a ball, a car, a cow etc. A car is a subgroup of the class vehicle, but is a cover term for various types of cars. The level of detail depends on the context and the categorizing and labelling often differ among different languages/cultures.

Another example of *segmentation* is how we have divided up the world. The world has been segmented into countries, and each of these are often divided into even smaller communities to facilitate management and control. The borders are sometimes natural ones, e.g. seas, rivers and mountains, but most often located as a result of political decisions or wars. The borders are always where the strongest interest group wants them to be.

Segmentation of the acoustic speech signal can in many respects be compared with the "segmentation" of the world. In speech annotation phoneticians, linguists and "speech engineers" have taken the place of the generals and politicians. Since the speech signal is characterized by an intermixing of adjacent features to the extent that they merge with those of adjacent sounds, we have the same problems of selecting appropriate entities and agreeing on the criteria or conventions to use.

In this thesis work *segmentation of speech* is defined as dividing the speech signal into directly succeeding discrete parts; so-called traditional or *linear segmentation*, as exemplified in figure 1.1. The segmentation is based on the speech pressure waveform and/or quantities directly derived from this, such as zero-crossing rate, energy, and broad band spectrogram. To provide better segmentation cues, the parameter set can be further enhanced by e.g. the fractal dimension feature or "fractogram" which may provide extra information for segmenting voiced fricatives [Maragos'92], or perception based representations, e.g. [Seneff'84], which may amplify and sharpen the differences between segments.

Obviously the above description is not the only possible way of representing and annotating speech. For example, examination of the speech production process shows that the articulation and the air-flow/ voicing source works nearly independently of each other [Hunt'90]. Hence, it is natural to describe speech using at least both segmental and suprasegmental units, as discussed in chapter 3. Indeed, speech should be regarded as a multidimensional sequence, where many activities have contributed, and should therefore be annotated with overlapping segments, so-called *nonlinear segmentation*, using a hierarchy of segmentation levels.

Such a nonlinear speech annotation approach is applied in the ESPRIT ACCOR project [Marchal'92], where different sensors are used in the speech recording to provide laryngograph signals, oral and nasal air flows, and lingo-palatal contact patterns (EPG patterns), in addition to the sound pressure wave. These signals are segmented independently by marking off discontinuities.

People speaking the same dialect will almost always easily agree upon the number of syllables and which speech sound classes, phonemes, are represented in a given utterance in that dialect,

but they may have problems with agreeing on marking the boundaries between the phonemes and syllables. In speech perception, there is no need for segmentation because the discreteness as such has no perceptual function. What the listener needs to do is to identify segments and determine their temporal order, but neither of these activities presupposes a discrete partitioning of the speech input [Diehl'87]. Therefore, speech can be annotated with overlapping functions (e.g. as in temporal decomposition [Atal'83]) as exemplified at the top of figure 1.2, with transition segments shown at the bottom line of figure 1.2, or by indicating the centres of the selected units, e.g. [Erp'88].
However, the various alternatives for segmentation of speech will not be explored further in this thesis and the linear segmentation approach will be followed.

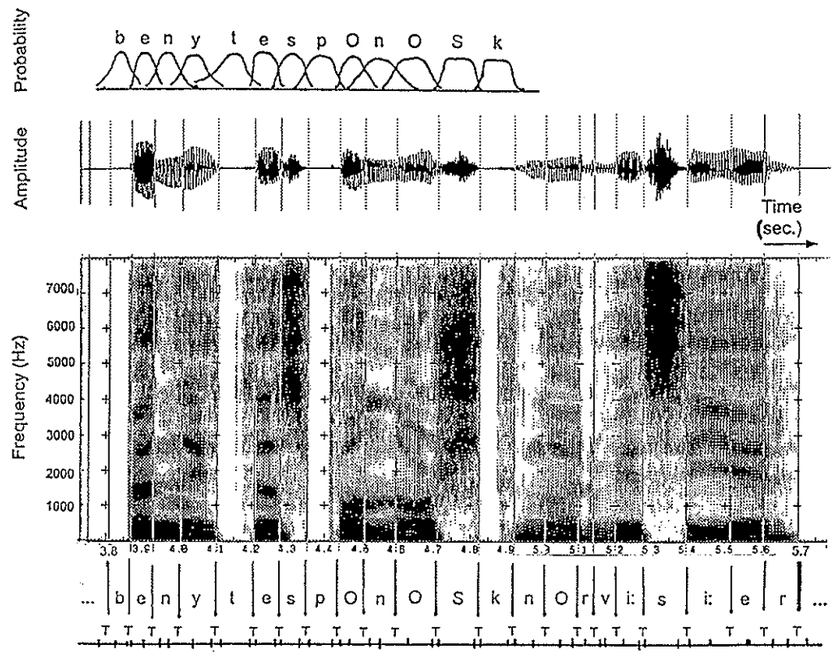


Figure 1.2. *The speech pressure waveform and the broad band spectrogram of the Norwegian sentence "benyttes på norsk når vi sier" ('is used in Norwegian when we say'), i.e. the utterance in figure 1.1 expanded. The + signs in the spectrogram are separated by 0.1 sec. horizontally and 1000 Hz vertically.*
-The top line indicates that perception cues for a phoneme can be found both in the preceding and the following phoneme(s). The vertical axis indicates the probability of evidence for each phoneme.
-The bottom line demonstrates how the transition segment, T, between two phonemes can be marked explicitly.

1.2 WHY SEGMENTATION AND LABELLING OF SPEECH?

In this section several applications of segmented and labelled speech are outlined. It will be shown that the relationship between phonological units, e.g. phonemes, and their acoustic realisations serves as a basis for speech processing techniques such as automatic speech recognition (ASR), text-to-speech (TTS), and speech coding, as well as for insight into human speech production and perception. However, we have to keep in mind that the tasks mentioned require a clear definition of the segmentation and labelling methods that are resorted to. This will be further discussed in chapter 4 and 6.

1.2.1 SPEECH RESEARCH

Speech is the most important and natural means of human communication. The speech signal contains the speaker's intended message. Satellite images of the earth or seismic signals also contain information, but not a message. We are all experts in talking and listening. We communicate with very little effort and the processes involved are managed nearly automatically, even though every utterance and speech sound is unique and may vary greatly between speakers and for the same speaker on different occasions. However, we have not acquired much explicit knowledge about these operations yet. One fundamental question in speech research is thus how we manage to transfer our ideas via speech, or via the so-called "speech chain"; which is the chain of events from the conception of a message in the speaker's brain to the arrival of the message in the listener's brain. This question may be broken down into three discrete parts:

-*How is speech produced?* That is, how are the ideas or feelings transformed into a language code? How are neural instructions to the muscles in the articulatory organs and lungs planned and triggered and how do the muscles actually work?

-*How is the information/message coded in the speech signal?* What are the acoustic cues for the speech sounds and the codes or conventions used in speech communication?

-*How is speech perceived?* That is, we want to find the signal transforms when the speech pressure waves reach the ears and proceed into the brain and finally impose the impression of meaning.

These questions are by no means fully answered. Therefore, more specialized studies are carried out in all fields concerning or related to the speech communication process, such as psychology, psycholinguistics, phonetics, phonology, dialectology, sociolinguistics, speech therapy, signal processing, mathematical modelling and computer science.

One major source of information is a huge, multilingual, carefully designed and annotated speech database containing samples of different speakers and speaking styles. Huge databases give more reliable statistical data and generalisations can be drawn and general acoustic-phonetic speech knowledge is more easily obtained. Below are listed some research areas in which segmented and labelled speech databases can be useful:

-Speech and language theories can be validated and compared in quantitative terms.
-Multilingual research such as e.g. comparison of languages at an acoustic-phonetic level.

-Development of speech material complexity measures.
-Determining temporal aspects of speech sounds, e.g. phoneme duration in different contexts.
-Data concerning phrase structure and stress realization.
-Quantifying how different contexts influence the phoneme realisation.
-Evaluating the acoustic bases of the phonetic transcription.
-Development of a continuous speech phonology.
-Discovering the transfer of information via sound by examination of the acoustic form of speech and the acoustic contents of phonological (linguistic) units, i.e. which acoustic features seem related to various speech sounds.
-Discovering the perceptual units of speech.

1.2.2 SPEECH TECHNOLOGY

In speech technology sections of the above mentioned speech chain are attempted substituted by a computer. In computer generated speech, or text-to-speech (TTS), the speaker is replaced by a computer which is appropriately programmed to convert a text string to speech. Speech coders make it possible to store and transmit speech more efficiently. In automatic speech recognition (ASR), the computer takes the listener's role. An ASR system may either perform what it is asked to do or it tries to convert the speech signals to a written text equivalent.

The issues of speech research, speech recognition, speech synthesis and speech coding are often treated separately. But the fields are closely related by the same underlying processes of speech and hearing. A unified approach based on these sources of knowledge and technologies together has therefore been proposed to improve understanding and products. For instance by regarding ASR and TTS as inverse processes, analysis-synthesis systems are constructed where the analysis part is an ASR system and the synthesis part is a TTS system, as exemplified in figure 1.3. Since only an index for each of the pre-defined phonological units is transmitted, speech coding at very low bit rates can be achieved [Flanagan'72],[Soong'89],[Flanagan'91].

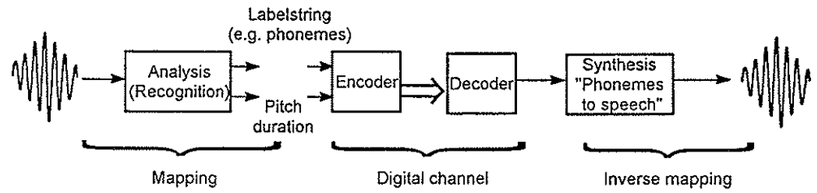


Figure 1.3 *Speech coding by combining ASR and TTS.*

Many procedures are proposed for ASR. In this thesis the term ASR is used for statistically based systems for transforming the speech signal into the corresponding sequence of linguistically defined units. These systems are characterised by three major modelling steps: *unit modelling*, i.e. the use of statistical modelling techniques to characterise the acoustic properties of the chosen unit, *word model composition*, i.e. the construction of word models from the unit

models, and *language modelling*, i.e. the description of constraints among the words in the language.
TTS synthesis is regarded as the process of converting a string of text into speech waveforms, based on a conversion of orthographic text to a phonological transcription and a concatenation of pre-recorded and stored speech units.

In order to improve the quality of TTS systems as well as the performance of recognition systems, huge databases of annotated speech of different complexity are believed to be needed.

For ASR, segmented and labelled speech databases are useful for training, or at least initialising, the statistical models for the selected recognition units. Especially for large vocabularies and continuous speech recognition, the selected recognition unit is smaller than a word to allow sharing of training material across words. Since many phones can not be uttered in isolation they must be extracted from a context, i.e. segmented and labelled speech corpora are needed.
In addition, an annotated speech corpus helps ASR developers to find what errors are made and perhaps why the errors occur, and to determine the influence of phonetic decoding upon other levels of recognition. Speech databases recorded and annotated in a standardised manner can serve as common testbeds for evaluating the performance of different ASR systems.

Annotated speech databases are not as necessary for TTS systems as they are for ASR systems. However, smaller and maybe more carefully designed and annotated speech databases are required for learning, e.g. by training statistically based TTS, the segmental phonetic rules in continuous speech, such as phoneme durations, stress realizations, assimilations, and elisions. Also, prosody and voice characteristics need to be investigated in order to achieve a more natural sounding synthetic speech. For standard comparison and assessment of synthesizers; especially segmental evaluations, carefully annotated speech databases are valuable.
Different voices can be implemented in TTS if a library of speech segments from different speakers exists.

Hence, annotated speech databases are needed for speech research and for the development of speech technology. In the last few years huge speech databases have thus been collected in laboratories all over the world. Some are multi-lingual, such as the ESPRIT-SAM [Fourcin'90] project's EUROM0 [Grice'89] and EUROM1 [Sherwood'92]. The American TIMIT, RM1, RM2, and ATIS databases are collected within the DARPA project [Pallett'90]. Huge national databases are also complied in Japan, e.g JSDB [Sagisaki'90], ADD [Ehara'90] and ETL [Tankana'90], and France, e.g. BDSONS [Carre'86] and BREF [Gauvain'90]. Before, most of the recordings were lab-speech, but spontaneous speech is now recorded, e.g. ATIS [Pallett'90], MIT VOYAGER [Socolf'90] and ADD [Ehara'90].

Therefore, the specific problem of segmenting and labelling speech signals has become a research area of its own.
Table 1.1 below provides a brief description of some available speech databases.

Name:	Quantities CDs	Hrs	GB	Spkrs	No. & size of sp. units	Transcript detail	TA[1]	Speaking Style	Rec. Env.[2]	SR (kHz)	Sponsor	Avail.
TI Digits	3	~14	2	326	>2500 digit seq.'s	word	N	Read	QR	20	TI	now
TIMIT	1	5.3	0.65	630	6300 sentences	phon.	Y	Read	QR	16	DARPA	now
NTIMIT	2	5.3	0.65	630	6300 sentences	phon.	Y	Read	Tel	8	NYNEX	now
RM1	4	11.3	1.65	144	15024 sentences	sent.	N	Read	QR	20	DARPA	now
RM2	2	7.7	1.13	4	10608 sentences	sent.	N	Read	QR	20	DARPA	now
ATIS0	6	20.2	2.38	36	10722 utterances	sent.	N	Spon/Read	Ofc	16	DARPA	now
Switchboard (Credit Card)	1	3.8	0.23	69	35 dialogs	word	Y	Spon Conv	Tel	8	DARPA	now
TI-46	1	5	0.58	16	19136 isol.words	word	N	Read	QR	16	TI	now
Road Rally	1	~10	~0.6	136	dialogs/sentences	keyword	Y	Spon/Read	Tel	8	DoD	now
Switchboard (Complete)	30	250	15	550	2500 dialogs	word	Y	Spon Conv	Tel	8	DARPA	12/92
ATC	9	65	5.0	100	30000 dialogs	sent.	Y	Spon	RF	8	DARPA	12/92
MapTask	8	34	5.1	< 256	128 dialogs	sent.	Y	Spon Conv	Ofc	20	HCRC	12/92
MARSEC	1	5.5	0.62	?	53 monologs.	phon.	(?)	Spon	varied	16	ESRC	12/92
ATIS2	6	~37	~5	351	12000 utterances	sent.	N	Spon	Ofc	16	DARPA	1/93
WSJ-CSR1	18	80	9.2	>124	38000 utterances	sent.	?	Read	Ofc	16	DARPA	1/93

Notes: 1. TA = Time-Aligned with waveforms
2. QR=Quiet Room, Ofc=Office background, Tel=Telephone line, RF=Radio freq. transmission

Table 1.1 *Some speech databases (provided by Linguistic Data Consortium, October 1992).*

1.3 PROBLEMS WHEN SEGMENTING AND LABELLING SPEECH

1.3.1 IS SEGMENTATION OF SPEECH POSSIBLE?

One point of view is that segmentation of the acoustic speech signal is impossible because neither the articulatory processes nor the acoustic speech signal are composed of discrete segments. Instead, adjacent gestures overlap and merge and so do the phonemic cues in the corresponding speech signal. We adjust our articulation organs from one "target" position for a phoneme to the next and get coarticulation effects. However, transitions between the targets also constitute important perceptual cues for the phonemes and for instance diphthongs are mainly characterised by a transition.

Although we only approximate the target positions and much information is contained in the transitions, this is obviously enough to trigger the impression of segments and phonemes. In our perception we reconstruct the spoken message from the speech signal. This effective and highly automated reconstruction process gives us the impression that it is possible to describe speech as a discrete sequence of phoneme sized segments. Also our writing system based on a linear string of letters may influence our belief in a sound-by-sound utterance.
Hence, the discrete entities may exist at some processing level in our brain, but not in the

acoustic signal.

A more pragmatic view is to give people what they want to have. Since annotated speech databases are needed in speech technology and research, we should try to mark operational and useful boundaries although we know that it is impossible from a theoretical viewpoint. This view was also adopted in the ESPRIT-SAM project, summarized by Adrian Fourcin, in [Fourcin'90,p.12]:

> *"So, although the precise assignment of discrete categories, for different sound classes, to the continuous speech signal is an impossible task - since the subjective level of labelling is not compatible with any physical set of exact temporal stretches of the signal - the consistent correlation is of real value".*

Hence, speech annotation tries to make this correlation explicit.

Also, if we look at the acoustic signal transformed to a spectrographic representation we may perceive it as containing a set of acoustically homogenous regions. The term homogeneous has here to be understood as a smaller change within a segment than between segments, and also to cover same kind of change, such as same formant rising. That is, the interpretation of the word homogenous is here language dependent and not purely acoustically based. Certainly, changes may well be observable during any individual state that one examines, e.g. formant transitions in a diphthong, but this does not detract from the recognisability of the boundaries between the diphthong and its surrounding phonemes. So, in this respect, segmentation of continuous speech is a practical possibility.

The question whether segmentation of the speech signal is possible or not will be left unanswered here. It seems more useful to accept that segmentation of speech in some sense is illogical and impossible, but nevertheless needed in practice. It is a trade-off between what we want to do, what we should do, and what we can do. While keeping this in mind, a pragmatical approach to the segmentation and labelling issue is selected. That means defining the type and level of segmentation and labelling in a way that suits the intended use of the speech material, making the requirements explicit, becoming aware of the problems involved, and performing the annotation in a way that fulfils the requirements as best can be.

1.3.2 PRACTICAL PROBLEMS

Accurate segmentation and labelling of the speech signal is a complicated and difficult task. It involves a proper and conscious representation of our knowledge of e.g. the phonology, acoustic phonetics, and articulatory phonetics of a given language. Since many different experts such as engineers, phoneticians, and linguists contribute in developing speech technology, it is essential to ensure that we use the same words for the same concepts. Hence, first the *terminology* has to be clarified.

Secondly, the speech signal to be annotated has to be described according to *recording conditions* (e.g. anechoic chamber vs. noisy environments), *speaker characteristics* (e.g. sex, size, and voice-quality), and *speaking style* (e.g. labspeech vs. spontaneous speech and isolated words vs. continuous speech), see figure 1.5. Factors such as *speaking rate* and *speech complexity* will also influence the annotation.

Thirdly, different *annotation levels* should be defined and described and the number of levels

used should be argued for. One annotation strategy has to be selected and made explicit. The annotation strategy must state whether it is the actual speech signal or the speaker's intended meaning that is annotated and how the different sounds are segmented and labelled. *Unexpected speech signals* such as hesitations, rare allophones, speaker idiosyncrasies, false starts, and discontinuities should also be accounted for.

Fourthly, decisions must be made for how the actual work should be organised. That is, choice of *equipment* (hardware and software), *information sources* (e.g. listening and temporal information (i.e. the speech pressure waveform) vs. additional spectrograms and energy contours), *transcribers* (e.g. background and how many), and *methodology for checking* the annotation. The *intension of the research* (i.e. purposes of data collection and annotation) should be stated.

Finally, it has to be clarified how to deal with the two main problems encountered when trying to annotate the speech signal, namely the *segmentation* problem and the difficulty related to *labelling* a given segment.
Some practical problems regarding the segmentation and labelling process are:

A. There are no discrete segments in articulation or in the corresponding acoustic signal. Thus the continuous speech signal has to be mapped onto a discrete phonemic transcription. This mapping between acoustic events and a linguistic representation is complex, non-linear, irreversible and only partially understood [Barry'90a]. The boundary placements will thus be *arbitrary* and must be *defined* according to some meaningful strategy; e.g. placing boundaries at certain discontinuities in the waveform of spectrogram.

B. Phonological units such as phonemes are language dependent. Therefore, one symbol may have different interpretation in different languages. Within one langauge an acoustic event may also be associated with more than one phonemic symbol. The phenomena of phonemic overlap, i.e. phonemes sharing the same biallophone, may add more difficulties to the speech labelling. For instance in the Norwegian word "skatt" ('treasure') /k/ is realised as an unaspirated and voiceless [k] and perceived as /k/, and in "ett godt" ('a good (one)') /g/ will often in initial position after a pause also be realised as an unaspirated voiceless [k], but still perceived as a representation of /g/.

C. Although we manage to identify some cues and features which we perceptually regard as essential for identifying a segment, how can we decide which cues are most important for the segment/ sound in question? In figure 1.4, six acoustic cues are indicated which work together to influence our perception of a voiced plosive.

D. There are no international agreements regarding explicitly defined rules for how segmentation and labelling of speech should be performed. Hence, the cues are weighted differently depending on the intended use of the annotated speech corpora, and various people also weight the cues differently even for use in the same application.

E. Even if one has a defined strategy, e.g. for annotation of speech into phonemes, it is difficult to follow this strategy consistently when segmenting and labelling manually. This is partly due to coarticulation and assimilation effects which smear the characteristics of one phone over the adjacent phones. Listening to one such phone segment in isolation may give a quite different impression than when listening to it in context (see e.g. [Cole'92]). Another reason for the difficulty of following one strategy consistently is that the acoustic realisations of phonemic

events are highly variable as a function of speaker, environment, context and occasion, as illustrated in figure 1.5.

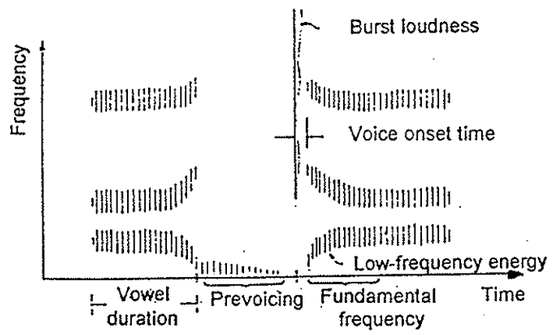


Figure 1.4 *A schematic spectrogram of the utterance /AdA/ showing six acoustic cues to voicing for plosives (after [Klatt'86, p.64]).*

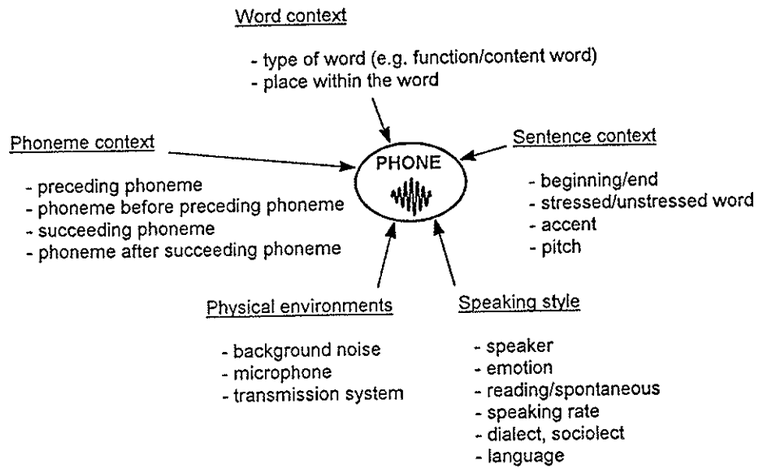


Figure 1.5. *Some contextual effects on the speech sound realization.*

1.4 AUTOMATIC SEGMENTATION OF SPEECH

There are three major drawbacks with manual speech annotation. Firstly, it is extremely time consuming (see chapter 4). An increasing amount of annotated speech data is needed, but the tremendous work load often constrains the amount of speech recordings that can be annotated[1]. Secondly, no standard multi-lingual procedure for speech annotation has been defined yet. Too few segmentation criteria are made explicit, resulting in subjective, labeller-dependent annotations which are difficult to reproduce, both with regard to the placing of boundaries and the labelling of the speech segments.
Finally, the manual segmentation and labelling of speech is prone to human random errors and inconsistencies.

By contrast, *automatic annotation procedures* are by definition free from human interaction. Vast amounts of speech recordings can thus be segmented and labelled by applying a fixed set of objective criteria in a consequent manner. The accuracy of such annotation may be poorer than for human performance, but the errors are more systematic and can hence be taken into account when using the annotated speech material. Since the criteria causing the errors are explicit, some of the errors can be corrected.

The main aim of this thesis work is therefore to develop an algorithm which automatically aligns a given transcription with the actual speech signal, so that the labels are successively allotted to the segments. *Only the segmentation process is thus automatised and the term automatic segmentation of speech, ASS, will be used*[2]. If the labelling of speech also should be done automatically, a complete subword based ASR system would be required.

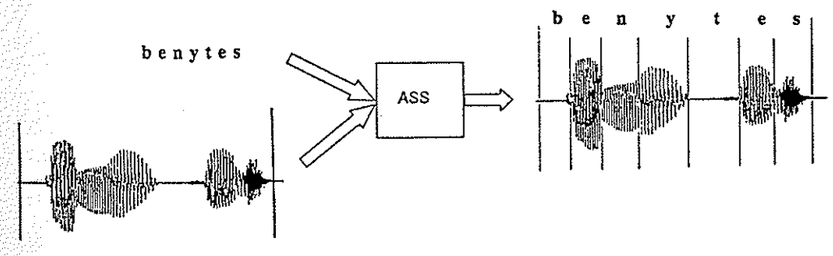


Figure 1.6 *The inputs and outputs to an automatic segmentation of speech, ASS, system, also called a semi-automatic label alignment, SALA, system. The inputs to the ASS algorithm are an endpointed speech waveform and a broad phonetic transcription, and the output is the segment boundaries for each label.*
The speech waveform is 0.55 sec. excerpted from figure 1.1.

---

[1] For instance, at the Advanced Telecommunications Research (ATR), Japan, 6 women annotate speech on a full time basis to meet the demand for training material for automatic speech recognition [Sugiyama'92].

[2] The process described here is automatic phonemic segmentation, but an automatic acoustic segmentation which is often performed prior to the label alignment or phonemic segmentation, is also an ASS method (see chapter 6).

Within the SAM project the above described approach is referred to as SALA - Semi-Automatic Label Alignment. This is a misleading term since the label alignment per se is automatic. The reasons for using the term *semi-automatic* can be: (i) SALA refers to alignment, i.e. the labels are prespecified (ii) the endpoints of the utterance are given, and/or (iii) the ASS may be too inaccurate for certain applications and manual correction of the boundaries may prove necessary.

In the development and testing of the ASS algorithm the given transcription was the string of labels produced manually by visual and auditory access to the signal. A further automation of the annotation process can be achieved if the ASS program is able to take an auditory transcription or a transcription made by a text-to-phoneme (TTP) converter as input. The ASS algorithm will then become more complex because it has to account for what phonemes the different speakers may miss out or insert.

1.5 SUMMARY

In this chapter segmentation and labelling of speech is defined and argued for. When segmenting and labelling speech we have to make theoretically doubtful compromises due to the nature of speech, generally termed as contextual effects. When automatic speech recognition systems fail to recognise the same labels or when automatic segmentation procedures place the boundaries at other instances than the manual ones, coarticulation is often given as the explanation; i.e. a circular argument.

The segmentation strategy and labelling units selected for different speech corpora, depend upon the intended use of the annotated speech material. However, in order to share annotated speech material among different sites and languages, a standardised, multi-lingual speech annotation strategy has to be defined. This thesis work is intended to contribute towards this standardisation.

The main aim of this thesis work is to develop an algorithm for automatic segmentation and labelling of speech. A goal for the development and refinement of automatic segmentation algorithms is to produce segment boundaries with accuracy comparable to the performance of human labellers[3].

[3] As regards accuracy in terms of coincidence with manual segmentation, see evaluation in section 6.3. For instance, 90-95% coincidence within ±20 ms deviation from the manually segmented ones may be considered a reasonable goal.

Chapter 2

SPEECH PRODUCTION AND PERCEPTION

For development of simple speech products we may achieve good results quickly by trial and error or by applying intuitive understanding of relevant principles. However, in order to segment and label speech scientifically and to improve the automatic production and recognition of speech, we first have to acquire knowledge about speech and language and about how humans produce and perceive speech. We still know too little about these processes and more research remains to be done before we e.g. can develop algorithms for spoken dialogues with computers.

Human language is an oral-auditory communication system. This is a convenient way to communicate: we can do something else while speaking, we do not need any artificial equipment, we speak with very little effort, and speech is very flexible.
Through writing we can communicate across space and time more directly. Printed text is made up of a fixed set of discrete, stable and context-independent symbols and there is space between the words. Speech sounds on the other hand, vary continuously, are never realized identically, the cues overlap, merge and depend on the context, and there are usually no pauses between words in an utterance.

Many researchers have examined the speech production processes, the speech signal and the hearing and the speech perception processes e.g. [Malmberg'68], [Flanagan'72], [Fant'73], [Schafer'79], [Ladefoged'82], [Taylor'90] and [O'Shaughnessy'90]. No attempt is made here to give a thorough survey of these investigations, but rather a framework useful for the succeeding discussion of segmentation and labelling and mathematical modelling of speech will be provided.

2.1 SPEECH PRODUCTION

The description of speech in this thesis is confined to speech sounds made on a pulmonic egressiv airstream mechanism. This means that the respiratory system expels air from the lungs through the larynx, the pharynx and the vocal tract (i.e. the oral and nasal cavity). If the vocal cords are close to each other and the muscles properly strained, the air stream may cause them to vibrate; they move apart and together quasi-periodically. The strongest output of this excitation measured at the lips occurs just after the airflow from the lungs is suddenly stopped as the cords are pulled together. This characterises voiced sounds, and the rate of vocal cord vibration is measured as the fundamental frequency F_0, also called the pitch. For these voiced sounds the pharynx and vocal tract constitute a resonance cavity where some frequency components are amplified and others attenuated. These frequencies are seen as intensity peaks, called formants, and valleys in the spectrum of for instance a vocoid. While we are talking the shape of the resonance room is continuously adjusted. This constant changing affects the frequency resonances and thus causes different sound qualities.

If the vocal cords are apart the air has relatively free passage from the lungs up through the pharynx, and the acoustic energy may be generated by the turbulence resulting from a constriction created by the tongue or the lips. This is typical for the unvoiced sounds; especially voiceless fricatives as /s/, /S/ and /f/. A combination of the friction excitation and the voiced excitation generates the voiced fricatives, e.g. /z/ and /v/. Plosives, e.g. /p/ and /b/, are produced by first making a complete closure in the vocal tract, then pressure is built up behind the closure and finally the closure is released and the air is expelled as a plosive burst.

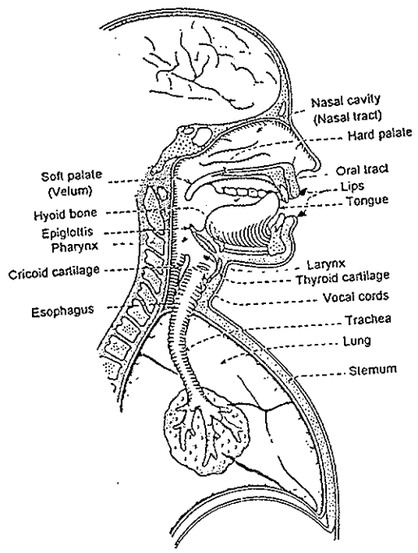


Figure 2.1 *The speech production organs (After [Flanagan'72, p.10]).*

We usually talk because we want to express our ideas or feelings to other people; that is, we talk with someone. Hence we must regard speech both from the speaker's and listener's point of view. As *speakers*, we know or try to find out how much the listener knows about the language, culture, and the subject, and we adjust our choice of words, speech rate and articulatory effort accordingly. We do not speak too fast or use too troublesome words but we do not talk too slowly or too simplified either. Through social activity we have acquired a certain level of communicative competence so that we are able to choose words and style suitable for the actual situation.
Given these constraints on the communication the speaker often likes to convey his ideas with the *least possible articulatory effort* [Ladefoged'82]. Coarticulation and assimilation effects, such as reducing the differences between phones to a minimum and leaving out segments, are partly results of this ease of articulation principle.

Listeners, on the other hand, wish to perceive the meaning from the speech signal with *least possible perception effort*. Hence, if two phonemes differ only by few features they will probably

be realised more distinctly in word positions where they both are allowed to occur than where only one of them can occur in a language [Ladefoged'82].

Listeners do not have to recognise all words of an utterance to understand the meaning; due to redundancy often only understanding of some keywords are needed. Those parts of the utterance that contain important information, the *keywords* or *content words*, are therefore pronounced more carefully and with more vowel contrast than parts with a low information load [Beinum'91]. *Function words* like "a, and, are, for, in", are generally unstressed and indistinctly articulated and their pronunciation depends heavily on context [Rabiner'89b]. It is also shown that change in the grammatical function of a given segment implies a change in the acoustic properties of this segment [Lourdes'91].

2.2 SPEAKING STYLES

In the previous section the production of speech was schematically described. However, all humans speak more or less differently due to differences in anatomy, age, sex, dialect, articulatory habits, sociolect, and even the same person cannot produce e.g. two words absolutely identically twice. Consequently, all human speech sounds are unique. In addition, we change our *speaking style* according to the given situation and who we are talking with. Our mood and physical condition also affect our speech.

Different speaking styles influence the speech signal and accordingly the segmentation, labelling and ASR processes. Although everybody speaks differently, speaking styles used for speech processing analysis may roughly be grouped into: *lab-speech*, *professional discourse*, and *spontaneous speech*. In the following these three speaking styles are described and compared.

Lab-speech is written text carefully read aloud e.g. in an anechoic chamber or silent office environment[1], often by a professional reader. The reader has rehearsed the given text and reading mistakes are excluded. The informant is asked to read in a neutral fashion. All extralinguistic sounds, such as stomach rumbling, breath noise, and lip smacking are usually removed or attenuated as much as possible.

ASR systems have till recently been trained on lab-speech. Choice of reading mode gives the ranking: isolated words - connected speech - continuous speech. For the *isolated word* speech material one single word is read carefully at a time, and the words become stressed and there are no coarticulation effects from other words. In *connected speech* a list of words, or concatenated words, such as digit strings for telephone numbers, are read aloud with no pause between the words. Consequently there will be coarticulation and assimilation effects over word boundaries and the words become less stressed. *Continuous speech* is meaningful sentences read aloud. Usually the sentences are read without any connection to other sentences, as for e.g. the TIMIT database [Pallett'90], and the filler sentences in the EUROM1 speech corpus [Sherwood'92]. Isolated sentences tend to be read with the same sentence intonation; for instance always lowering the pitch towards the end of the sentence.
For these three categories; isolated words, concatenated words and isolated sentences, the readers often rehearse so that they memorize the text. Thus, in some sense this is not reading and the

[1] Speech recorded in adverse environments will not be treated in this thesis.

informant only has to concentrate on careful pronunciation.

A *continuous passage* on the other hand, is a meaningful story read aloud. It contains more than one sentence and the sentences are related. For this type of recording it is not usual to memorize the whole text. Compared with the other reading styles the prosody in a continuous passage may differ and stronger assimilations and reductions may occur.

By the term **professional discourse** we mean speech as used in formal professional situations, e.g. reading news on radio or TV. It is a monologue produced for a large audience. The newsreaders are well prepared and they tend to seem spontaneous but neutral. They are taught not to hesitate unnecessarily, because hesitations make the speech less fluent and waste time necessary for communicating information. They are advised not to correct minor slips of the tongue, because the audience will often ignore small mistakes in pronunciation. If the newsreaders use teletext, i.e. text prompted near the camera, their reading style becomes very similar to lab-speech.
These constraints on the reading result in a rather fast speaking style with almost no hesitations and a large amount of information is transmitted in a controlled time.

A **lecture**, defined as a prepared talk to a present audience also fits into the professional discourse speaking style category. The speaker may have the text available, but he is rarely reading. Most often the content is planned in advance, but not the exact form. This leads to a more relaxed and natural speaking style than newsreader style.

Spontaneous speech covers the styles used in situations where it is natural to talk. Thus, it is not one special style, but rather an oscillation between different ones. For many commercial applications of ASR spontaneous speech has to be acceptable as input.
One common procedure for collecting spontaneous speech material is so-called triggered monologues, where the (unprepared) informant is asked questions which require extensive answers. For instance, the recordings can be arranged as an interview about childhood, career, professional life etc. in a studio [Blaauw'91], [Delgado'91], or as interviews of ordinary people in the street or sportsmen shortly after a competition. Another possibility for collecting spontaneous speech material is dialogues/discussions where a group of people are chatting with each other and may forget about the microphone.

Many more processes are involved in spontaneous speech than when reading a text aloud. Before we articulate a sentence, we have performed various cognitive activities such as creating a message, retrieving the right words to convey the intended message, and formulating a structured sentence with words of right grammatical classes in the right order. Since we have rather little time to organise and refine our spontaneous speech we often produce incomplete sentences with much redundancy. As far as formal grammatical rules are concerned, spontaneous speech is thus more prone to errors than labspeech. When reading a text aloud, the verbal planning is easier[2], and hence we can concentrate more on the articulation.

Up till recently most of the speech material used in speech research and ASR has been texts read aloud. Phonological theories have been based on lab-speech, and accordingly researchers often found what they predicted. ASR systems trained on lab-speech are shown to fail in practical

[2] The message and its syntactic and lexical form are determined by the text; not by the speaker.

applications, basically because people speak differently from lab-speech. Now some research activities have shifted from the study of language as a system towards a study of language as used in communication. Both for speech research and speech technology the analysis of phonological rules in spontaneous speech has become more important.

Humans perceive differences in speaking styles even when lexical, syntactic and semantic structures are identical. In an experiment, [Blaauw'91], spontaneous speech was recorded and auditorily transcribed, including all silent pauses and vocal hesitations (i.e. filled pauses, repetitions, false starts, and extensively lengthened syllables). Emotionally neutral monologues were constructed, embedded in a context equal or similar to the original context, and then read aloud by the same speaker. Afterwards the sentences were played in random order and listeners judged whether an utterance was spontaneous speech or lab-speech. 86% of the spontaneous speech utterances and 79% of the read aloud sentences were correctly classified. No single cue, but rather a combination of cues seemed to give the impression of the speaking style. For example, the read aloud sentences tended to be more fluent, and had higher average fundamental frequency (F_0=157 Hz) than the spontaneous speech (F_0=130 Hz).

We still know too little about speech and speaking styles to be able to decide the style directly from the speech signal. But parameters such as *speaking rate* (i.e. number of words per second when both speech and pauses are considered), *articulation rate* (i.e. number of words per second during effective speech time), filled and unfilled pauses, speech errors, reductions, duration, timing, and pitch are measurable and can be used as indicators to characterise speaking styles. However, since many of these parameters change during an utterance, only average values can be employed. E.g. if we compare articulation rates in lab-speech, the fastest speaker does not necessarily have the highest speaking rate in all parts of the text; the so-called "lack of uniformity" phenomena [Fant'91]. In addition, there is also a wide variation in habitual rate of articulation (see also Appendix B and Appendix C).

2.3 HEARING AND SPEECH PERCEPTION

So far speech in terms of how it is generated has been discussed. But the receiver part of the speech chain is at least as important as the transmitter part. Actually, the rules for speech production may have evolved to ensure acoustic signals appropriate for the auditory system. It is not unreasonable to assume this since the human auditory system is similar to that of other mammalians and it was developed long before speech communication. Our hearing is also useful for other tasks than filtering and transformation of speech signals, e.g. for spatial orientation and detection of sound sources.

Insight into the perception processes may help us to look for and select proper cues and units in segmenting and labelling speech signals. It is important to utilize such knowledge in other fields of speech technology too. For instance in TTS or speech coding it is a waste of effort trying to reproduce features below human perception thresholds. ASR systems should emphasise features that are audible to humans, since a speaker is unlikely to control features that he cannot perceive. Below are discussed briefly some fundamental hearing and speech perception processes.

2.3.1 HEARING

The speech pressure waveform propagates through the air and finally part of it reaches our ears and sets the eardrums and the organs in the middle and inner ear into motion.

The external channel terminates at the eardrum. Its form forces frequency components at about 3 kHz into resonance[3]. Partly because of this amplification (12-15 dB [Shaughnessy'90]) we are most amplitude sensitive to frequencies around 3 kHz.

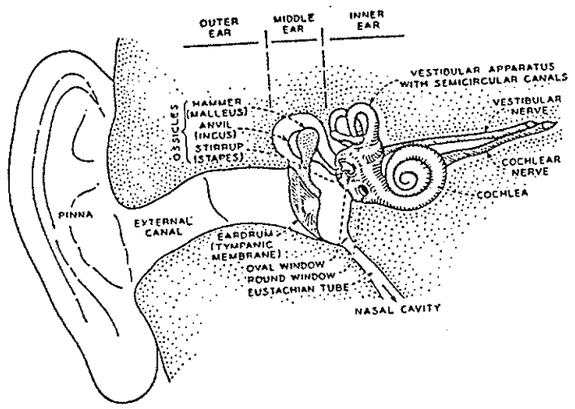


Figure 2.2 *Schematic diagram of the human ear [Flanagan'72, p.87]. The drawing is not to scale: the middle ear and inner ear are enlarged.*

When the eardrum vibrates the neighbouring ossicle bones have to vibrate too. In the middle ear the pressure is amplified (about 30 dB [Shaughnessy'90]); partly because the hammer increases the force of the eardrum movements by about 40% [Flanagan'72], and partly because this force is transferred onto the stirrup which has a much smaller surface area than the eardrum (*pressure=force/area*). Frequency components above 1 kHz are attenuated by about 15 dB/octave due to the low-pass characteristic of the middle ear [Flanagan'72], [O'Shaughnessy'90].

The vibrating stirrup generates volume displacements of the cochlear fluid, which in turn lead

[3] The external canal can be modelled as a "quarter-wavelength" resonator, where the length of the resonator l is related to the wavelength λ_n by $l=(\lambda_n/4)(2n-1)$. Since the speed of sound in the medium, c, is related to frequency and wavelength by $c=f_n \lambda_n$, the natural frequencies f_n are given by $f_n = {}^{[(2n-1)*c]}/_{(4*l)}$ [Flanagan'72], [Zue'89]. Hence, if $l \approx 3$cm and $c \approx 34000$cm/sec, the first resonance frequency is $f_1 \approx 2800$ Hz.

to local deformations of the basilar membrane[4]. The frequency resolution along the basilar membrane is best at low frequencies and maximal deformation on each point of the membrane depends on the frequency; the so-called characteristic frequency. Consequently, different hair cells on the basilar membrane will bend causing an ionic current. The receptor potential response of a hair cell to acoustic tone bursts have a d.c. component and a "synchronous" component [Zue'89]. The synchronous component decreases when the stimulus frequency increases. The succeeding nerve fibres are able to give a synchronous neural firing-rate to an input stimulus of frequencies below 4-5kHz. The neural firings then propagate from the cochlea through the auditory pathway to the brain.

The ear is basically a frequency analysis instrument. Because the ear is not able to follow high frequently stimuli synchronously, we have less frequency resolution at higher frequencies. This is compensated by a better temporal resolution at high frequencies.
The acoustic properties of speech signals suit these principles. The fact that the higher formants of voiced sounds have larger bandwidths is not noticed by the listener, but small variations in the lower formants are perceived. For instance formant transitions of a vowel constitute strong cues to the identification of the following plosive. Voiceless sounds on the other hand, are characterised by high frequency energy with a rather rough spectral structure, but they contain events that are sharply defined in time. This property helps us to distinguish e.g. /s/ from /t/ (because the nerve fibre rate response is strongest to sudden-onset stimuli [Zue'89], such as the burst onset of a plosive).

This suggests that in manual segmentation tasks we should use another approach to the *voiced* sounds, which require good *frequency* resolution at low frequencies (below 2kHz), than to the *voiceless* sounds which require good *temporal* resolution at high frequencies (above 2kHz).

Another important property of the hearing processes is that strong frequency components mask the ear's response to weaker components. There are two main types of masking; temporal and simultaneous. In *temporal masking* the strong component obscures a weaker one that follows (forward masking) or precedes (backward masking) at the same or a neighbouring frequency. For instance when a weak fricative, e.g. /f/, follows a vowel, the beginning of the fricative may be masked. Especially short fricatives can be completely obscured. Backward masking can suppress important cues in the burst of a plosive, when this precedes a stressed vowel [O'Shaughnessy'90].
In *simultaneous masking*, also called *frequency masking*, the strong component suppresses weaker components occurring at the same time but at other frequencies. Most often lower frequencies mask the higher frequency ones. The intensity tops of the spectrum, mainly the three lowest formants F_1, F_2 and F_3, of a voiced sound are important cues to the impression of that sound [Klatt'82]. This is partly because the valleys between the formant tops are suppressed by simultaneous masking effects.
The effect of temporal masking decreases with time distance and frequency masking decreases as the frequency separation between the components increases, but the decrease is less when the weaker component has the highest frequency [Hunt'90].

[4] The cochlea consists of three chambers: scala vestibuli, scala media and scala tympani. Reissner's membrane separates the two first ones and the basilar membrane separates the two last ones, i.e. scala media and scala tympani.

The masking effects are utilized for instance in many speech coding schemes to determine optimal perceptual noise shaping and to allocate bits only to signal levels which are possible to perceive, e.g. [Johnston'88], [Flanagan'91], [Jayant'92]).

2.3.2 SPEECH PERCEPTION

Speech has proved to be a useful tool for transmitting ideas from one person to another. However, the speech signal itself does not resemble either the articulatory gestures that produce it or the auditory impressions it gives rise to. That is, articulatory gestures are only represented in the speech signals they produce; they are not directly mapped onto the signal. On the other hand, we perceive speech as a discrete sequence of phoneme sized segments, although the sounds are not separated in the speech signal. *This discreteness is thus imposed on the stimulus by the perceptual abilities of listener.*

It is from the speech signal that listeners have to extract the correct phonetic and prosodic cues, derive the speaker's intended meaning and make the appropriate associations. In this reconstruction process we exploit our experience and knowledge about the cultural background and language (e.g. vocabulary, syntax, semantics, and redundancy), speaker idiosyncrasies, the subject in question and so on. Indeed, **perception** can be defined as "*an interpretation in light of experience*" [Dew'77, p.6].

Based on knowledge sources mentioned we generate a set of expectations of what is meaningful in the given circumstances. Cues extracted from the speech signal together with visual impressions are thus used to confirm and refine or reject what we expected.
Predictable[5] words, as function words, are thus easier to perceive correctly. Since few phonetic cues are needed to perceive these words and since they often are not important for the meaning, they are often pronounced rather indistinctly.

Perception experiments, e.g. [Taylor'90], have demonstrated that prediction is an important part of speech perception. For instance, words uttered in a sentence are recognized more easily than those uttered in isolation, and words in the middle and at the end of a sentence are recognized more easily than those at the beginning. Phonemes tend to be incorrectly perceived more often when uttered in a redundancy-free context than in a redundant context [Ohala'90].
Masked or mispronounced sounds will, due to this expectation effect, be more or less automatically recognized as correct sounds if they lead to plausible words. This is often referred to as **phoneme restoration** or **phoneme correction**, e.g. [Ohala'90].

All human languages are acquired, i.e. we gradually learn the set of conventions needed to communicate meaning in our society. People talking the same dialect have developed the same **phonemic filter**, which enables them to extract the same cues from the speech signal and to weight and combine these cues to derive the intended messages. *Perception of speech sounds is thus a product of the external stimulus and the listener's internal phonemic filter.*

As stated in section 2.2, no human speech sound can be reproduced identically. Hence, it would

[5] A word is predictable if it e.g. often occurs in general use, is syntactically constrained within a sentence and semantically constrained by the topic.

be impossible to let all sound variation carry meaning. Instead, we categorize the sounds into sound classes such as phonemes, i.e. we suppress and ignore irrelevant sound peculiarities with our phonemic filter and perceive only the phoneme identity. Listening tests, e.g. [Repp'84], indicate that this **categorical perception of phonemes** is most prominent with consonants. Vowels are perceived more as "/e/-like" or "/i/-like" [Hunt'90].

In one study, [Kuhl'83], human listeners were to identify synthetic CV syllables (two formant stimulus). The syllables varied in 15 acoustically equal steps to cover /b&/, /d&/ and /g&/, and they were played singly and in random order. The results depicted as the whole line in figure 2.3, show that the listeners perceived three distinct phonemic categories, i.e. they identified stimuli 1-4 as /b&/, 5-11 as /d&/, and 12-15 as /g&/ and *not as a continuum of sounds.* Even for CV's near the "phoneme boundary" almost no casual categorisations appeared. The dotted line shows the performance when the listeners had to discriminate between two stimuli. This result indicates that it is easier to discriminate two sounds that are perceptually related to different phonemes than if they "belong" to the same phoneme. This phenomenon is often referred to as the **phoneme boundary effect** [Kuhl'83], [Taylor'90].

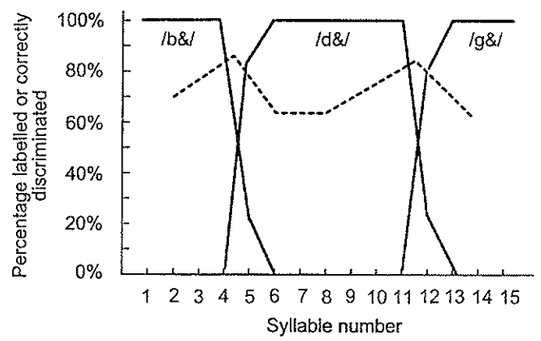


Figure 2.3 *Categorical perception of phonemes, whole line and phoneme boundary effect, dotted line. After [Kuhl'83, p.1005].*

Several researchers have tried to explain how we manage to perceive phonemes categorically from inconstant acoustic cues in the speech signal. For instance the *motor theory of speech perception* [Liberman'57], claims that we extract the meaning from speech by exploring our knowledge of how we articulate the sounds. Hunt, [Hunt'90, p.6], argued that "if the motion of the articulators could be derived directly from the speech waveform, it might provide a particularly good representation of the perception of speech". He claimed that speech perception processes are difficult to explain without resorting to a speech production model.

However, the fact that humans also perceive colours categorically [Kopp'68] and that monkeys categorically perceive speech sounds [Kuhl'83], cannot be explained by the motor theory. This theory also makes it problematic to explain how children are able to distinguish phonemes before they articulate them properly themselves.

characterised by local friction and complete closure respectively[1].

By means of a *phonetic transcription* we can transcribe the speech signal as accurately as we wish, or manage, using diacritic symbols together with the usual set of symbols, such as IPA [IPA'89] or SAMPA [Wells'92]. In phonetics we describe the signal using a continuous scale for the features, as opposed to the phonemic classification into discrete items. In such narrow transcriptions of for instance the vocoids based on the Cardinal vowel system, we may first plot the positions of the vocoids in the vocoid trapeze when pronounced in isolation by a given speaker, see figure 3.1. When the vocoid quality is changed in continuous speech, deviation from the "isolated position" can be marked by diacritics. Contoids are usually described by place and manner of articulation, i.e. where the main sound source (primary articulator) is placed, and if the sound is voiced/voiceless, retroflexed, has total closure/narrow passage etc. It may be problematic to decide how accurate is accurate enough, and consistently stick to this level of accuracy. Row B of figure 3.2 exemplifies two levels of "narrow phonetic" transcription where the diacritic symbols are IPA ones.

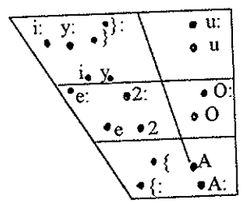


Figure 3.1 *The vocoid trapeze for speaker AFN, (see chapter 4), with dialect background from the southern part of Vestfold county, Norway (After [Foldvik'92]).*

Phonology (phonemics) describes the system, e.g. the syllable structure, and the function of the speech sounds within a language. The *phoneme* is defined as the smallest linguistic unit that has distinctive function in a language. Hence, the phoneme is not a sound but an abstract linguistic class. We classify the speech sounds into phoneme classes according to the function the phonemes have in the given language. By means of phonemic transcription we cannot choose the level of accuracy, but classify the speech sounds using a predefined set of abstract, linguistic symbols.
The phonemes are grouped into vowels and consonants, where the vowel can be defined as the core of the syllable in the language (see figure 3.4). This definition may give rise to some problems, for instance when a nasal or a lateral becomes syllabic in a nasal or lateral plosion.

By **phonemic transcription** we mean the phonemic representation a word is given in e.g. a lexicon/dictionary. This represents the symbolic interface between the speech signal and the orthography which is needed both for ASR and TTS. However, this is a phonemic description of a word when carefully pronounced in isolation, also called *strong form* or *citation form*, and

[1] However, a nasalized sound may be regarded as a vocoid since the definition does not exclude sounds where the air escapes both through the mouth as well as the nose.

it is not always clear what is the ideal or perfect pronunciation, and this may differ from one person to another. Another problem is that we wish to segment and label continuous speech which is much more characterised by assimilation and reduction effects both within words and over word boundaries.
The phonemic transcription of continuous speech is within the SAM community called "**broad phonetic**" in order to distinguish it from the citation phonemic transcription [Barry'90 a], and this term is also employed in this thesis, see e.g. row A in figure 3.2. With broad phonetic labelling each phonemic label will cover more acoustic-phonetic variability than citation phonemic labels.
Hence, the phonetic transcription is here termed as "**narrow phonetic**" transcription.

Broad phonetic transcription has several advantages compared with narrow phonetic transcription. Firstly, the former is much faster when transcribing manually, and even though the classification may seem cruder, many phonetic details can be supplemented by allophonic rules governing the language in question. Secondly, the phonemes represent the smallest number of distinctive phonological classes to be recognized in each language; often in the order of about 40, i.e. it maximises phonetic information with a minimal set of symbols. In comparison, there are theoretically 40^2=1600 different diphones, about 4500 demisyllables and 20000 different syllables (in some languages at least). Thirdly, the phonemes correspond more closely to the lexicon entries. Finally, the categorical perception of phonemes (see section 2.3.2) indicates that even if each phoneme can be represented by a range of sounds, listeners tend to perceive them categorically as realizations of one and the same phoneme[2].

No matter how narrowly we transcribe, or how small units we choose for segmenting and labelling the speech signal, the annotation units are always an abstraction. Hence, for some purposes we wish to describe the speech signal with purely **physical labels** which represent no abstraction. Each type of analysis method provides its own set of labels[3]. Based on figure 3.2, labels as periodicity, aperiodicity, zero-crossing density, F_0, silence, high frequency noise and energy related parameters as low-frequency energy, high-frequency energy and total energy, can be selected. Here we may get overlapping segments, although this is not the case for the examples shown in row D of figure 3.2.

In many sciences, and especially in natural sciences, we want to make verifiable measurements and state our knowledge in terms of numbers. Ladefoged [Ladefoged'82], suggests that an ideal transcription of an utterance should describe the target position of the articulators of each sound and a specification of the rules for moving from one target to the next. The articulatory targets should then be stated in terms of numerical values of distances between the vocal organs. Another possibility is to specify the sounds in terms of percentage values of the prime features as shown in [Ladefoged'82, p.267].

[2] Also, the phoneme seems to be a natural unit for probably most mother tongue speakers with reading and writing ability. This statement may be supported by counting sound based errors, or slips of the tongue, where phonemes are overwhelmingly represented (about 90% of the occurrences reported by [Taylor'90]).

[3] For instance the Electropalatography (EPG) [Nordli'91, p.61] and Magnetic Resonance Imaging (MRI) picture sequences [Nordli'91, p.52] in contrast to the waveform and spectrogram representation depicted in figure 3.2.

For speech-knowledge acquisition and rule development we may need a more detailed annotation of speech than the narrow phonetic transcription, but still cruder than the physical annotation. For example in a broad phonetic annotation certain acoustic landmarks such as the boundary between closure phase and release of a plosive might be marked. Generally, this is called the **acoustic-phonetic level** of speech annotation [Barry'90 a], and is applied e.g. for the annotation of the TIMIT database [Seneff'88]. At this level of annotation the speech signal is divided into discrete, acoustically homogenous parts, and these segments are marked with phonetic labels. Since such purely acoustically homogenous speech segments do not exist, we have to interpret homogenous as a more or less stable portion, using a threshold for defining this more or lessness. This leads to a language independent definition of homogeneity, and segment boundaries can be automatically found entirely based on the speech pressure waveform. Convenient labels at this level can be friction noise, stop closure, stop release, aspiration and broadly classifying a quasi-periodic signal portion as vocoids or voiced contoids. These are all language independent terms. However, the interpretation of acoustic homogeneity given in section 1.3.1, was language dependent, and often these acoustic-phonetic segments are marked in such a manner that the narrow phonetic and broad phonetic segmentation and labelling can be automatically derived from them. In this case the segmentation and labelling becomes language dependent also at the acoustic-phonetic level of transcription.

In row C of figure 3.2, an acoustic-phonetic segmentation and labelling is depicted. The segments are:

1. front, half closed vocoid, 2. voiced closure, 3. voiceless stop closure, 4. release burst, 5. aspiration, 6. central vocoid, 7. voiceless friction noise, 8. back, open-mid vocoid, 9. nasal, 10-13. as 2-5 respectively, 14. front, close vocoid, 15. voiced stop closure, 16. nasal, 17. voiced palatal approximant, 18. front, close vocoid, 19-22. as 2-5 respectively, 23-25. as 3-5 respectively, 26. central (close-mid) vocoid, 27. nasal, 28. central vocoid, 29. creaky voice, 30. front, close vocoid, 31. nasal.

For TTS purposes, the speech databases should be segmented and labelled both into detailed acoustic-phonetic and prosodic entities [Pols'90]. **Prosodic annotation** is important e.g. for developing more natural sounding synthetic speech as well as for improving automatic speech recognition and understanding systems. Since prosody is a cover term for accent and intonation, prosodic labels have to describe miscellaneous cues, such as duration of vowels, stress and tones of syllables and intonational patterns for different dialects and attitudes. The prosodic level of annotation is hence not one single level. Instead, we can describe prosodic cues using all four levels explained above, as for instance **physical parameters** such as duration and intensity, or **acoustic-phonetic** description of whether the fundamental frequency is rising or falling. **Phonetic events** such as rise, fall, and stress can also be added. Turning to broad phonetic or citation phonemic transcription, these categories usually depend on the prosodic and phonemic theory subscribed to by the labeller, but entities as tone unit, nuclear tone, sentence accent, pitch accent and juncture are proposed for transcription of multi-purpose speech databases [Barry'90 a]. Work is going on to establish an international standard of prosodic annotation of speech analogous to IPA for phonetic segments, e.g. SAMPROSA (SAM Prosodic Alphabet) [Gibbon'90] in Europe and TOBI (Tones and Break Indices) [Silverman'92] in USA.

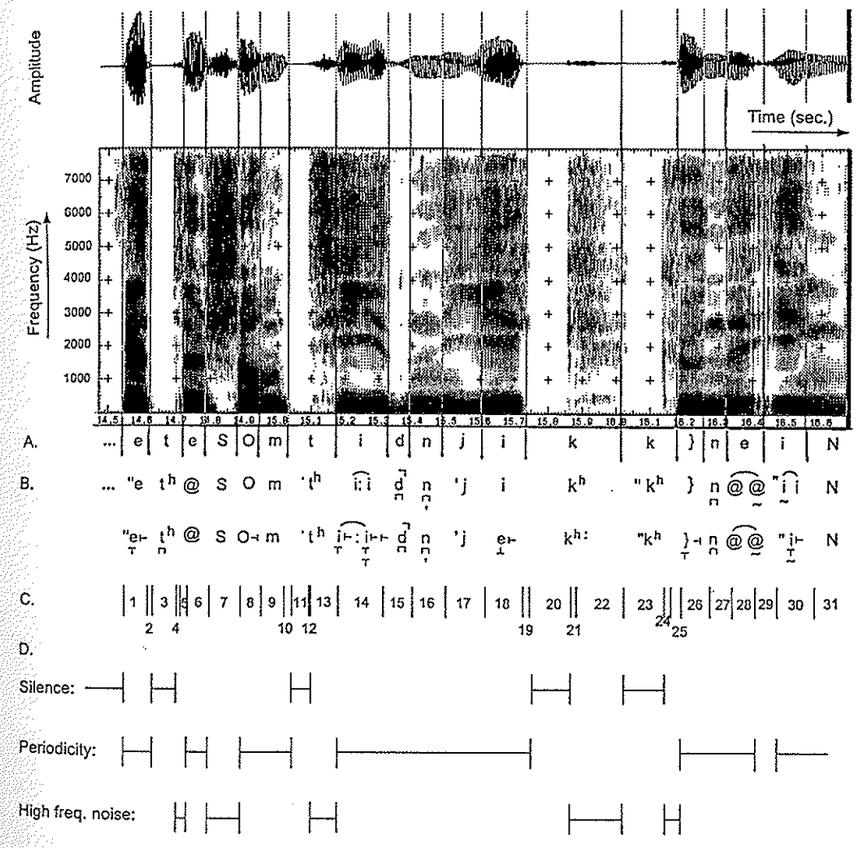


Figure 3.2. *At the top the waveform and the broad band spectrogram for the Norwegian sentence "etterhvert som tiden gikk kunne ing(en)" ('as time passed nobody could') read aloud by a male speaker are depicted. The + signs in the spectrogram are separated by 0.1 sec. horizontally and 1000 Hz vertically.*
This sentence is manually segmented and labelled at different levels:
A. Broad phonetic, B. Narrow phonetic, C. Acoustic-phonetic, D. Physical.
In the narrow phonetic level IPA diacritic symbols [IPA'89] are used.

3.2 SUBWORD REQUIREMENTS

The subword units may represent a phoneme, a more complex spectral evolution such as diphones and demisyllables, some sub-phonemic units such as closure and burst release in a plosive, or acoustically defined units.

The intended use of the subwords to some extent sets the requirements for the subwords. As mentioned in section 1.2, for TTS a small, but carefully annotated speech database is desirable, whereas for ASR purposes, a vast amount of more coarsely annotated data is required. However, in all types of speech segmentation and labelling the criteria and conventions used should be *explicit* and *applied in a consistent manner*. Only then it is possible to use the annotated speech material for different applications, to exchange annotated speech material, to incorporate new speech knowledge and speech data later, and to reproduce and to automate the annotation process. Unfortunately we have still got too little acoustic-phonetic knowledge to fulfil these requirements.

For annotation of **general purpose speech databases** the subword units should be:
-*Well defined*, so that they can be automatically and reliably segmented [Huckvale'90]
-*Flexible*, so that new representations can be derived from the selected ones [Barry'90 a] (e.g. a syllable can be split into phonemes and concatenated syllables can constitute a word).
-Suitable for *acoustic-phonetic observations* [Nakatsui'86].

The fundamental condition for a practical use of such speech databases is that the speech can be annotated in a reproducible manner. For use in multi-site data collections the annotation strategy has to be convenient and relatively easy to learn.

Appropriate subwords for **ASR** should be:
-*Consistent*, so that different instances of a unit have similar characteristics [K.F.Lee'90b].
-*Economic*, so that the subword inventory is small, about 100 subwords [Colla'89b].
-*Comprehensive*, so that most of the contextual effects are contained within the unit [Rosenberg'83].
-*Suitable for concatenation*, so that words can be composed and recognised as a sequence of the selected subwords [Soong'89].

Consistent subwords, i.e. subwords that are realised relatively similarly in different contexts, are fundamental in order to get maximum discrimination between the different subwords. In ASR we face the trade-off that larger subword units are consistent, but can be difficult to train, whereas smaller units are trainable but inconsistent. That is, we have to strike a balance between the accuracy of representation and the quality of models.

Appropriate subwords for **TTS** should be:
-*Suited for concatenation*, so that they can be concatenated without too complex smoothing.
-*Economic*, so that the inventory is small enough, about 1000 subwords, to be read by a single person in a uniform speaking style.

For synthesis of unrestricted text speech is usually produced by a concatenation of subword realisations. In TTS there is a trade-off between maximizing speech quality and minimizing memory space, complexity, and computation time [O'Shaughnessy'88].

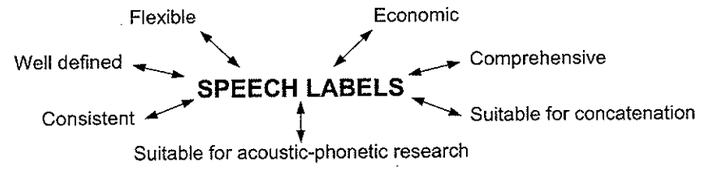


Figure 3.3 *General requirements for the speech labelling units.*

Some of the subword requirements stated above overlap. Consistent units have to be well-defined and insensitive to context, and are thus comprehensive.
As new knowledge is acquired and new needs crop up, the label set should be easily extended to incorporate the new phenomena [Barry'90 a].

3.3 SUBWORD UNITS

Humans tend to perceive the continuous speech signal as discrete sounds following each other in temporal order. We are therefore misled to think of a phoneme size entity as the natural segmentation unit for speech, and phonemes are often called the segmental unit of speech. However, since there is rarely such separation in the acoustic signal, the selection of units in speech technology is more directed by what is practical and useful in the intended application of the annotated speech material.

The main difference between subword units is whether they are defined merely based on the acoustics, i.e. acoustic subwords, or if they are language specifically, or phonologically defined, i.e. phonemic subwords. In this section some of these units are described.

3.3.1 ACOUSTIC SUBWORDS

The acoustic subwords differ from the phonemic subwords enumerated in this chapter in that they are not directly related to a phonemic symbol. It is thus difficult to apply the requirements in section 3.2. to the acoustic subwords. The motivation for this type of subwords is that phones are regarded as being composed of stationary and transitional segments which can be represented by consecutive acoustic subwords.

The acoustic subwords, also called acoustic segment units [C.-H.Lee'88] or output from a segmenting acoustic processor [Bahl'83], can also be given a phonetic interpretation (see section 7.2.3). The acoustic segments can then be assigned broad phonetic class labels, or the phoneme segment can be split into acoustic subwords, as proposed e.g. in [Barry'90b]. In the latter case, e.g. a plosive segment is divided into contiguous segments denoting closure (voiced or voiceless), silence, burst, and aspiration.

For some applications the number of acoustic segments per time unit is selected to be approximately equal the expected average number of phonemes per time unit, i.e. phoneme sized acoustic subwords are detected. For instance when segment boundaries are placed in the middle of acoustically stationary areas, the automatic segments may approximate diphone-like units [Roucos'82].

There are many advantages of segmenting speech into acoustic subwords. Firstly, the speech segments are characterized by acoustic, language independent properties, which can be derived automatically. That is, the calculations are entirely based on signal processing and hence there is no need for explicit modelling or any prior phonological knowledge of the language. Secondly, the automatic subword generation is deterministic in that identical waveforms will be segmented into the same acoustic subword. Thirdly, the acoustic segments often contain highly correlated frames and can hence be quantised, i.e. represented by less data, without losing essential information. This property can be utilized in the preprocessing stage of ASR to compress the amount of data before the recognition step.

Experiments with acoustic subword based ASR, e.g. [Wilpon'87],[Soong'89],[Svendsen'89], gave better results than word based ASR even on a small vocabulary with the words uttered in isolation. This encourages the use of acoustic subwords as basic units in continuous speech recognition. Approximately 250 acoustic units are needed for representing English for an ASR task [Rabiner'89b], (this number depends on the segmentation strategy applied).

Satisfactory vocoding results in very-low-bit-rate speech compression systems based on acoustic subwords are also reported, e.g. [Roucos'82], [Feng'91], [Honda'92]. When the speech signal is segmented into stationary sections, it is possible to optimize the coding parameters in each zone, and thus improving the quality and reducing the transmission bit rate.

However, it is difficult to associate the acoustic segments with phonemic labels in order to design a lexicon for the matching stage of ASR, because there is no one-to-one correspondence to any phonological subword unit. It is therefore necessary to generate the word lexicon based on a sequence of acoustic subwords, e.g. [C.-H.Lee'88], [Svendsen'89], [Soong'88], [Soong'89]. For TTS, a direct transformation from text to speech without any intermediate phonetic or phonemic transcription may be difficult.

3.3.2 PHONEMIC SUBWORDS

Below some frequently used *phonemically* defined subwords are presented and discussed with respect to some of the requirements listed in section 3.2. and how they are applied in real applications.

Phonemes

As mentioned in section 3.1., a **phoneme** can be defined as *the smallest linguistic unit that has distinctive function in a language.* That is, the phoneme is an abstract unit defined for a given language. By contrast, a **phone** is a real, unique speech sound realisation, and an **allophone** is a class of phones which have the same information-bearing parameters, or distinctive features, either perceptually or acoustically within a given language. One phoneme has one main allophone

which is least sensitive to contexts, and various context dependent biallophones. For instance in American English, /t/ is most often realised with aspiration in word initial position, and as a voiced flap between vowels.

To label a section of the speech signal with a phoneme symbol can be regarded as a *sound quantisation.* All the phones that do not make meaningful distinctions among themselves within the language are quantised to the same class, i.e. to the same phoneme, even when they are acoustically quite different.

Since the phoneme is an abstract subword unit it is difficult to identify the phoneme and its boundaries acoustically, and it is hard to decide which acoustic cues to weight most when identifying a phoneme (e.g. formant transition in vowel preceding a plosive). In addition, some sounds may be allophones of more than one phoneme. This *phonemic overlapping* will imply difficulties for e.g. phoneme based ASR.
For ASR it is important that the subword unit is *economic.* A number of 35 to 50 phonemes within a language imply that statistically models for the phonemes can be sufficiently trained with just a few hundred phonemically balanced sentences. However, one statistical model for all realisations of a phoneme will include a huge variance, i.e. it may overgeneralize, and hence the phoneme is *not consistent* in the sense stated in section 3.2.

Phonemes in e.g. function words are often articulated indistinctly, or even omitted, and are therefore not representative instances of the phonemes. In phoneme based ASR such words may be modelled separately with so-called phoneme-like units [Rabiner'89b].

Since coarticulation effects and junctures can not be represented by phonemes, a TTS-system based on phonemes has to smooth the stored phoneme-segments rather heavily at their boundaries to avoid discontinuous, unintelligible speech. That is, the phoneme is *not suited for concatenation.* However, for automatic generation of diphone libraries used in TTS, speech is often first segmented into phonemes, e.g. [Hemert'87], [O'Shaughnessy'88], [Ottesen'91].

Allophones

Allophones supply information about syllable and/or word boundaries. They represent more directly some complex phonetic variation and coarticulation effects, and hence acoustic-phonetic and coarticulation rules may not be as necessary as for phonemes when applied in ASR and TTS. Compared with phonemes, the allophones will reduce the complexity of the interpolation algorithm in TTS. In addition, many allophones are more easy to characterize and identify acoustically, (see e.g. figures 4.4 to 4.8 with different /r/ allophones in Norwegian).

The fact that the phoneme is realized as different allophones in different syllable and stress contexts is by many ASR-designers regarded as noise, whereas others have exploited this fact as a constraint and a knowledge source to improve their ASR system. Church, [Church'87], for instance, describes a parsing algorithm used in continuous speech recognition which takes a string or a tree of segments labelled with phonetic features as input, and uses the allophonic information to decide syllable boundaries and thereby makes it easier to match the lexicon.

However, it is difficult to identify the allophonic cues automatically, and for ASR of English maybe as many as 1000 different allophones are needed [O'Shaughnessy'88].

Diphones

Since transitions between phonemes are disregarded in phoneme models, one possible solution is to model these transitional regions explicitly, for instance by diphone models. A **diphone** is *a segment of speech that contains the last part of one phoneme, the transition, and the first part of the next phoneme*[4]. That is, the diphone includes some coarticulation rule information and the transitional information which is necessary for many sound identifications (e.g. information cues for the consonant are often given by the formant transition in the surrounding vowels).

However, the inventory is relatively large; N phonemes gives theoretically N^2 diphones. It is hence more problematic to get enough training material for ASR purposes. The number of diphones is 1444 for English [O'Shaughnessy'88], but e.g. in a practical TTS system the number can be reduced to about 1200 by excluding non-occurring and rare diphones and exploiting symmetric spectral patterns such as in "-is-" and "-si-".
Another problem is that most phonological rules, as currently written, are not easily applied to diphones. (It is a open question how difficult it would be to rewrite all phonological rules with diphonic symbols and concepts).

Triphones

The triphone can be defined as a natural extension of the diphone definition above: *The speech signal corresponding to two subsequent diphones, i.e. from the last part of one phoneme, the whole following phoneme, and the first part of the third subsequent phoneme.* This unit is very rarely applied in practical systems due to memory limitations: N phonemes gives theoretically N^3 triphones, i.e. 50 phonemes implies 125000 triphones. For a practical ASR system for English this number can be reduced to approximately 10000 triphones [Rabiner'89b].

Syllables

The **syllable** may be defined as *a phonological, structural unit, describing the characteristic combination of vowel+consonant (VC) in a language giving combinations as e.g. VC, CV, CVC or V only or C only.* However, the syllable is "a unit of speech for which there is no satisfactory definition" [Ladefoged'82, p.285]; i.e. the concept should be more rigidly defined.

We often think of a syllable as a vowel or syllabic nucleus and its functionally related neighbouring consonants, see figure 3.4. As discussed in chapter 1, it is often easy to agree on the number of syllables, and to locate the core. A more difficult problem is to agree on the border between the syllables. As far as segmentation and labelling is concerned, the syllable

[4] Where to place the diphone boundary depends on the given diphone type and the intended use. Some TTS systems cut in the middle of all phonemes except for plosives where the boundary is placed just before the release. Others mark the start of the diphone after 40% of the duration of the first phoneme and the diphone end at 60% duration of the last phoneme e.g. [Ottesen'91], or they use special rules such as e.g. placing the boundary 50 ms after the start of the vowel in consonant-vowel diphones [O'Shaughnessy'88].

suffers from the same problem as all the other phonologically defined units, namely that of no direct correspondence to the acoustic speech signal. In addition, the syllable is not well defined phonologically either [Ladefoged'82], [Sivertsen'88].

Phonologically, the ambiguities are tried solved by using phonotactic rules in the language or by defining constraints. One definition could be (1) *the principle of maximal open syllable*; i.e. the syllable should end with a vowel. If this is not possible to achieve due to phonotactic constraints, then use (2) *the principle of minimal coda*, or equally, *maximal onset*[5]. For instance, according to (1) the English word "employ" should be divided into "e+mploy". But since "mp" is not allowed to begin a syllable in English due to phonotactics, we get "em+ploy".

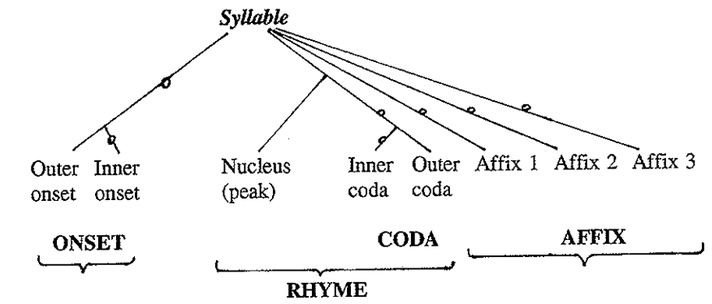


Figure 3.4 *Division of one-syllable words into typically chosen constituents of the syllable and demi-syllable (this is a combination and modification of the figures in [Zue'89, p. H-102] and [Ruske'85, p.596]). Branches marked by o are optional. The first two example words are English, whereas the last one is Norwegian showing that /sp/ is regarded as only one obstruent cluster.*

[5] Some researchers, e.g. [Kahn'86], avoid some of the problems by using the expression "ambisyllablicity" for a consonant belonging simultaneously to two syllables, as /n/ in the English word "pony". However, it seems as if all syllable definitions have to allow both exceptions and ambiguities (see also [Schotola'84]).

One benefit with using syllable units is that currently used phonological rules, or sound-change rules, for e.g. stress and length, mainly depend on the syllable structure [Zue'89]. In other words, the syllable gives compact and explicit formulation of many phoneme realization rules. It is also assumed that most of the contextual variations of phonemes are due to the influence of other phonemes within the same syllable. Hence, coarticulation effects are captured in the syllabic unit, but both ends are still affected by the context, and often it is difficult to locate the precise boundaries between syllables.

The syllable is employed as the basic unit in some ASR systems because it combines the best of phoneme based ASR, such as small symbol inventory[6] and the advantage of the whole-word template-matching method with no need for explicit representation of many contextual effects. In addition, the nucleus of the syllable can be automatically located using for example energy-contour, temporal loudness contour, or spectral information [Weigel'89]. In a phoneme based ASR system, the syllable unit may restrict the possible phoneme sequences, and can serve to control the choice of potential allophones [Gupta'86], [Church'87].

"Pseudo-syllables" or *Phonetically defined syllables.*

A number of phonetic definitions of the syllable have been discussed (e.g. [Sivertsen'88], [Ladefoged'82]). One articulatorily based definition is based on the chest pulse, which is a contraction of the muscles of the rib cage that expels air from the lungs. This measure is difficult to use in a practical automatic segmentation or ASR system. Others have tried to utilize the fact that the mouth is most open for vocoids, and since these mostly constitute the syllable nuclei, the degree of mouth opening could indicate syllables. However, due to general measurement problems, speaker variability, syllabic consonants etc., it has been found difficult to incorporate the mouth opening criterion in a practical annotation or ASR system.

Acoustically defined syllables can be found by locating extremities in the intensity or loudness of the speech signal (e.g [Colla'89b], [Fournol'89], [Weigel'89]). Actually, most of the words in e.g. English are pronounced with intensity peaks corresponding to syllable nuclei. But there are some exceptions; as for example the word "hilly" /hili/, which is realised with only one intensity top but contains two (phonological) syllables [Ladefoged'82].

Auditory definitions of the syllable are based on the sonority. This phonetic definition is based on how we perceive sounds, and can therefore also be language dependent. As for the articulatory definitions it is difficult to measure sonority, but roughly we can rank the phoneme groups by increasing sonority as shown in e.g. [Church'87], [Ladefoged'82] and [Zue'89]:

stops < *fricatives* < *nasals* < *liquids* < *glides* < *vowels*

The sonority constraint on the syllable structure is that the sonority must decrease from the nucleus towards the margins; i.e. the number of syllables in an utterance equals the number of

[6] In many languages there is a very large inventory of syllables, e.g. more than 20000 can be found in English, but only about 4400 are sufficient to describe nearly all English words [O'Shaughnessy'88]). Other languages such as Japanese consists of only 101 different syllables, and are thus well suited for syllable based ASR.

sonority peaks and the syllable boundaries are always placed at minima in sonority[7].

Demi-syllables (half-syllables)

Despite the problems regarding the syllable, many researchers have used demi-syllables as a part of their ASR system. This unit retains many of the good properties of the syllable, but with a reduced inventory of symbols[8].

A **demi-syllable** is the part of the syllable extending from syllable boundary to syllable nucleus, *final demi-syllable*, and the part extending from nucleus to next syllable boundary, *initial demi-syllable* [Ruske'85]. The nucleus is split in the middle, i.e. one boundary is placed in the vowel; which is the point where the effects of coarticulation are reduced and articulatory movements are minimal (as opposed to the rapid transitions at many phoneme boundaries). The demi-syllable includes coarticulation effects, and at least one border lies in a "stable" area; i.e. the demi-syllables are suitable for concatenation. According to this definition, many demisyllables are diphones.

The ASR training problem can be reduced by splitting up the demisyllables into their vowel and consonant clusters [Ruske'84], [Schotola'84]. In this way the huge inventory of different demisyllables can be avoided while preserving the advantages of demisyllable segmentation. This means that the *demisyllable can be regarded as a segmentation and processing unit but not as a decision unit for recognition.*

3.4 OTHER UNITS

Phonetic features

A **phonetic feature** is *a phonetic property that can be used to classify sounds in a language.* The phonetic features are thus not subwords, but rather a set of labels which describe segments of the speech.

Each feature has an acoustic correlate. This correlate, or property, is assumed to give rise to a pattern of response in the auditory system that is qualitatively different or distinct from the corresponding response pattern associated with other features. For instance, in order to differentiate the phonemes in English, Ladefoged [Ladefoged'82] proposed the following

[7] The model fits in many cases, but there are exceptions. E.g. the strong fricative /s/ does not match this hierarchy at all, as shown in the words "its" and "stress" (one phonological syllable, but two peaks of sonority). Another problem with the sonority is that the loudness of some sounds may vary from person to person.

[8] The inventory size can be reduced by about a factor of 5 from that of syllables [Rosenberg'81]. Still there is a large number of units: In English there are about 100 initial consonant clusters, 15 vowels /diphthongs and 200 final consonant clusters, which gives a total of 4500 demisyllables [O'Shaughnessy'88]. It is argued that a very large English vocabulary can be produced from fewer than 2000 demisyllables [O'Shaughnessy'88]. The number of syllables can still be reduced by a factor of 2 if the apical consonants /s,z,t,d,T/ are treated independently [Rosenberg'83], i.e. a combination of demi-syllables and phonemes. For instance the English word "tax" should be divided into /tA/+/Ak/+/s/.

features: Voice, Place, Stop, Nasal, Lateral, Sibilant, Height, Back and Syllabic. He also claimed that a speech segment is either voiced or voiceless, thus Voice is a binary feature. The feature Place on the other hand is multivalued in that it has to distinguish between labial, dental, alveolar, palatal and velar, places of articulation. All features are articulatorily described except Sibilant which represents the amount of high-frequency noise in a sound. Labelling speech segments using these features is enough to classify the phonemes, but are too rough for a narrow phonetic transcription.
When features are used in ASR or automatic speech segmentation and labelling, as e.g. in [Dalsgaard'89], all the multivalued features are divided into binary features as in [Chomsky'68], and they have to be measurable from the acoustic speech signal.

There are many advantages connected with using features as annotation units. Firstly, only 20 features are needed to perform the meaning differentiating function in any language. Secondly, the representation has usually more features than the minimum number that are needed to distinguish the utterance from possible competitors, i.e. there is redundancy so that there is room for variability in the acoustic representation of the utterance. Thirdly, the features have articulatory correlates as well as acoustic or perceptual ones (each lexical item is assumed to be represented in the mind of a speaker/listener in terms of patterns of features). Through proper selection of properties that describe spectral relationships, these properties can be speaker independent since they do not depend on the speaker's vocal tract or average fundamental frequency. Finally, words that have different meanings have a different representation in terms of binary features, except for homonyms which have the same expression but different meaning.

A practical problem with the features is that the acoustic properties that qualify as correlates of phonetic features tend to be *relational* and not absolute. Another problem is that the features overlap in time. Additional language specific rules are thus needed to select one phonemic label for a given segment.

Words

In languages as Norwegian and English the word unit corresponds to a group of letters without interval in the orthography, but in e.g. French "du beurre" is regarded as one phonological word [Grice'90]. The definition of the word unit is thus language dependent.

The word is the smallest unit that contains phonological, semantic, as well as syntactic information. Used in ASR it eliminates an entire level of recognition activity, in that we do not have to concatenate the recognised units in order to look up in the lexicon; that is, this unit is exactly what we want to recognize. In addition, word models are able to capture within-word contextual effects, not only the immediate phonetic contexts as in triphones and syllables.

Using words as ASR units works very well for recognizing small vocabularies of words spoken in isolation, e.g. [Pan'85], [Rabiner'79], [Rabiner'81]. Also in connected word (digit) recognition, words as the basic unit provide good results, e.g. [Rabiner'80, 81, 88], [Myers'80]. When there is sufficient training data, word models generally achieve the best ASR performance.

However, there are a substantial number of words, e.g. over 300000 in English where about 50000 can be considered common [O'Shaughnessy'88], and since word models lack generality in that each word needs its own model, there is no possibility of sharing data in the training

process. For big vocabularies we need a vast amount of data because each word should appear several times in each phonetic context. When word models are used in ASR, the users have to produce many repetitions of each new word that is added to the vocabulary. This is extremely time consuming and inconvenient. In continuous speech there is no pause between the words so that this unit must also be segmented.

Words can only be used in TTS for small vocabularies, for instance in restricted task applications, due to the memory requirement and the problem of adding new words.

Context-dependent phonemes

In order to develop an effective ASR device whose basic recognition unit is the phoneme, we first have to establish the large set of relevant context-sensitive rules, making them explicit, and encoding them in the recognizer. Alternatively, the phoneme can be modelled as triphones[9] where single phonemes are modelled conditioned on the immediate left and right neighbouring phoneme. This takes into account the most important coarticulatory effects, and is therefore much more consistent and detailed than the "context free" phoneme models. Since triphone models are specific phoneme models, they can be interpolated with better trained but less appropriate context-independent phoneme models or models with only right context or only left context, as e.g. in [Lee'88]. So-called generalized triphones which exploit that many phonetic contexts are very similar by clustering the triphones before training the models, are also suggested [Lee'88].

Even though coarticulatory effects are most pronounced for immediately adjacent sounds, they may also influence the sound realisation several phonemes away from the observed phoneme. The context dependency can thus be modelled as word-dependent phonemes, where each phoneme within a word is modeled as depending on the word in which it occurs [Chow'86]. Word-dependent models will alleviate the drawbacks mentioned for context-independent phonemes in e.g. indistinctly articulated function words. In this way many models for each phoneme are obtained. This can be regarded as allophonic representations. The drawback is that many realizations of each word are needed in order to make robust models.

Combined units

Since no single subword fulfils all requirements, either a multi-level annotation or a hybrid approach based on several subword units and/or words should be employed, and thereby utilizing the best properties from all of them. For instance, in [Cravero'86], it is claimed that the optimal subword unit for ASR is a combination of phonemes and diphones. Hence, when the transition between two sounds is considered significant for the recognition of two sounds themselves (e.g. a plosive followed by a sonorant), the corresponding diphone is included in the set, otherwise the transition model is realized by appending the two phoneme models. The use of single phonemes and the elimination of unnecessary diphones reduces the unit inventory to a size for which good estimates of model parameters can be obtained with acceptable size training sets (this approach gave an optimal set of 22 stationary and 101 transitional elements).

---

[9] A triphone is here interpreted as a phoneme in a given fixed phoneme context (see "Triphones" in section 3.3.2 and in the Glossary).

Another step in this direction is the diphone-like segments defined e.g for Italian [Colla'89 b] as (i) "steady-state" sounds, such as vowels and nasals, (ii) transition between two sounds, such as between consonant and vowel and between consonant and sonorant, and finally (iii) "triphone" transitions, such as consonant-glide-vowel. This can probably be optimized further, e.g. by labelling the function words with single symbols.
One may also combine demi-syllables and phonemes, or, rather than making an a priori choice of the decision units, the units should be determined automatically by using training methods according to the nature of the phonemes and their contexts.

Prosodic units

The prosody accounts for much of the variability in the speech signals and conveys some of the information necessary for recovering the intended meaning of an utterance. Prosodic units are at least the length of a syllable, and are used to express syntactic structures, to carry speakers attitude, and to organize the structure of discourse. Typical units for describing *speech melody* are tones and intonation, and units describing *speech dynamics* can be duration, stress and rhythm. Annotation of bigger prosodic units such as stress groups, metric feet, and intonation groups are often used for e.g. improving the quality of TTS systems.

3.5 SUMMARY

In this chapter different levels of speech annotation and some frequently used annotation units have been discussed (other units such as dyads, CVC-words, logatomes, and morphemes are also used for speech research and speech technology, but are not discussed in this thesis). The main problem is that the acoustic subwords have no direct mapping to the phonemically defined units, whereas phonemic subwords provide no direct mapping to the acoustic signal. All the units have their advantages and disadvantages, and *none of them fulfil all requirements* listed in section 3.2.

The properties of the phonemic subword units discussed in this chapter can be roughly summarized as in table 3.1:

Size	**Advantages**	**Disadvantages**
Large (i.e. larger than a phoneme)	Comprehensive Consistent Less context dependent, i.e. a token is more representative	Large inventory
Small (i.e. phoneme size or less)	Economic Larger units can be derived	Context dependent realisation, i.e. inconsistent

Table 3.1 *Some properties of the phonemic subwords discussed in this chapter.*

In general, units larger than a phoneme will contain more contextual and transitional phenomena than smaller units. In this respect larger units will provide a more accurate representation of the speech. It is claimed [Fournol'89] that it is easier to locate demi-syllables than e.g. phonemes with automatic segmentation techniques, because it is more difficult to "tune" the procedures for smaller units. However, increasing the unit size also increases the number, and since the annotated training databases still are limited there will be fewer tokens of each unit. A small subword inventory provides a manageable set of templates and we get many realisations of each subword to e.g. train statistically based subword models and hence robust models are obtained.

As regards concatenation quality, which is especially important to the TTS systems, diphones and demi-syllables are the best units because the segmentation and thereby the concatenation is performed in spectrally stable portions of the speech signal. The number of these units is also manageable for the TTS purpose.

In this thesis the **phoneme** is selected as the basic subword unit for speech annotation. The pros and cons of this unit are discussed in section 3.1 and 3.3.2, where the facts that there are few phonemes (*economic*) and that they can be used as building blocks for any of the bigger subwords, are considered the most important ones. Actually, the phoneme subword annotation is *flexible* in that bigger units such as diphones, triphones, demi-syllables, syllables, and words can all be derived automatically from a phoneme-sequence. By means of phonetic and phonological rules the phonemes can be divided into smaller segments, e.g. a plosive segment divided into a pause and a burst segment. Another important pro is that the phonemes map most directly to the lexicon entries.
The phoneme, i.e. the SAMPA symbol set, was also selected within the SAM-project because the intended use of the annotated continuous speech database was "analytic work in speech recognition and synthesis assessment" [Erp'89a, p.305].

Chapter 4

MANUAL SEGMENTATION AND LABELLING OF SPEECH

The ultimate goal for this thesis work is to develop an automatic segmentation algorithm for segmentation and labelling of a multi-lingual speech database. However, in order to train and evaluate automatic procedures, manually annotated speech corpora are needed as reference sets. It is thus essential that the manual segmentation and labelling is done as consistently and repeatable as possible, that the conventions used are explicitly known, and that shortcomings and inconsistencies in manual annotation are highlighted.

Ideally the annotation strategy should be standardised. However, no single strategy for manual segmentation and labelling has been standardised yet. A starting point for a standardisation is then to make explicit and compare existing conventions for different languages, clarifying what they have in common and where they differ.

In this chapter a multi-lingual speech corpora is investigated with respect to recording protocols and annotation conventions. One possible approach to broad phonetic annotation of a continuous passage is then exemplified by a detailed description of the manual segmentation and labelling of Norwegian.

4.1 TOWARDS ONE COMMON ANNOTATION STRATEGY?

Although almost every speech researcher has performed some manual annotation of speech, not many detailed annotation strategies are published. Also very little effort has been spent on examining the reliability of manual annotation criteria between labellers within languages and across languages. This is strange because the manually annotated speech material is used e.g. to measure the correctness of automatic speech segmentation (ASS) and automatic speech recognition (ASR) procedures.

The motivation for comparing annotation conventions of different languages was threefold:

1) To find which annotation criteria are common and which differ in order to develop a common multi-lingual speech database specification. Since no standard set of segmentation and labelling conventions exists and everybody makes their own set of criteria, the first step in developing a common annotation approach is to make approaches explicit and compare existing ones.

2) To find, on a phonetic basis, phonemes that represent approximately the same sounds in

different languages in order to increase the training material for the statistic models in the ASS algorithm. That is, we want to examine the correlation between the phonemic labels and the real acoustic signals that they represent in different languages.

3) To check if the broad phonetic segmentation and labelling is performed in such a manner that cross-comparison between languages at the same level of annotation is possible. That is, is it reasonable to regard the manual annotation as reference material?

In the following two subsections the speech material selected for investigation in this thesis is described and then different annotation conventions proposed for typical phoneme transitions are compared.

4.1.1 EUROM0 AND SAMPA

The aim of the ESPRIT-SAM project [Fourcin'90] was to define and develop applications of multilingual standards for assessment, evaluation, and cross-comparison of speech technology products such as automatic speech recognition systems and speech synthesis systems.
Thus, within the SAM project the multilingual EUROM0 [Grice'89] and EUROM1 [Sherwood'92] speech databases have been compiled by using common hardware [Lindberg'89] and software for speech recording e.g. EUROPEC [Zeilinger'91], and speech analysis, PTS [Caerou'91]. The software packages and the computer readable SAM Phonetic Alphabet, SAMPA [Wells'92], were developed within the SAM project and are now increasingly used as *de facto* standards in Europe.

For EUROM0, recordings of Danish, Dutch, English, French and Italian were compiled on a CD-ROM. The different languages were recorded in different sites using common recording protocols. However, the recordings got different levels of background noise because some sites recorded in anechoic chambers whereas others recorded in quiet office rooms. This may have influenced the succeeding annotations of some recordings. E.g. the high-frequency noise in the Danish EUROM0 recording has been claimed to be difficult to distinguish from word final friction [Grice'89, p.43].
For each language single digits, digit triples and a continuous passage were recorded. The continuous passage was read by two men and two women for each language. The continuous passage was a meaningful story read aloud, i.e. it contained more than one sentence and the sentences were related, see also chapter 2. The Norwegian EUROM0 text is shown in Appendix B.3 and the other EUROM0 texts are listed in [Grice'89, pp. 9-14].

The corresponding Norwegian and Swedish continuous passage recordings have also been compiled, but these are not stored on a CD-ROM[1].
For the investigations in this thesis work the continuous passage recordings of English, Italian, Danish, Norwegian and Swedish were selected. These recordings will be referred to as the respective languages' *EUROM0 recordings.*

One or more native phonetician(s) from each language segmented and labelled the continuous

[1] The Norwegian and Swedish continuous passage recordings are available in the same format as the other recordings from National Physics Laboratory, United Kingdom.

passages at a broad phonetic level with the computer readable SAMPA symbols by access to speech analysis facilities such as the speech pressure waveform, spectrum and spectrogram in addition to auditory information[2]. (For English, spectral information was not used [Erp'89a]).

In addition to the annotation problems discussed in section 1.3.2, there are at least two problems that have to be dealt with in order to achieve the goal of comparing languages at a common annotation level. Firstly, SAMPA is phonemic and thus used according to the analysis of distinctive sound oppositions within each language. In Appendix A the relationship between the IPA and SAMPA symbols is given, *but the IPA symbols represent phoneticly defined sounds which are language-independent, whereas SAMPA is used for language-dependent sound categorisations.* The acoustic realisations covered by one SAMPA symbol will thus vary between languages, and within one language it will cover all allophones of that phoneme.
Secondly, if the annotations of EUROM0 were to be performed at the same annotation level, not only the symbol set but also a common segmentation strategy should have been agreed upon. For instance the Dutch only indicated the centre of the phonemes, whereas the others segmented according to the definitions used in this thesis.

4.1.2 COMPARISON OF ANNOTATION CONVENTIONS

Some annotation strategies have been proposed, e.g. for EUROM0: English [Barry'90 b], Swedish [Nord'90], Italian [Cosi'90], Norwegian [Kvale'91], Danish [Dyhr'92] and for TIMIT: American English dialects [Seneff'88]. Since no general guidelines exist for manual segmentation and labelling, everyone makes his or her own conventions. These conventions are based on what is felt natural according to the phonetic school of the labeller and what is the intended use of the annotated speech material. If the conventions are strictly adhered to, all the segmentation and labelling is correct.

The Swedish, Norwegian, and English EUROM0 annotations were performed at a level fairly similar to the *broad phonetic* labelling as described in chapter 3. Each phoneme in the perceived phoneme sequence was segmented whether acoustic cues were seen in the waveform or not. However, if acoustic cues were seen the phoneme boundaries were placed at these. The annotation of the Danish EUROM0 recordings on the other hand, was performed at some *acoustic-phonetic level* where an abrupt change in the signal was marked whether it indicated a phoneme boundary or not. *In manual cross-comparison tests and in the assessment of ASS algorithms these differences in strategy have to be taken into account.* The acoustic-phonetic approach will presumably give much better correspondence to automatically placed boundaries than the broad phonetic approach.

Across languages some types of sound transitions, and thereby also segmentation problems, are common and should hence be treated uniformly. Below some **typical annotation problems** in the EUROM0 material are listed together with the suggested solutions for placement of segment-boundaries and choice of labels (language code in parentheses). In addition, some examples from the annotation of the TIMIT database [Seneff'88] are provided. Since not all annotation strategies were well documented, some of the examples are based on a limited investigation of the foreign

[2] Analysis of the Norwegian EUROM0 is given in Appendix B, and the annotations of the Danish, Swedish and English EUROM0 are discussed in Appendix C.

speech material. For some languages and some labellers there were phoneme transitions where no single segmentation rule was followed consistently.

**Voiced sound initialises a sentence*, e.g. /.../-/i/, the boundary is placed:
-at the first positive zero-crossing of vocalic voice period (Eng)
-at the zero-crossing preceding a strong positive going peak (Swe)
-at the first amplitude (bigger than the background noise) seen in the waveform (Nor).

**Voiced-voiceless transition*, e.g. /i/-/s/:
-all friction is included in the fricative (Eng),(Dan)[3],(Ita)
-all voicing is included in the vowel (Nor)
-symmetrical; some voicing is included in the fricative and some friction is included in the vowel (Swe).

**Voiceless sound followed by silence*, e.g. /s/-/p/, the boundary is marked:
-at the last positive amplitude (bigger than the background noise) seen in the waveform (Nor), (Eng)
-at the abrupt change in the intensity in the spectrogram.

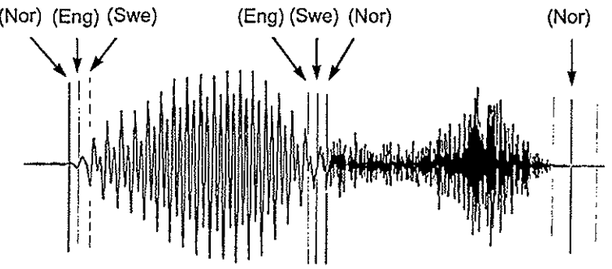


Figure 4.1 *The waveform of /... i s / excerpted from /...isprOke/ (=in the language) uttered by a Norwegian male speaker, exemplifying the three transitions described above. The end boundaries marked with Nor are 200ms apart. At the /s/-p/ transition an uncertainty area is indicated (see section 4.2.1).*

**Plosive initialises an utterance*, e.g. /.../-/p/, the start of the plosive is marked:
-50 ms before burst begins (Nor),(Eng)
-where burst begins (Dan),(Swe)
-300 ms before the end boundary of the plosive (Ita).

[3] The Danish convention was that a fricative starts with the onset of friction, "even if voicing is still present" [Dyhr'92]. In practice this principle is not consistently followed (see e.g. the first sentences of speaker BLD).

**Two immediately succeeding plosives with no audible release in the first one*:
-the first plosive includes the small variation seen in the waveform before a proper closure is established, the second plosive denotes the silent closure and the burst (Eng)
-the boundary is placed rather randomly in the silence area (Dan)
-the first plosive denotes the closure, i.e. the silence segment, and the second plosive denotes the burst and aspiration (TIMIT).
-not a problem because both plosives are released (Nor),(Swe).

**Extralinguistic sounds* e.g. lipsmacks, breath, and stomach rumbling:
-segmented and labelled with the silent pause symbol (Nor),(Swe)
(consecutive silence segments are merged (Nor)).
-some types of extralinguistic sounds are explicitly segmented and labelled (Eng).

**Epenthetic silence* which are not perceived by listening but found by visual inspection of the waveform (appears most often when a voiceless sound precedes a voiced sound):
-if the silence portion is longer than 15 ms it was marked as silence (Nor),
otherwise; the silence portion is included in the voiceless sound (Nor),(Swe),(Eng),(Ita)
-segmented and labelled with an own symbol, i.e. not the normal silence symbol (TIMIT).

**Epenthetic sounds*, which are not perceived in context but found by careful inspection:
-included in the neighbouring phoneme which has most features in common with the epenthetic sound (Nor), (Swe), (Ita).

**Geminate phonemes*:
-merged into one segment if no boundary cues are seen in the waveform or spectrogram (Nor)
-merged into one segment (TIMIT), (Swe).

**Glottalisation[4] / Creaky voice*:
1) Between vowels:
-symmetric approach, i.e. place boundary in the middle of the creaky area (Nor)
2) Between vowel and plosive:
-include all the creaky voice in the vowel (Nor)
-include all the creaky voice in the plosive (Eng).

**Auditorily based decisions* (when no visual acoustic landmarks are found):
-a documented careful listening technique is applied (Nor),(Ita).

**Diphthongs* (which are phonemes):
-the whole segment is marked with a diphthong symbol (Nor),(Swe),(Eng)
-the weak vowel is assigned almost $^1/_3$ of the whole diphthong duration (Ita).

[4] In English there are two types glottalisation: (i) Glottal reinforcement where a glottal stop precedes a normally definable closure and release, and (ii) Glottal replacement of the stop category, e.g. intervocalic /t/.

**Marking uncertainty in the annotation*
(i.e. avoids the drawbacks of forcing every decision to be an all-or-nothing choice):
-no such marking (Nor),(Eng),(Ita),(Swe)
-the segment is marked with two labels and an asterisk, e.g. /f*R/ (Dan)
(i.e. the boundary between /f/ and /R/ is not marked at all).

**Devoicing of voiced phonemes:*
-the symbol for the voiced phoneme is kept (Nor), (Eng), (Dan?)
-partly devoiced: the voiceless part is included in the voiceless neighbouring sound (Swe), (TIMIT)
-totally devoiced: new symbol for the perceived sound (TIMIT)

**Phonemic assimilation where the voicing feature is not the only one changed:*
-the new perceived phoneme is labelled (Nor),(Swe),(Eng),(Dan),(TIMIT).

The two last examples show that different voicing realisations of a phoneme do not alter the phonemic labelling, whereas e.g. "can be" is labelled /k{**m**bI/. That is, if only the voicing features differ from the expected phoneme realisation the *intended* phoneme is labelled. For all other phonemic assimilation the *actual* speech signal is labelled.
It may also seem inconsistent that a plosive is restricted to the closure and the burst, although the formant-transitions in the neighbouring vowels are important cues for the perception of the plosives, whereas for consonants with a constriction phase, such as /v/ and /j/, the formant transitions are often included in the consonant segment.

To **conclude**, it is difficult to lay down detailed guidelines for the annotation of speech, due to the great variability and intrinsic complexity of the speech signal e.g. different speakers, with differences regarding phonation, articulation and dialect, speaking styles and languages. For instance in the case of different phonation types, a breathy voice may cause too much frication for a fixed, detailed rule for the voiced to unvoiced sound transition.
Most of the phonemes are realised differently in different languages and dialects and often have allophonic contextual variants within one and the same language. Therefore, conventions have to be made to also include these variants (see e.g. discussion of the Norwegian r-sound in section 4.2.2.).

However, the comparison of annotation approaches in this section may serve as a starting point for the development of a standard procedure for manual multi-lingual speech annotation.

4.2 ANNOTATION OF THE NORWEGIAN EUROM0 RECORDINGS[5]

4.2.1 GENERAL APPROACH

For Norwegian, two males; AFN and TGN, and two females; SHN and TBN were recorded[6] in an anechoic chamber. All the four informants come from the South-Eastern part of Norway, and all read the passage in standard South-Eastern Norwegian. They are researchers at the University of Trondheim, between 30 and 56 years old. Three of the speakers were used to the anechoic chamber and the recording situation; one of them, AFN, is a phonetician[7].

We[8] used the analysis program WAVES on a SUN workstation, where the waveform and a broadband spectrogram were displayed. Our point of departure was to *annotate what we heard and saw with the predefined phonemic labels in Appendix A and mark the endpoints for each phoneme*. Thus, we only employed the diacritic sign for length since length has a phonemic value in Norwegian. Stress and tone were not marked in the manual annotation.

We knew the text, had the written version in front of us, and listened to 3-5 words at a time. We then performed the segmentation on the basis of visual inspection of the zoomed waveform and the spectrogram and on *careful listening* (see further details below).

To develop an annotation strategy, we first segmented and labelled a speech material similar to the Norwegian EUROM0 recording. For each phoneme to phoneme transition our goal was to find reliable acoustic cues in the waveform and spectrogram for marking the boundary. We then noted the cues and how we applied them. Next time the same phoneme pair occurred we tried to apply the same cues and strategy. According to the discussion in the preceding chapters, we cannot claim that the selected cues and strategies are the correct ones, but by making them explicit and by applying them consistently, it will at least be easier to correct the procedure when new insight is acquired. For example when segmenting a plosive surrounded by vowels we know that one of the most important perceptual clues is the formant transition within the vowels. But still we did not include this formant transition in the plosive, as can be seen e.g. in the /i/-/d/ transition in figure 3.2.

In order to make the multilingual annotated speech database more useful, the manual segmentation strategy and the symbols employed should be well documented. As far as labels are concerned, the consonant symbols are mostly selected according to the IPA scheme of classification, but the phonetic content of the vowels varied a lot among dialects and languages

---

[5] Parts of this work are also published in [Kvale'91].

[6] We simultaneously recorded the corresponding laryngograph signal. This equipment may affect the larynx activity slightly. The laryngograph signal is not utilized in this thesis work.

[7] The speech waveforms in figure 1.1, 1.2, 1.6, 3.2 and 4.1 are all excerpted from the recordings of speaker AFN.

[8] I performed the manual annotation of the Norwegian EUROM0 recording together with the Norwegian phonetician Arne Kjell Foldvik.

for the languages investigated. It is therefore appropriate to indicate the typical phonetic content for each dialect by plotting the vowel symbols in the cardinal vowel trapeze, as shown for speaker AFN in figure 3.1.

For some phoneme-to-phoneme transitions there were no clear visual acoustic cues available for marking boundaries, and it is tempting to mark such a sound pair with one symbol only. E.g. for annotation of Danish, it is suggested that all /j/+vowel segments should be labelled with single diphthong symbols [Dyhr'92].
However, if neither the spectrogram nor the waveform indicated the endpoint for a phoneme, we used a **careful listening** technique where we iteratively listened to varying portions of the speech signal, as illustrated in figure 4.2 below.

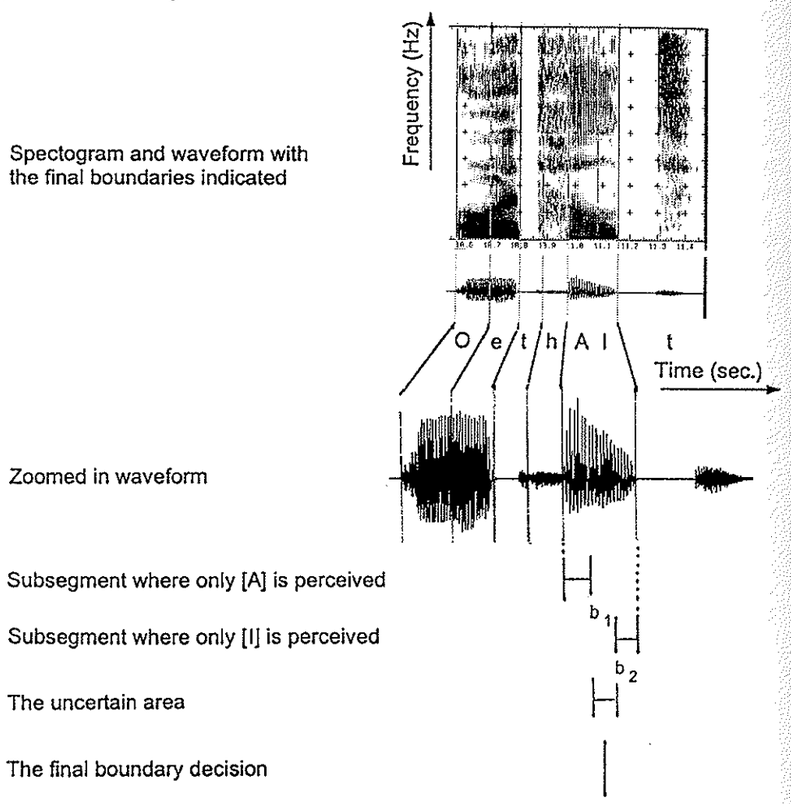


Figure 4.2 *The careful listening technique.*
At the top a spectrogram and waveform of "og ett halvt" (=and a half) manually segmented and labelled /OethAlt/. Then the waveform is zoomed in and the careful listening technique is exemplified. The + signs in the spectrogram are separated by 0.1 sec. horizontally and 1000 Hz vertically.

In figure 4.2 a /I/ succeeds an /A/, which is often seen as a smooth transition in the spectrogram and waveform. We then marked the start of /A/ and the end of /I/ following the conventions in section 4.2.3. Then we listened from the beginning of /A/ and successively moved the cursor to the right and listened to the segment. When a slight influence of /I/ could be heard a preliminary boundary, b_1, was marked. Ideally, (from the transcriber's point of view), only /I/ should be heard when listening to the segment from b_1 to the end of /I/, and b_1 would then be the final decision. But due to coarticulation effects we would often hear an /AI/-like sound. A new boundary, b_2, where only /I/ was heard was then established. The segment between b_1 and b_2 is the **uncertain area**. The final segment border was marked in the middle of this area; which would often coincide with maximum formant transitions seen in the spectrogram.

As discussed in chapter 2 we actively reconstruct the spoken message based on the perceived speech signal together with our knowledge about the vocabulary, syntax, dialect, habits of the speaker, context and paralinguistic cues. In the manual annotation task we also utilised other knowledge sources, e.g. in the segmentation of the word "tall" (=number): Because of coarticulation it was difficult to divide the /AI/ segment. Careful listening indicated an uncertain area with a rather long duration. But the phonemic structure of this word is a short /A/ followed by a long /I/ so in this case $^1/_3$ of the /AI/ segment was allotted to /A/ and the rest to /I/.

Very short segments with a rapid change in amplitude at the endpoints will give disturbing auditory impressions. It sounds as if the segment starts with a plosive. Therefore, we may perceive one sound when listening to a whole meaningful sentence, and a different one when listening to short segments containing one or two phonemes. For example, in the word sequence "etterhvert gått" (=gradually gone), the /g/ at the beginning of the second word is perceived as voiced plosive. However, when listening to an isolated part the stop sounds more like a [k]. This illustrates our ability to restore phonemes to make an utterance meaningful to us, as discussed in chapter 2. Since we performed a phonemic labelling we regarded this [k] as an allophone of /g/ and accordingly labelled the segment /g/.
Generally, voicing and frication properties do not always exhibit the theoretically defined features for the phonemes, basically due to synchronisation difficulties between the articulatory system and the larynx sound source. In our phonemic speech annotation such variations were not taken into account as exemplified with the /g/ above.
When friction and voicing were ruled out we applied our convention of labelling a segment with the phoneme which was most auditorily alike, i.e. the phoneme perceived when listening to the segment both in isolation and context. This principle was for instance applied to reduced vowels or if the result of an assimilation process was a new phoneme, e.g. "landbruk" (=agriculture) pronounced more like /lAmbr{:k/ than /lAnbr{:k/.

Our convention was that if we heard a sound in context it should be segmented and labelled, even if we could not see any acoustic cues in the waveform or the spectrogram or could not hear the sound in isolation. The audible, but invisible phoneme was then squeezed in between the two surrounding phonemes by taking a couple of pitch periods from each of them. This is typically what is done for e.g. the /v/ (cf. fricatives in section 4.2.2 and see figure 4.12).

Generally, the waveform is the most accurate source of acoustic cues for marking segment boundaries. Thus information from the waveform was applied whenever possible, as seen e.g. for the transition between voiced and voiceless sounds below. Boundary cues in transitions between voiced phonemes, e.g. between vowels and nasals, are often most easily seen in the spectrogram, but then the time instant when the boundary is placed is not that accurate.

The uncertainty in the segmentation varies a lot. For example, the transition between the vowel and nasal usually appears as a very sharp boundary in the spectrogram; e.g. with a clear difference in intensity as seen in figure 3.2, but the transition can at times be smooth and with no particular acoustic cues displayed. This varies for the same speaker and between speakers and follows no regular pattern.
When the transition between two phonemes can be seen as an abrupt change in some acoustic cue, the uncertainty area is less than 5 ms. However, when we resorted to the careful listening technique the area of uncertainty could amount up to as much as 40 ms. A measure such as mean uncertainty for boundary placement is hence not a very useful one.

Most of the boundaries between phonemes were easy to define in the spectrograms due to abrupt differences in intensity or voicing. The rest, less than 30%, had to be treated more carefully; either by some rule based on phonetic knowledge or by careful listening.

In a pilot experiment we indicated the reliability of the boundary placements as reported for the French EUROM0 in [Erp'89a]; i.e. with 0=absolute unreliable, 1=fairly reliable, and 2=reliable. However, we found it too difficult to be consistent in this indication of reliability.

4.2.2 THE PHONEME CLASSES

In this section a general description of the phoneme classes defined in table A.2 in appendix A is given and some peculiarities for each phoneme class are discussed. Note that although the discussion in this and the next section applies to many realisations there will always be exceptions. Contextual allophonic variants of phonemes are at the end of this section exemplified by the r-sound.

Fricatives: The /j/ and /v/ phonemes are not realised as fricatives in Norwegian, but rather as approximants. The /v/ is easy to identify when it is marked by an abrupt change in the intensity in the spectrogram, as in /ive/ in figure 4.12. But usually /v/ is very short and is identified only by its modification on the formant structure of the neighbouring vowel(s) as in /nvi/ in figure 4.12. Because /v/ is a labiodental it has a lowering effection on the formants of the neighbour vowel(s). The boundary was marked at the steepest point in the formant-transition.

Of the other fricatives, /f, s, S, C/ were realised voiceless, but /h/ was sometimes realised voiced and sometimes voiceless as depicted in figure 4.3. Most often the /h/ is not pronounced at all, as in the word "opphøyd" /up2yd/, (in this word the /h/ elision was unexpected because the /p/ should be released with aspiration, i.e. as [p^h]).

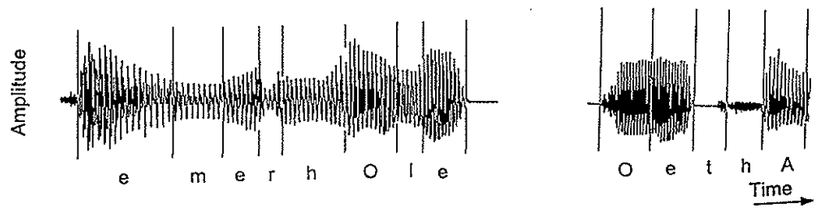


Figure 4.3 *Voiced and voiceless /h/-allophones extracted from speaker AFN.*

Lateral: The /l/ usually becomes voiceless towards the end when a voiceless sound follows it, as in the word "alt" (=everything). Sometimes the first part of /l/ is devoiced as in "slik" (=such) when a voiceless segment precedes the /l/. We included these unvoiced parts in the /l/-segment.
(The transition between voiced and voiceless lateral is often very sharply marked in the waveform and the spectrogram, and could hence be accurately detected by simple signal processing procedures)

Vowels: The distinction between short and long vowels is entirely phonemic and thus not based on the real duration of the segments. When marking length of segments we made these assumptions: All stressed Norwegian syllables contain an element of length which is realised by a long vowel e.g. /e:/, a diphthong e.g. /{i/, or a short vowel followed by one or more consonant(s) e.g. "oss", ('us'), /Os/, "ost", ('cheese'), /ust/. Thus only long single vowels are marked for length.

In continuous speech the vowels in unstressed syllables tend to be reduced or neutralised to a schwa, [@][9], or they simply disappear. An example of this is the word "som", which is pronounced [sOm] in isolation (citation form), and [s@m] or [sm] in continuous speech.
As can be seen in the Norwegian Phoneme Inventory, Appendix A, the schwa is not a phoneme in Norwegian. Since the transcription was based on what we saw and heard, we would chose the phoneme symbol which was most auditorily alike; most often /e/.
Later on we listened through the database once more and marked all neutralised vowels as schwa @. This resulted in 180 schwas, of which 167 substituted /e/, 7 /e:/, 3 /i/, 2 /O/ and 1 /{/.

In sequences of vowels which do not constitute diphthongs the transition between the vowels is often realised by a short period of creaky voice, specially at word and morpheme boundaries when the following vowel is stressed, as between the words "kunne ingen" (=could nobody) in figure 3.2. The segment boundary was put in the middle of the creaky area that could be seen in the spectrogram. In some instances a short period of silence appeared between the vowels. This pause was not perceived when listening in context, but was clearly seen in the waveform and therefore transcribed as silence.

Diphthongs: In a phonemic transcription the two vowels which constitute the diphthong have to belong to the same syllable. For example in the word sequence "skrive i" (=write in) we cannot transcribe /skrivei/, but /skrive i/, even though it phonetically might be identical with the pronunciation of the diphthong /ei/.
In diphthongs of short duration the spectrogram showed little formant transition, whereas between two equal vowel phonemes or within one vowel there could be a large formant transition (see /tidn/ in figure 3.2 or /nes/ and /sin/ in figure 4.9).

Nasals: Some nasals look very short in the spectrogram, but often they are realised by nasalising the preceding sounds. We looked for intensity changes in the spectrogram when we segmented the nasals, e.g. /in/ in figure 4.9.

Plosives: Because this was a phonemic transcription, /b d g/ were chosen as symbols even when they sometimes were partly or fully devoiced.

[9] The SAMPA symbol @ represents an unstressed central vowel, a schwa.

When a homorganic nasal or lateral follows the plosive, as in "atten" /Atn/ (=eighteen) and "bøddel" /b2dl/ (=executioner), there is a nasal and lateral release respectively of the closure phase. The plosive may then be just a segment of silence. Since diacritic symbols are not used in our annotation, the syllabification of /n/ and nasal release in "atten" are not marked; i.e. only /n/ is marked.

One difference between English and Norwegian pronunciation can be seen in the case of two subsequent plosives where both will have an audible release in Norwegian, while only the last one will be audibly released in English. For instance the word "act" will in English be pronounced [{kt^h], and "akt" in Norwegian [Ak^{h}t^h]. Hence, we marked the end boundary for /k/ when the aspiration disappeared.

R-sounds:[10]

The most common pronunciation of /r/ in Norwegian is as an apical alveolar tap, i.e. the tongue tip touches the alveolar ridge and makes a short closure phase. On the spectrogram the closure phase of the tap shows up as a low frequency voicebar very similar to the closure phase of the retroflex, lateral flap /rL/. Figure 4.4 below depicts the waveform and the broad band spectrogram of the words "åre" /O:re/ (=oar) and "åle" /O:rLe/[11] (=crawl).

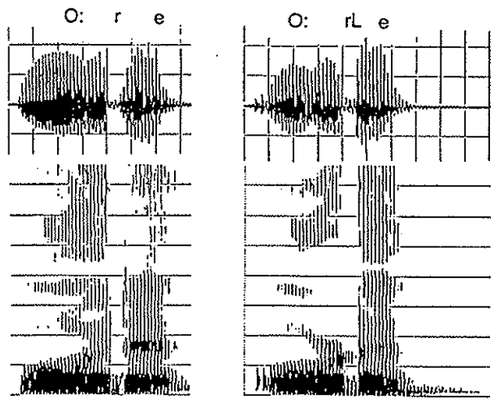


Figure 4.4 *Waveform and broad band spectrogram of /O:re/, /O:rLe/ pronounced in isolation by a male speaker. Neighbouring vertical lines are 50ms apart. In the spectrogram the frequency difference between neighbouring horizontal lines is 1 kHz.*

[10] Parts of this discussion are also published in [Kvale'92].

[11] In Norwegian /rL/ is a /l/-variant, but due to the acoustic similarity with the apical tap speakers with other phonemic filters (e.g. British and American) perceive the /rL/ as an /r/-variant. /rL/ does not occur in the Norwegian EUROM0 recording.

We thus assumed that the tongue movement for the alveolar tap is a symmetric one, i.e. the apex takes the same time to move up to the alveolar ridge as away from it after the closure phase. We therefore included 10 ms, or approximately one pitch period, on each side of the /r/-segment in addition to the portion of less intensity in the spectrogram to include the tongue tip movement.

However, the symmetry was only found for /r/ in intervocalic position and in vowel-/r/-voiced fricative /v/ position. Accordingly, some few exceptions from the 10ms-addition-rule mentioned above had to be made:

A. After a plosive burst or a voiceless fricative as in the words "tro" /tru:/ (=belief), "dro" /dru:/ (=left) and "fri" /fri:/ (=free) a period of voicing and formant structure is seen in the spectrogram before the tongue tip reaches the alveolar ridge. This is also the case for /r/ in sentence initial position as in "ren" /re:n/ (=clean).
This voiced portion which was often much longer than the 10 ms we defined for the symmetrical /r/ realisation was included in the /r/ segment, as seen in figure 4.5. In some words, especially where a voiceless plosive precedes the /r/, as in "prinsipp" (=principle) and "prate" (=chat) the voiced period is perceived as an epenthetic @ sound between the plosive and /r/. This @ was not transcribed but included in the /r/ segment, see also /sifrene/ in figure 4.10.
For some speakers the /r/ may also be partially devoiced particularly if the preceding phoneme is a voiceless sound.

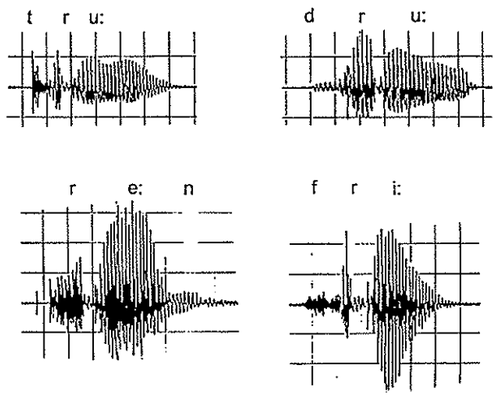


Figure 4.5 *Waveforms of /tru:/ and /dru:/, /re:n/ and /fri:/ pronounced in isolation by a male speaker. Neighbouring vertical lines are 50ms apart.*

B. When /r/ precedes a voiceless fricative, voiceless plosive or a nasal, a short period of voicing appears after the /r/ closure as seen in figure 4.6 and in /fO:rfO/ in figure 4.10. This effect may also occur at the end of sentences.

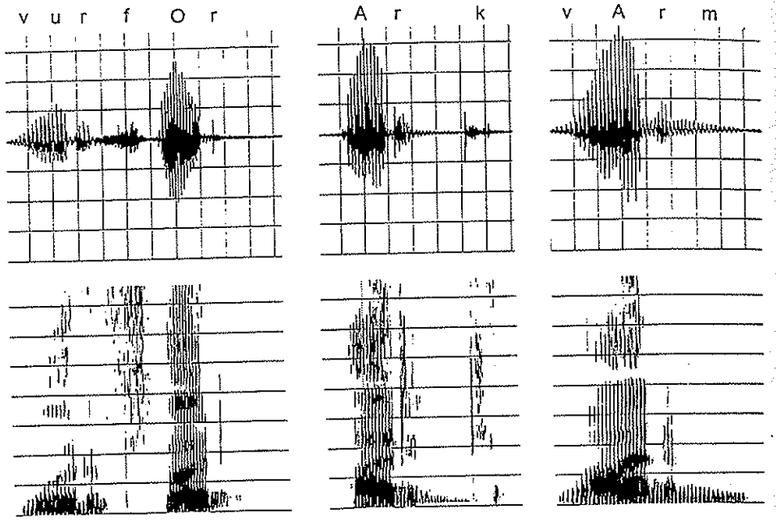


Figure 4.6 *Waveform and broad band spectrogram of "hvorfor" /vurfOr/ (=why), "ark" /Ark/ (=sheet of paper), and "varm" /vArm/ (=warm) pronounced in isolation by a male speaker. Neighbouring vertical lines are 50ms apart. In the spectrogram the frequency difference between neighbouring horizontal lines is 1 kHz.*

In figure 4.5 and 4.6 we notice that even when one side of the /r/ does not fulfil the symmetric assumption one pitch period of the neighbouring phoneme should still be included in the /r/-segment.

C. At the end of sentences the amplitude of /r/ normally decreases evenly towards the end as seen in /vurfOr/ in figure 4.6, and -/er/ in figure 4.7. If the voice source is turned off early the last part of /r/ becomes devoiced. This voiceless part was also included in the /r/-segment.

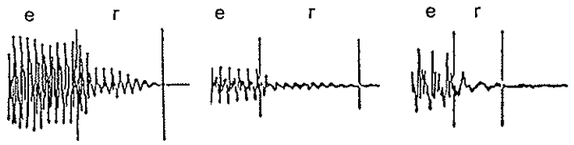


Figure 4.7 *Waveform of -/er/ pronounced by three speakers. Taken from the end of a longer word at the end of an utterance in the Norwegian EUROM0 recording. The time axis differs from that in the other figures concerning the r-sound.*

There is no standard pronunciation of Norwegian and the pupils' right to use their own dialect pronunciation in school is stated in education laws. Depending on the speaker's dialect background, the /r/-phoneme is produced as an apical tap or trill, a postalveolar approximant, a uvular tap or trill, or a post-palatal, velar or uvular fricative or approximant [Foldvik'87]. However, only apical tap realisation occurred in the Norwegian EUROM0 recording.

In figure 4.8, an apical alveolar tap, a dorso uvular approximant, and an apical trill are compared in intervocalic position. The uvular approximant with no clear acoustic boundary cues makes the segmentation more difficult. The uvular /r/ when preceded or followed by an unvoiced sound is realised as a fricative.

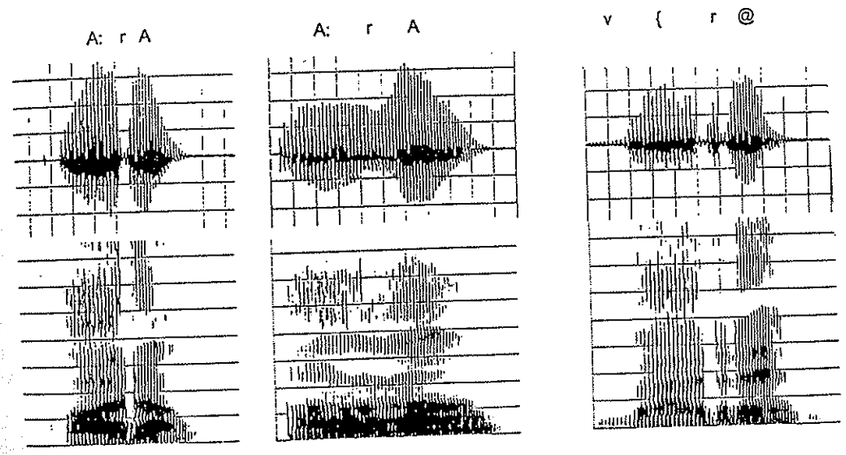


Figure 4.8 *Waveform and broad band spectrogram of /A:rA/ pronounced in isolation by a male speaker, with an apical alveolar tap to the left and an uvular dorsal approximant to the right, which can be considered as the two articulatory extremes of Norwegian /r/-pronunciation. The apical trill is exemplified to the right with the word "verre" /v{r@/ (=worse).*
Neighbouring vertical lines are 50ms apart. In the spectrogram the frequency difference between neighbouring horizontal lines is 1 kHz.

4.2.3 TRANSITIONS BETWEEN PHONEME CLASSES

In the following the segmentation strategy applied for Norwegian is described for phoneme classes.

Vowel / Nasal: Sometimes there was no doubt about the boundary between the vowel and the nasal because of the abrupt change in intensity in the spectrogram. At other times we put the boundary where formants F_2 and F_3 became much less intense, and F_1 of the vowel continued with the same intensity into the nasal. If no such change occurred the decision was based on careful listening which in this case was particularly difficult because of the nasalisation. Since nasalised vowels are not phonemic in Norwegian, as opposed to e.g. French, the speaker may lower the velum long before he finishes the articulation of the vowel. This may give a very long period of coarticulation of the two phonemes. Examples of this transition are seen e.g. in figure 3.2; /}n/, in figure 4.9: /en/, /}n/, /in/ or in figure 4.10; /en/.

Nasal / Vowel: The same problem occurred as for the transition vowel/nasal, but now the boundary was placed when F_2 and F_3 became much more intense in the spectrogram, see e.g. /ne/ and /me/ in figure 4.9.

Nasal, Vowel or Lateral / Silence or voiceless plosive: The boundary was placed where the voicing finished, i.e. when there was no visible amplitude in the waveform, see /l.../, /nt/, /et/[12] and /n.../ in figure 4.9 and /Ok/ and /ek/ in figure 4.12.
In some instances the vowels became devoiced due to preaspiration effect of the succeeding voiceless plosive. This unvoiced period was included in the vowel, i.e. we defined the plosive as beginning when a complete closure was reached and no amplitude was seen in the waveform.

Nasal or Vowel / Voiced plosive: We placed the boundary where a intensity drop was seen in the spectrogram and a significantly smaller amplitude appeared in the waveform, as seen e.g. in /ng/ in figure 4.9. A portion of low amplitude waveform was thus included in the voiced plosive, and often the voiced waveform was seen during the whole plosive as for /d/ in figure 3.2.

Vowel / Voiceless fricative: The friction starts before the voicing of the vowel is finished, which can be seen in the waveform as a high-frequency wave superimposed on the smooth waveform of the vowel. We adopted the convention that all voicing belongs to the vowel even if the friction has started. An accurate boundary decision based on the waveform is thus performed, see e.g. figure 4.10. Looking at the spectrogram we placed the boundary where F_2 and F_3 disappeared and the frication was visible, even though we noticed that F_1 of the vowel continued into the frication area (spectrum with high frequency noise component).

[12] Notice that in the case of /et/ the closure phase of the plosive is longer than the 50 ms defined for a plosive in the beginning of an utterance. This has to be accounted for e.g. in pause detectors in ASR.

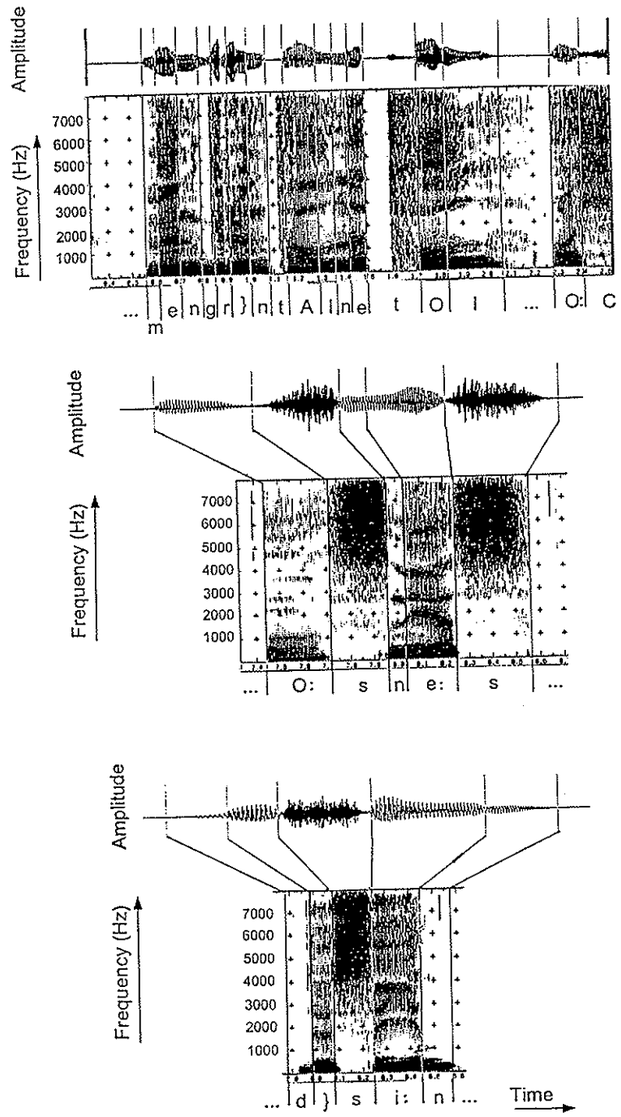


Figure 4.9 *Examples of the segmentation conventions with waveform and spectrogram of utterances of speaker AFN. The + signs in the spectrograms are separated by 0.1 sec. horizontally and 1000 Hz vertically.*
At the top: "men grunntallene tolv og tj(ue)", labelled /men gr}ntAlne tOl ... O:C/.
In the middle: "og snes", labelled /O:sne:s/.
At the bottom: "dusin", labelled /d}si:n/.

Nasal / Voiceless fricative: In the spectrogram the nasals are characterised by low frequency intensity, F_1 or murmur (see also [Zue'89],[Ladefoged'82]). The frication starts before the intensity of the murmur has begun to decrease. At this point it is not possible to hear the frication due to masking effects, although it is visible in the waveform and spectrogram. When F_1 of the nasal becomes less intense it is possible to hear the frication, and we placed the boundary here, even if F_1 continues into the frication area (as for Vowel/Fricative).

Lateral / Plosive or Fricative: The same procedure as for the nasals was followed.

Vowel / Lateral: This was often difficult to segment due to smooth transition in the spectrogram and waveform, and hence careful listening was applied (figure 4.2). But in e.g. /Ol/ in figure 4.9, an intensity drop of F_2 was noticed in the spectrogram and used as a boundary cue.

Voiceless Fricative / Vowel: This was on the whole easy to segment, i.e. an abrupt change was seen in the spectrogram and the unvoiced-voiced boundary was based on the same convention as for the voiced-unvoiced transition described above. Examples are depicted in figure 1.2, figure 4.9 and figure 4.10.
Some speakers even had a brief segment of silence between the phonemes. This pause was included in the fricative as seen in /si/ in figure 1.2 and figure 4.9.

Voiced Plosive / Vowel: If there were no separate acoustic cues for the burst of the voiced plosive the boundary was placed immediately after the visible beginning of the vowel, i.e at the first positive amplitude in the waveform (e.g. /be/ in figure 1.2 and /rdi/ in figure 4.10).

Lateral / Nasal or Vowel: Usually there were little information in the waveform and spectrogram, as in /ln/ in figure 4.9, so the segmentation was based on careful listening which was as difficult as for the vowel/nasal transition. However, sometimes there was a clear difference in intensity visible in the spectrogram and the segment boundary was placed there.

Silence / Plosive: The burst of the plosive was easily detected but where the closure phase for the plosive starts was not clear from the waveform or spectrogram. Accordingly the convention was established that the plosive started 50 ms before the beginning of the burst, as in figure 1.2, or 50 ms before the beginning of the voicing, as in /d}sin/ in figure 4.9.

Fricative / Plosive or Silence: The frication decreases smoothly towards the closure of the plosive or towards the silence. We marked the boundary where the last positive amplitude was seen in the waveform as in /sp/ in figure 1.2 and /s.../ in figure 4.9.

Plosive / Fricative: The aspiration of the plosive continues into the friction of the fricative, but usually it was possible to detect an intensity difference in the spectrogram.

Voiceless Plosive / Vowel and Silence / All classes (except plosive): This was generally very simple to segment due to the abrupt changes in the spectrogram and waveform, as in /tO/ in figure 4.9.
For the silence to voiced sound transition the boundary was placed at the first amplitude (bigger than the background noise) seen in the waveform (e.g. /... e/ in figure 3.2, /... O/ in figure 4.9 and /...i/ in figure 4.12).

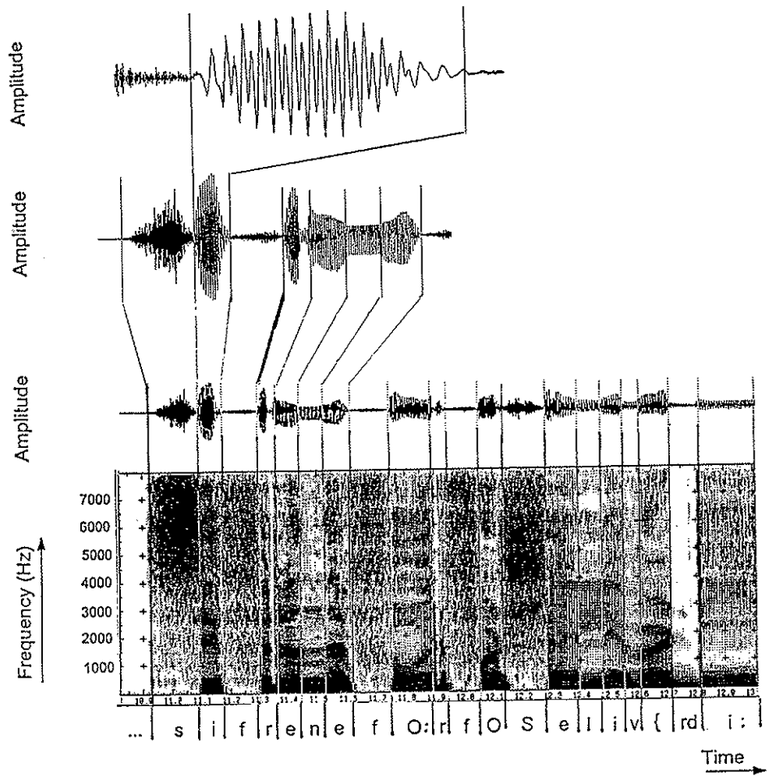


Figure 4.10 *Manual segmentation and labelling of "siffrene får forskjellig verdi" ('the numbers get different values'), read aloud by speaker AFN and labelled as /sifrene fO:r fOSeli v{rdi:/.*
The + signs in the spectrogram are separated by 0.1 sec. horizontally and 1000 Hz vertically.

4.2.4 CONSISTENCY

To check how consistent the annotation had been and to measure properly how fast it is possible to work, the manual segmentation and labelling of about 2 minutes of the speech material of speaker AFN was performed three months later by the author. It took 135 minutes to segment and label 748 segments, i.e. 5.5 phonemes per minute. Compared to the previously annotated material the labeller had omitted two phonemes and three pauses were left out. On the other hand, it was also found that the first annotation had left out two pauses. This shows that segmentation and labelling will entail some errors, in this case 4 errors out of 748, or about 0.5%. Otherwise, the labelling was identical. The agreement between the two sets of segment boundaries is shown in figure 4.11 showing that 96.5% of the boundaries coincided within ±20ms deviation. These results should be taken into account when assessing the accuracy of ASS algorithms.

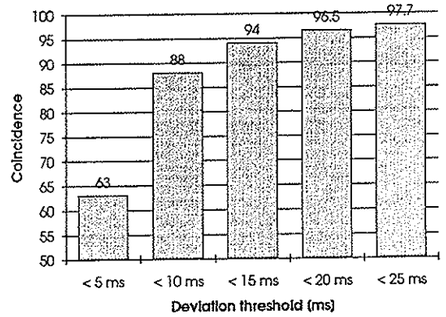


Figure 4.11 *Coincidence between two manual segmentation and labelling of the same material performed by the same labeller with a 3 months' interval.*

A similar experiment is reported for 13 sentences read by a Dutch professional male reader [Erp'88]. A *given* narrow auditory phonetic transcription (with 158 labels) was time aligned by two phoneticians independently. The labellers had agreed upon a common set of segmentation criteria. When one of the labellers segmented the same speech material about three months later, 24% of the boundaries differed more than 10ms from the previous placements; i.e. 12% more than in my experiment. These differences were claimed to be due to lack of explicit segmentation criteria and random errors. Another reason may be that it is more difficult to align the narrow auditory phonetic transcription, e.g. more detailed criteria are needed, than performing phonemic labelling and segmentation. The small speech material may also contain more difficult transitions to segment.
The difference between the two labellers exceeded 10ms for 28.5% of the boundaries (i.e. for 45 boundaries). This shows that the first experiment is fairly representative for the consistency between labellers.

For Italian, three experts segmented and labelled 10 continuous sentences [Cosi'91]. The segmentation reliability was computed by regarding one of the labellers at the time as reference and comparing this with the two others. (Test and reference setting were circulated to cover all possible combinations). Here, 738 out of 789 boundaries (93.5%) were within ±20ms deviation from reference, which is comparable with the results in figure 4.11.

In a cross-comparison test [Erp'89b] a Dutch phonetician aligned a *given* phonemic transcription with speech by indicating the *centres*[13] *for each phoneme* of the first two sentences of the English speaker JH, and the first sentence of the French speaker MD and Danish BL in EUROM0. The labeller spoke perfect English, moderate French but had no knowledge of Danish. The centre positions were then compared to the computed midpoints of the segment boundaries placed by native phoneticians in each language. For English less than 1% differed more than 20ms, for French 15% differed more than 20ms, and for Danish 13% differed more than 20ms.

This comparison across languages shows that segmentation of speech is particularly difficult when the labeller does not know the common realisations of the phonemes in the actual language, such as the Danish "stød". The deviations found were thus explained as a result of language specific aspects, random human variations, and the use of different equipment.
Another reason for the difficulties with defining the centre of phonemes may have been that many allophones are not symmetrical and give different segments depending on which acoustic parameters the decision is based on (see e.g. [Kvale'92]).

The very small speech material and the comparison of segment centres make it inconvenient to compare the results with my experiment.

4.2.5 TRANSCRIPTION LEVELS

Our automatic segmentation algorithm needs a string of labels as input, see figure 1.6. For each given label sequence it computes the corresponding segments. This implies that a transcription of the actual utterance has to be performed **prior** to the automatic segmentation. (The transcription is often given in manual segmentation tasks also, see e.g. [Erp'88]).

Within the SAM project, the automatic segmentation procedures were trained and tested with the phoneme string produced in the manual segmentation and labelling task. This was only a preliminary solution. For future applications, e.g. for annotation of the huge EUROM1 database [Sherwood'92], a transcription based on minimal human effort is desirable and also necessary. However, other types of transcription will put an extra computation load on the automatic segmentation algorithms in that they will have to allow for deletions, changes, and insertions.

In order to measure the deviation from the "correct" transcription, the following transcriptions of the Norwegian EUROM0 recording were compared:

1. **Text-To-Phoneme (TTP):** This is the output from a Norwegian TTS program [Ottesen'91].
2. **Citation:** A transcription of words as spoken in isolation (see chapter 3.1).
3. **Phonotypical:** A transcription predicting what would be written down in an auditory transcription of fluent continuous speech, including assimilations and elisions.
4. **Auditory:** An auditorily based transcription, without access to a visual representation of the signal.
5. **Manual:** This is the phoneme string produced in the manual segmentation and labelling task with both auditory and visual information available. This annotation is regarded as "correct".

[13] The "centre" of a segment is here defined as the most salient points, i.e. roughly the middle. For plosives, centre is the point just before the burst [Erp'89b].

The five transcriptions employed the same phoneme inventory, i.e. the Norwegian SAMPA set. The three first transcription approaches predict the phoneme sequence from the orthographic text without access to what is actually spoken. In this prediction it is in some cases difficult to decide which symbol to use, e.g. when transcribing the short /e/-sound which has got several allophones but which has to be transcribed by one of the two symbols /e/ and @. A similar problem is encountered when transcribing vowels in the /u/-/O/ area. Problems also arise when selecting one of several alternate pronunciations, e.g. the capital Oslo pronounced /uslu/ or /uSlu/.
The citation, the phonotypical, and the auditory transcription were performed by a trained Norwegian phonetician.

The five transcription levels are exemplified below by the first part of the first sentence of the Norwegian EUROM0 recording:

Orthographic:	/ i språket kan vi skrive uendelig mange ord med et lite sett ... /
TTP:	/ i sprO:k@t kAn vi: skri:v@ }:@n@li mAN@ u:r me et li:t@ set ... /
Citation:	/ i: sprO:k@ kAn vi: skri:v@ }:end@li mAN@ u:r me: et li:t@ set ... /
Phonotypical:	/ i sprO:k@ kAn vi skri:v@ }endli mAN@ u:r me et li:t@ set ... /
Auditory - AFN:	/ i sprO:k@ kAn vi skri:v@ }endli mAN@ u:r me et li:t@ set ... /
Auditory - TBN:	/ i sprO:k@ kAn vi skriv@ }endli mAN@ u:r me et li:t@ set ... /
Manual - AFN:	/ i sprO:ke kAn vi: skri:ve }@ndeli mANe u:r ... met li:t@ set ... /
Manual - TBN:	/ i: sprO:k@ kAn vi skrive }endli mANe u:r me:t li:t@ set ... /

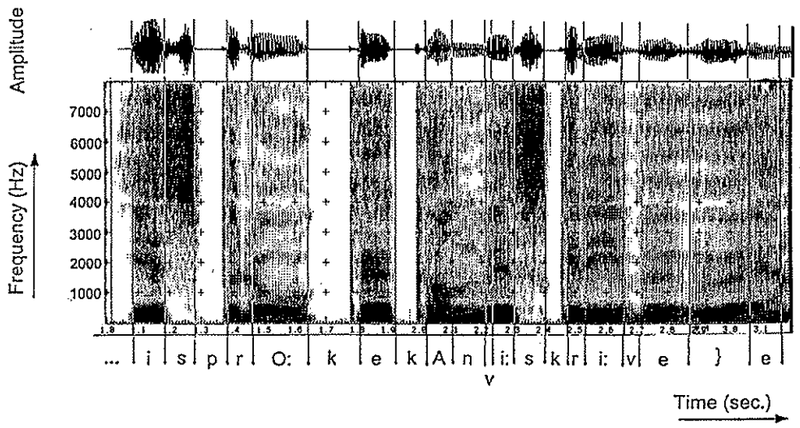


Figure 4.12 *Waveform and spectrogram and manual segmentation and labelling of "i språket kan vi skrive" read aloud by speaker AFN.*
The + signs in the spectrogram are separated by 0.1 sec. horizontally and 1000 Hz vertically.

Comparing the number of phonemes:[14]

The first row of table 4.1 shows that the Citation Transcription and the TTP-program contain more phonemes than the other transcriptions. The main reason for this is that neither of them accounts for assimilation effects over word boundaries. For example, if in fluent Norwegian speech one word ends in /r/ and the next starts with the plosive /t/ or /d/, the plosive becomes a postalveolar one and the /r/ is elided. In this way the words "benytter ti" becomes [b@nyt@r ti] in TTP and citation form, and [b@nyt@ rti] in the others. Another reason is that in TTP the mute /t/ in the definite singular neuter form of nouns will be transcribed. A third reason is that the citation transcription and TTP have many more instances of /h/ and /r/ than the others forms of transcription.
The number of phonemes in the auditory transcription differed by only one phoneme compared to the manual annotation of speaker AFN, but for speaker TBN 20 more phonemes was detected in the auditorily based transcription. In the manual segmentation and labelling geminate sounds were not notated with two symbols if there were no acoustic cues in the spectrogram or waveform to justify this, see e.g. "med et" in the example in figure 4.12 above. Also, in the auditory transcription we may reconstruct sounds that are elided, cf. chapter 2.

	Manual (mean)	Auditory TBN	Auditory AFN	Phono-typical	Citation	TTP
No. of phonemes	1125	1132	1145	1132	1184	1190
No. of long vowels	138	77	81	88	153	105

Table 4.1 *The number of phonemes and long vowels in auditory transcription of speaker TBN (Auditory-TBN) and AFN (Auditory-AFN) and in the Phonotypical, Citation and TTP transcription compared with the mean value of the manual segmentation and labelling of the four speakers.*

In the second row of table 4.1 we see that the number of long vowels differs in the transcriptions. This is due to differences in the transcription strategies. The manual segmentation and labelling is strictly phonemic, i.e. long vowels are marked as long whether they are acoustically long or not. In citation form and TTP each word is stressed and hence more long vowels can be expected. The auditory and phonotypical transcriptions are "narrow" in the sense that long vowels are only marked as long when perceived long.

The citation transcription and TTP (and also the Auditory and the Phonotypical transcription) have a higher number of [@] and accordingly fewer /e/ than the manual segmentation. This indicates that we have been too restrictive when transcribing [@] in manual segmentation, and also that perception of different vowel qualities varies. For instance cues which provide phonemic distinctions in a language are much more noticeable than those that do not (cf. chapter 2). Since [@] is not a phoneme in Norwegian it may have been more difficult to categorise.

[14] A more detailed analysis of the annotated Norwegian EUROM0 speech material is given in Appendix B.

Plosives, nasals, laterals and fricatives have almost identical number of occurrences irrespective of transcription, while for vowels this is not the case. One reason for this is that consonants are perceived more categorically than vowels (cf. chapter 2).

The difference between label strings may be measured by a string match procedure usually applied for evaluation of automatic phoneme recognisers, see section 7.5. Regarding the manual labelling as reference, this string match calculates the number of deletions, substitutions, and insertions in the TTP-generated label string[15]. For instance for the first sentence of AFN shown in figure 4.12 (i.e. to the first /.../), the TTP generated label string contained 24 phonemes which corresponded to the manual annotation of speaker AFN, whereas 6 phonemes were substituted, 1 phoneme was deleted, and 1 phoneme was inserted.

For all the 127 sentences in the manual annotation of speaker AFN 34 TTP-generated sentences were identical and 992 (out of 1144) phonemes corresponded, whereas 149 phonemes were substituted, 3 deleted, and 50 phonemes inserted.
In a similar comparison with the manual segmentation of speaker TBN, 8 out of 72 sentences were identical and 949 (out of 1112) phonemes corresponded, whereas 162 phonemes were substituted, 1 deleted, and 80 phonemes inserted.
As regards the accuracy measure defined in section 7.5, the TTP-generated label string obtained 82.3% accuracy when compared to manual segmentation of AFN and 78.2% accuracy when compared to manual segmentation of TBN.

The TTP-generated label string corresponded best to the manual segmentation of speaker AFN. The high number of substitutions was mainly due to different transcriptions of short and long vowels and to the substitution of /e/ by [@]. The number of insertions confirmed that the TTP-label string represents an "over-articulation" compared to "normal" articulation.

Comparing time used for annotation

The manual segmentation and labelling session took about 34 hours, which equals 2.2 phonemes/minute. This included some breaks, discussions and waiting for spectrogram computing. The most time-consuming part was the careful listening. Those parts of the recordings where there were few clear acoustic cues accounted for most of the discussions and approximately 30% of the time spent on manual segmentation. The speed of segmentation was increased to 5.5 phonemes per minute when the spectrograms had been computed first, the procedure was known, and the decisions were taken by one person only. In comparison, [Høhne'83] and the French GRECO-PRC Labelling Taskforce, referred to in [Perennou'89], used about one minute to segment and label one phoneme.
Auditive transcription of two speakers took about 4 hours or 9.5 phonemes/minute. Citation form and phonotypical transcription was done in 2 hours i.e 19 phonemes/minute. TTP on the other hand, is totally automatic.

[15] In order to make correspondence with the manual segmentation the TTP-generated label string was divided into the same sentences as for the manual segmentation of the given speaker. The label string used for manual annotation contained in this case [@], see Appendix B.

4.3 SUMMARY

In the beginning of this chapter three reasons for investigating different annotation strategies were given. First we compared the segmentations of some typical phoneme transitions in order make a platform for development of common segmentation conventions. We found that each labeller had his own strategy, but often it differed only slightly from the others. Thus, based on this comparison different labellers may agree upon a multi-lingual segmentation standard which covers most of the phoneme-transitions.
The second point was to find the correlation between the phoneme symbols and the sounds in each language in order to increase the training set for a statistically based automatic segmentation algorithm. We found that the consonant symbols mainly covered the same acoustic realisations in the different languages, whereas the vowel realisations differed more. Details of this comparison will be discussed in chapter 7.
Finally we checked if the annotation was performed at the same level in the different languages. We found that although the starting point was a phonemic annotation, some sounds and some transitions between sounds were annotated by applying different criteria. Most often no reasons for these deviations from the broad phonetic level were given.

Our annotation strategy and the segmentation and labelling criteria applied to Norwegian was described. A comparison of two annotations by the same labeller with three months' interval gave an indication of the upper limit of accuracy that an automatic segmentation procedure can achieve. The coincidence was better than reported for similar experiments elsewhere. However, very few such comparisons have been performed and none are directly comparable due to differences in speech material and in type and level of annotation.

In comparing different transcription levels, the TTP transcription seems most promising as a front end to an automatic segmentation algorithm in that it is automatic and does not differ too much from the correct manual labelling. It can also be further improved by e.g. ruling out the mentioned mute /t/ and some occurrences of /h/ and /r/.

Since many different conventions exist and humans always exhibit some inconsistencies and errors, a reasonably good automatic segmentation method which is always consistent may, for many purposes, be preferable. An automatic segmentation can at least be used as a preprocessing step to suggest boundaries for the human labeller.

Chapter 5

MODELLING SPEECH

The speech signal carries redundancies, i.e. it carries more information than is necessary for speech perception, and for manual and automatic speech segmentation. Since the most important features for perceiving speech are contained in the spectrum, the speech waveform can be effectively represented by spectral parameters. The power spectral density in a short time interval, the *short-time spectrum*, of a voiced speech sound basically consist of two components: (i) a rapidly changing *spectral fine structure* corresponding to the periodicity of the sound source, and (ii) a slowly changing *spectral envelope* which corresponds to the overall spectrum of the glottal source, the vocal tract resonances and anti-resonances, and the radiation characteristics.

The spectral envelope conveys the identity of the voiced sound. One method of estimating the spectral parameters from the speech waveform is by *nonparametric analysis*, such as short-time autocorrelation, short-time spectrum, complex cepstrum, and bandpass filter bank analysis. Since no model for the source of the signal is assumed, this analysis can be applied for any signal.

However, since we know that the signal is speech and we know a lot about the speech production, we can apply *parametric analysis*, where the speech *signal source* is modelled and the signal is approximated by adjusting the source model parameters. If the model is reasonable, the parametric analysis technique achieves a more accurate approximation to the signal than the nonparametric analysis.

Below, a model of the speech production, the *lossless tube model,* and an estimate of the model parameters by *linear predictive analysis* are described. This theory is derived from [Flanagan'72], [Markel'76], [Rabiner'78], [Furui'85] and [Orfanidis'88]. Then the cepstrum coefficients are defined and a cepstral filtering technique is discussed. At the end of this section the deterministic and stochastic *dynamic programming* algorithm is outlined based on [Cooper'81], [Bertsekas'87] and [Silverman'90]. The theory of Markov chains and general hidden Markov models (HMMs) are explained e.g. in [Rabiner'89] and [Picone'90]. Here, the notation of HMM needed for the discussion of automatic segmentation is introduced, and the fundamental problems to be solved in HMM-based ASR and ASS-systems are described.

5.1 THE LOSSLESS TUBE MODEL FOR SPEECH PRODUCTION

One approximation to the **vocal tract** configuration at a given time instant is the *lossless tube model*, which is a concatenation of equal length tubes with different cross-sections. Since the tubes are without loss, the air pressure wave is only altered at the junctions between two tubes with different cross-sections, where some air is reflected and the rest transmitted. The vocal tract model can thus be characterised by a set of areas, or equivalent reflection coefficients. Each

junction corresponds to a one-pole digital filter. The vocal tract can thus be modelled as a concatenation of such filters. With N tubes the vocal tract is represented with an all-pole transfer function[1] V(z) of order N, corresponding to $^N/_2$ complex poles or $^N/_2$ formants.

The lossless tube model approximation to the vocal tract is crude and neglects e.g. the losses due to viscous friction between the air and the wall of the tube, heat conduction through the walls of the tube, and tube wall vibrations, which mainly influence the bandwidth of the formants. Still, this all-pole model gives a good representation of the vocal tract effects for many speech sounds. For instance, many vowel spectra can be approximated by only 3-4 tubes. Models for nasals (and fricatives) should include both resonances and anti-resonances, i.e. both poles and zeros are required in the transfer function. However, the effect of zeros in a transfer function can be achieved by including more poles.

The **voiced excitation** is a glottal pulse of finite length, which gives a lowpass filtering effect in the frequency domain. That is, the glottal pulse should be modelled by zeros only, but a two-pole model for the voiced excitation G(z) is also suggested [Markel'76]. The **voiceless excitation** is represented with a random number generator and a gain parameter. However, only voiced or voiceless excitations are insufficient for modelling voiced fricatives, which are produced by mixed excitations.

Radiation effects can be modelled as $R(z)=R_0(1-z^{-1})$, i.e. a high-pass filter operation. A prolonging of the vocal tract has the same influence on the speech spectrum.

Although the vocal tract is constantly changing its shape during speech production, the main properties of the speech signal vary relatively slowly (see e.g. the speech waveform plots in figures 1.1 and 1.2). The speech signal can thus be approximated by a constant parameter vector within a period of 10 to 30 ms, called a *speech frame*. However, for sounds with abrupt changes such as plosives, this approximation is not a good one. Figure 5.1 below summarizes the modelling of speech production described in this section:

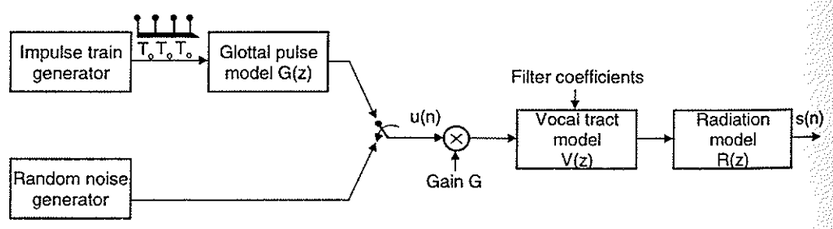


Figure 5.1 *General time discrete model for speech production, after [Rabiner'78, p.105], where the gain of the voiced and voiceless excitation is combined in the gain term G. The timevarying vocaltract is assumed to be fixed for periods of 10-30ms. s(n) is the speech amplitude at sample number n.*

[1] The z-plane representation is used here, see e.g. Glossary and [Oppenheim'75].

5.2 LINEAR PREDICTIVE CODING - LPC

The linear prediction technique was first used for speech analyses and synthesis by [Itakura'68] and [Atal'68]. Since then, linear predictive (LP) speech analysis has been frequently used in speech coding, TTS, ASR, and ASS because it yields robust, reliable, and accurate estimates of important speech parameters such as pitch, formants, and frequency spectra.
The LP analysis is related to the general speech production model in figure 5.1 above, but here the glottal pulse, radiation and vocal tract effects on the speech spectrum are combined with a single all-pole transfer function H(z). The excitation for voiced sounds is thus simplified to an impulse train generator with fixed frequency $F_0=^1/_{T_0}$. The relation between the input excitation U(z) and the output speech signal S(z) is then given by:

$$H(z) = G(z)\cdot V(z)\cdot R(z) = \frac{S(z)}{U(z)} = \frac{G}{1+\sum_{k=1}^{p} a(k)\ z^{-k}} \qquad (5.1)$$

where a_k, k=1,...,p are the filter weights, p is the filter order, and G is the gain term.

In the time domain the last relation of equation (5.1) above can be expressed as:

$$s(n) = -\sum_{k=1}^{p} a(k)s(n-k) + G\cdot u(n) \qquad (5.2)$$

where a_k, k=1,...,p now can be interpreted as the predictor coefficients and p as the predictor order.

Only the speech signal s(n) is recorded, i.e. the source signal u(n) is unknown, and a speech sample has to be approximated as a linear combination of the p past speech samples:

$$\hat{s}(n) = -\sum_{k=1}^{p} a(k)s(n-k) \qquad (5.3)$$

In voiceless sounds u(n) has a relatively constant energy within the analysis window, whereas in voiced speech the energy in u(n) is concentrated at the start of each pitch period. Because ignoring u(n) is valid for more time samples for voiced than for voiceless speech, the LPC modelling provides a better match to the voiced speech spectra [O'Shaughnessy'90].

The prediction coefficients are estimated by minimizing the expected squared prediction error

$$E\left\{ [s(n)-\hat{s}(n)]^2 \right\} = E\left\{ [s(n) + \sum_{k=1}^{p} a(k)s(n-k)]^2 \right\} \qquad (5.4)$$

by setting its partial derivative with respect to each prediction coefficient to zero. This leads to the Yule-Walker or *normal equations*:

$$\sum_{k=1}^{p} a(k)R(i-k)=R(i) \quad ; \quad i\geq 0 \tag{5.5}$$

where

$$R(k) = E[s(n)s(n-k)] \tag{5.6}$$

is the autocorrelation function, *acf,* of k lags for the stochastic time series s(n) with zero mean value.

Due to the time varying nature of speech, the acf can only be estimated from a limited portion of the signal, i.e. within an *analysis window* containing N samples. Different assumptions of the data outside the analysis window lead to different approaches for estimating acf; the two most frequently encountered are the *autocorrelation method* used in this thesis, and the *covariance method.*

In the **autocorrelation method** it is assumed that the speech signal is zero outside the analysis window. With non-zero values within $0 \leq n \leq N\text{-}1$ only, the acf is estimated as

$$\hat{R}(k) = \sum_{n=0}^{N-1-k} s(n+k)s(n) \qquad for \quad 0 \leq k \leq N-1 \tag{5.7}$$

and the prediction coefficients are estimated from:

$$\sum_{k=1}^{p} a(k)\hat{R}(|i-k|) = \hat{R}(i) \qquad for\ 1 \leq i \leq p \tag{5.8}$$

This acf matrix has a *Toeplitz structure* (i.e. it is symmetrical and all the elements along a given diagonal are equal), and is easily inverted by the *Levinson-Durbin recursive algorithm.* The resulting all-pole filter is guaranteed to be stable.

In the **covariance method**, on the other hand, no assumption is made about the observations outside the analysis window. All the speech data available is used, but the prediction error is set to zero outside the window from 0 to N-1. Here the resulting acf matrix will be symmetrical but not Toeplitz, and the equations are solved by Cholesky decomposition. This method may give an unstable inverse filter.

5.3 CEPSTRUM COEFFICIENTS

For ASR and speaker identification the LPC parameters or some LPC-derived representations, such as Line Spectral Pairs (LSP), Partial Autocorrelation (PARCOR), acf of LPC, and LPC-cepstrum are often used. The derivation of, and the relation between, these entities are shown in e.g. [Furui'85, pp.85-134]. In comparisons of different representations and distortion measures for ASR the *cepstral coefficients* and the *cepstral distortion measure* obtained good performance [Atal'74], [Atal'76], [Gray'76], and [Nocerino'85].

The cepstrum, or the cepstral coefficient, is defined as the inverse Fourier transform, F^{-1}, of the logarithm of the magnitude spectrum of a time series, e.g. for speech:

$$c(n) = F^{-1}(\log|S(e^{j\omega}|) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|S(e^{j\omega})|e^{j\omega n}d\omega \tag{5.9}$$

The relation between the source and the speech signal in eq. (5.1) can then be expressed as a sum:

$$F^{-1}\log|S(e^{j\omega})| = F^{-1}\log|U(e^{j\omega})| + F^{-1}\log|H(e^{j\omega})| \tag{5.10}$$

where the first term on right side corresponds to the spectral fine structure of the excitation, with a peak in the high-quefrency[2] region for voiced sounds. The second term on the right corresponds to the spectral envelope, with a concentration in the low-quefrency region, i.e. the lower cepstral coefficients. With the cepstral representation the spectral envelope and the fine structure can thus be easily separated.

The cepstrum parameters {c(n)} can be computed recursively from the LPC coefficients {a(n)} by [Atal'74]:

$$\left.\begin{aligned} c(1) &= -a(1) \\ c(n) &= -a(n) - \sum_{k=1}^{n-1}(1-\frac{k}{n})a(k)c(n-k) \qquad 1<n\leq p \\ c(n) &= -\sum_{k=1}^{n-1}(1-\frac{k}{n})a(k)c(n-k) \qquad n>p \end{aligned}\right\} \tag{5.11}$$

Cepstrum parameters derived from eq. (5.11) is called the *LPC-cepstrum.*

When the spectrum is estimated by nonparametric analysis the corresponding cepstrum in eq. (5.9) is often called the *FFT-cepstrum.*
Instead of estimating the spectral values at equal intervals on a linear frequency axis they are often computed at equal spaces on some kind of a logarithmic frequency scale, e.g. a mel-scale or a bark-scale, to better approximate perceptual properties. The bark-scale, which is based on the auditory critical bandwidth, corresponds to the frequency scale on the basilar membrane in the peripheral auditory system, whereas the mel-scale corresponds to the auditory sensation of

---

[2] In cepstral analysis the log spectrum is treated as a pseudo time series and its corresponding power spectrum is the cepstrum. To distinguish from ordinary spectrum the frequency in cepstral analysis is denoted quefrency.

tone height [Furui'85, p.232]. The mel-scale maps the frequency f as:

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{5.12}$$

That is, in the mel-scale the mapping is approximately linear in frequency up to 1 kHz and logarithmic at higher frequencies.
In this thesis the chosen filter bank analysis consisted of triangular filters equally spaced out along the mel-scale, and the mel-frequency cepstral coefficients (MFCC) {c(n)}, n=1,N, were calculated by the following discrete cosine transform applied to the filter bank outputs m_j:

$$c(n) = \sum_{j=1}^{p} m_j \cos\left(\frac{\pi n}{p}(j-0.5)\right) \qquad , \ 1 \le n \le N \tag{5.13}$$

where p is the analysis order and N is the required number of output parameters.

The spectral envelope derived from the LPC-cepstrum approximates the spectral peaks better than the FFT-cepstrum [Furui'85, p.68]. In chapter 7 we will check whether this property gives a more accurate automatic segmentation or not.

5.3.1 CEPSTRAL DISTORTION MEASURES

In order to discriminate between two sounds a short-time spectral distortion (or dissimilarity) measure is needed. With the cepstral representation the squared difference between two log power spectra S_R and S_T, can be computed directly by the *cepstral distortion measure* [Gray'76]:

$$\begin{aligned} d_2^2 &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |\log[S_R(e^{j\omega}) - \log[S_T(e^{j\omega})]|^2 \, d\omega \\ &= [c_R(0)-c_T(0)]^2 + 2\sum_{n=1}^{\infty}[c_R(n)-c_T(n)]^2 \end{aligned} \tag{5.14}$$

where the cepstral coefficients $\{c_R(n)\}$ represent the reference spectrum and $\{c_T(n)\}$ the test spectrum. To alleviate the dependence on the absolute energy level of the signal in eq.(5.14), the term $[c_R(0)-c_T(0)]^2$ can be removed. The upper limit in the summation of eq.(5.14) has to be set to M, where $M \ge p$ is a necessary condition[3] [Furui'85, p.237]. Since the lower order cepstral terms mainly describe the smooth spectrum corresponding to the vocal tract response, the effect of higher order cepstral coefficients should be minimised, either by truncating to a low number of coefficients, or by gradually truncating higher cepstral coefficients by weighting or *liftering* with e.g. a raised sine lifter [Juang'86]. Eq.(5.14) then becomes the *weighted cepstral distortion measure*:

$$d_{wcep} = \sum_{n=1}^{M} \{\, w(n)[c_R(n)-c_T(n)] \,\}^2 \tag{5.15}$$

where $w(n)$ is the weight applied to the nth cepstral coefficients.

[3] If $M < p$ it is probable that the distance value becomes zero even between different spectra and that the positive characteristic of the distance measure cannot be maintained.

Various lifters are proposed for improving ASR. With $w(n)=1$, the *Euclidean cepstral distortion measure*, narrow formants bandwidths are suppressed [Junqua'93]. With $w(n)=n$, the *Root Power Sum* (RPS) distortion measure, has an inherent suppression of the "spectral tilt" [Hanson'86], [Juang'86], and it represents the distance measure proposed by [Klatt'82] of perceived phonetic distance. With a proper choice of L the *sine lifter* can de-emphasize both the lower and higher order cepstral terms:

$$w(n) = 1+\frac{L}{2}\sin(\frac{\pi n}{L}) \qquad ; \ 1 \le n \le L \tag{5.16}$$

In [Junqua'93] it was shown that the sine lifter in eq.(5.16) with L=12 applied on 12 cepstral coefficients enhanced the first formant peak while bandwidths were suppressed. Figure 5.2 shows that sine liftering with L=12 applied on a cepstral vector with 12 Mel-scale cepstrum coefficients increases the cepstral values.

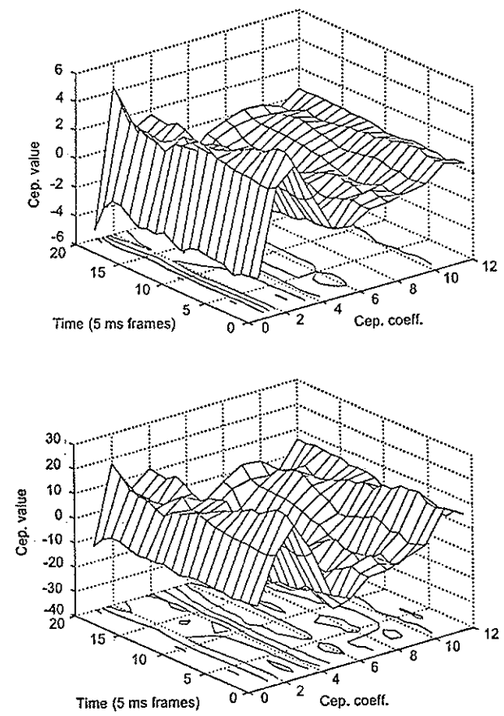


Figure 5.2 *12 Mel-scale cepstrum coefficients without (top) and with (bottom) sine liftering with L=12 for the 20 first frames (i.e. 100ms) of AFN, which is the vowel /i/. Note that the cepstral values are much smaller without liftering (different scale for the cepstral values in the two plots).*

The distortion measures in eq. (5.14) and (5.15) satisfy the requirements that a distance measure applied in speech processing should fulfil [Gray'76], i.e. they have a meaningful interpretation, they are symmetric, positive definite (the distance between two vectors is positive), and easy to evaluate. The cepstral distortion measures have shown good recognition performance in ASR systems, e.g. [Lee,K.F'88], [Rabiner'88].
Furui, [Furui'85, pp. 233-243], summarizes other LPC-based distortion measures and in Nocerino [Nocerino'85] a comparison of ASR performance as a function of distortion measure is given.

5.3.2 ADDING MORE PARAMETERS

So far only instantaneous representations of speech within short time slots are derived. However, for speech perception it is not the absolute values of the intensities that matter but the relative changes. E.g. the formant slope of a vowel preceding a plosive captures perceptual cues for the plosive. Relative changes of the formants are also less speaker dependent than absolute formant values.
The spectral evolution can be measured by the time derivative of the sequence of cepstral vectors, called the *delta cepstrum*. To derivate a discrete sequence is impossible, so the delta-cepstrum coefficient has to be approximated, e.g. by the first-order orthogonal polynomial coefficient or the linear regression coefficient [Furui'86a], [Soong'86], [Rabiner'89a]:

$$\Delta c(n) = \frac{\sum_{k=-K}^{K} k\,c(n+k)}{\sum_{k=-K}^{K} k^2} \qquad (5.17)$$

over a finite length window of (2K+1) frames centred around the current frame. The choice of window length is a tradeoff between accuracy and variance of the estimates.

Increasing the parameter set by higher order spectral parameters as *delta-delta cepstrum*, and by adding *energy* E, *delta-energy* ΔE, and *delta-delta energy* ΔΔE, will further improve the ASR performance, e.g. [C.H.Lee'91], [Wilpon'91].

However, the larger frame vector dimensionality implies estimating more parameters. Since the training material is limited, it is desirable to find which parameters contribute most to the automatic segmentation or the ASR performance. Bocchieri [Bocchieri'92], found by discriminative analysis that ΔE and ΔΔE were the most important parameters for ASR. The 10 next parameters were: c(0)-c(5) and Δc(2)-Δc(5). Since Δc(2)-Δc(5) characterize the dynamics of the formant structure better than Δc(1) and Δc(0), this result was as expected [Bocchieri'92]. (The Δc(0) is the derivative of the logarithm of the linear prediction residual and the Δc(1) is the derivative of the logarithm of the frame "spectral slope").

5.3.3 CEPSTRAL DOMAIN FILTERING

For speaker independent segmentation and recognition tasks it is important to extract relevant phonetic cues only; either by eliminating the individual speaker's characteristics or by compensating for speaker differences. The automatic segmentation experiments in chapter 7 will be performed on the EUROM0 recordings containing only 4 speakers for each language.

Therefore, some of the speaker-sensitive factors such as the overall *spectral tilt* for each speaker, should be suppressed in the front-end of the system. The overall spectral tilt is influenced by e.g. the recording equipment, glottal characteristics of the given speaker and the glottal effort of the speaker [Junqua'93].

High-pass or band-pass filtering of log subband energies has been shown to improve robustness of ASR to convolutional channel distortions, e.g. [Hermansky'91],[Hirsch'91],[Hanson'93].
Most channel distortions, and many kinds of additive noise, vary slowly compared to the variations naturally occurring in speech [Hanson'93]. Thus, filters which remove such slow variations in the parameter vectors have improved ASR rates.

Especially promising is the RASTA (**Rel**A**t**ive **S**pec**TrA**!) processing [Hermansky'91], which has improved the ASR performance within different acoustic environments because it made the spectral representation more robust to steady-state spectral distortions, e.g. [Hermansky'91], [Hermansky'92], [Hanson'93] and [Junqua'93]. A general RASTA-filter can be defined as:

$$T(z) = \frac{k\sum_{n=0}^{N}\left(n - \frac{N-1}{2}\right)z^{-n}}{1-\rho z^{-1}} \qquad (5.18)$$

Selecting k=0.1 and N=5, the filter in eq.(5.18) is the RASTA-filter applied in each frequency band of the analysis proposed in [Hermansky'92]. That is, the RASTA filter is an IIR bandpass filter where the moving average (MA) numerator is a first order linear regression filter, and the autoregressive (AR) denominator gives a leaky integrator effect.
Figure 5.3 displays the frequency and impulse response for this filter as a function of the pole value ρ. By decreasing ρ from 0.98 to 0.86 the lower cut-off frequency increases and the filter will suppress more slowly varying components. The sharp zeros in the frequency response at 28.9 Hz and 50 Hz (or π), remain unchanged with variations in the pole value.

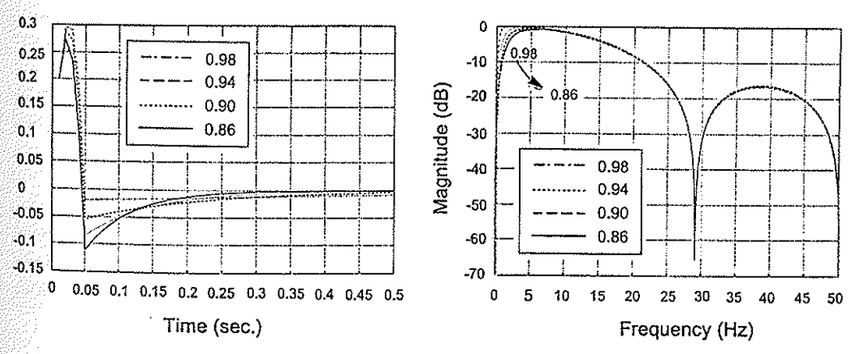


Figure 5.3 *Impulse response and frequency response of the RASTA filter for poles placed at ρ=0.98, 0.94, 0.90, and 0.86.*

The time constant τ of the RASTA filter varies with the selected pole value. From the impulse responses in figure 5.3 we can see that ρ=0.86 gives $\tau\approx$75ms, ρ=0.90 gives $\tau\approx$110ms, ρ=0.94 gives $\tau\approx$175ms, and ρ=0.98 gives $\tau\approx$500ms. That is, a pole closer to one implies a higher time constant, which implies summation over more previous values and therefore a smoother spectrum.

The RASTA filter applied in the log spectral domain attempts to remove convolutional noise by making the long term spectrum identically zero. However, such a filter may also give a similar effect when applied in other domains. Since the cepstrum coefficients are linearly related to the logarithmic spectrum, eq.(5.9) and eq.(5.13), the RASTA filter can be applied directly on the cepstral coefficients.

Recent studies [Hanson'93], have demonstrated that both band-pass filtering (BPF), i.e. RASTA, and high-pass filtering (HPF) applied on log subband energies improved the isolated word recognition for both labspeech and speech in noisy environments.

Encouraged by the ASR improvements obtained by RASTA-filtering several spectral representations, we will in chapter 7 investigate RASTA-filtering on cepstrum coefficients in the front-end for automatic segmentation. This is reasonable because:

(i) The filtering improved ASR on the noisy speech as well as labspeech (as EUROM0).

(ii) Since there was no consistent difference in improvements between filtering in the log subband domain or other spectral representations, such filtering on e.g. mel-scale cepstrum coefficients (MFCC) may yield the same positive effect.

(iii) Since BPF and HPF had almost the same effect, it is their suppression of very low frequencies which is important for ASR (because the frequency response of BPF and HPF is only similar at very low frequencies).

(iv) Although the problems of ASR and automatic segmentation are different, the techniques involved are similar.

In RASTA-filtering on cepstrum coefficients the MA-part of the RASTA filter estimates the delta-cepstrum according to eq.(5.17), and the AR-part subtracts from this estimate an exponentially weighted geometric mean of the cepstral coefficient over the past analysis frames. That is, the delta-cepstrum is high-pass filtered.

The *liftering* described in section 5.3.1 weights the cepstrum coefficients within a cepstral vector for each time frame. *RASTA-filter applied on cepstrum* on the other hand, bandpass filter the time evolution of each cepstrum parameter.
Thus, **in the cepstral domain liftering and RASTA filtering supplement each other.**

Figure 5.4 illustrates that RASTA processing on mel-frequency based cepstral coefficients smooths the time evolution of each coefficient and makes its mean value closer to zero (the DC-component is removed). A typical example is the mean of the first MFCC for the first sentence of speaker AFN, which is -2.0194 without liftering, -5.1807 with liftering, and 0.0049 with RASTA(ρ=0.94) filtering.

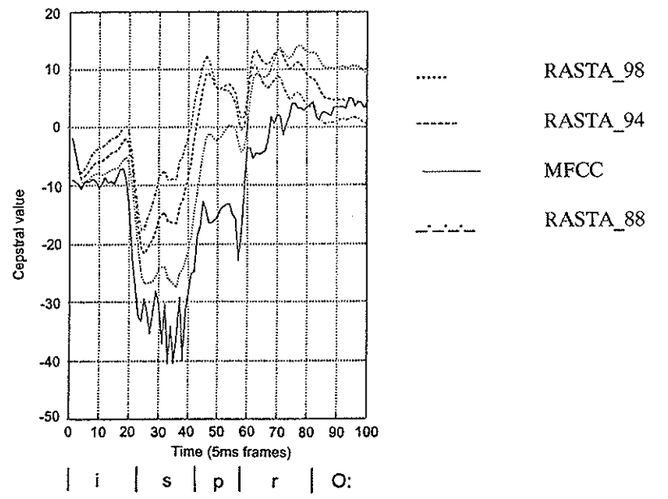


Figure 5.4 *RASTA filtering with ρ=0.98, 0.94, and 0.88 on the first Mel-scale cepstrum coefficient for each 5ms frame for the first 100 frames (i.e. 500 ms) of speaker AFN. (See waveform and broad band spectrogram in figure 4.12).*
In the bottom row the manual labelling is shown.

The first cepstrum coefficient, c(1), reflects the energy balance between high and low frequencies [O'Shaughnessy'90]. This fact is illustrated in figure 5.4 where c(1) has a higher value for the sonorants than for the fricative, with a rapid change in value at the transition between such sounds.
Although the RASTA-filtering smooths the time-evolution of cepstral coefficients, the changes at the phoneme boundaries are not altered.

Cumulative cepstral distortion

In order to explore the effect of RASTA-filtering MFCCs, we compared the *cumulative cepstral distortion* for ordinary MFCCs and RASTA-processed MFCCs. The cumulative averaged distortion $D(t)$ between cepstral vectors with time distance t was estimated from the 25 first sentences (about 7-10000 frames each of 5ms duration) for each speaker in the Norwegian EUROM0 as:

$$D(t) = \frac{1}{nfr}\sum_{k=1}^{nfr}\sum_{i=1}^{nc}(c_k(i)-c_{k+t}(i))^2 \qquad (5.19)$$

where *nfr* is the number of frames, *nc* is the number of coefficients in a cepstral frame vector

($nc=12$), and $c_k(i)$ is the *ith* coefficient within the cepstral vector for frame k.

The plot at the top in figure 5.5 shows that the cumulative cepstral distortion for the ordinary MFCCs first increases monotonically with *t*, but then flattens out. For speaker TGN this knee appears at about 140ms, for SHN at 175-200ms, and for AFN and TBN at about 225-275ms. This knee may be interpreted as "the point when, on average, the distance between compared spectral vectors exceeds the average length of the relatively steady part of speech" [Hermansky'92, p.88]. However, according to the average phoneme durations and the articulation rates for the Norwegian speakers listed in table B.2 in Appendix B, the steady parts of speech should be expected to be shorter, around 100ms.

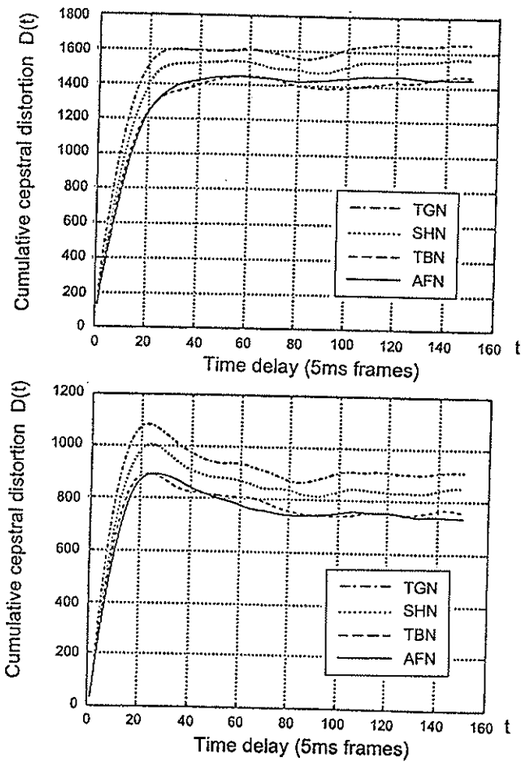


Figure 5.5 *Cumulative cepstral distortion, D(t), for the four speakers in Norwegian EUROM0, with ordinary MFCCs (top) and RASTA_94 processed MFCCs (bottom). The time delay t is given in 5ms steps.*
Note the different scale for distortion in the two plots.

The cumulative averaged distortion curves for the RASTA_94 processed coefficients are plotted in the bottom of figure 5.5 (ρ=0.94 corresponds to a time-constant $\tau\approx$175ms). Three differences between the two plots in figure 5.5 is noted. First, the maximum distortion is reached much **earlier** with RASTA-processed cepstrum, i.e. at about 130ms for AFN, 125ms for SHN, and 115ms for TGN and TBN. These values are closer to the expected duration of the stable portions of speech. Secondly, the maximum distortions are **smaller** than with ordinary MFCCs (different scales in the two plots). Finally, after the maximum is reached the cumulative distortion **decreases** before it flattens out.
Thus, the RASTA-processing of MFCCs may enhance boundaries between acoustically different segments of speech. If this observation is correct, RASTA-filtering of cepstrum coefficients may improve the automatic segmentation accuracy.

5.4 DYNAMIC PROGRAMMING

In order to perform phonemic segmentation automatically we have to discriminate between different phonemes (i.e. detect that we have encountered a new phoneme segment). It is therefore natural to employ techniques developed for ASR in automatic segmentation as well.
A **deterministic** approach to ASR is to compare the **t** parameter vectors of a test utterance **T** with a known reference pattern **R** of **r** frames by means of a pattern matching technique and a distortion measure. The utterances generally have unequal durations, and in different realisations of a word different sounds may be nonlinearly prolonged or shortened. A linear duration normalisation prior to the comparison will thus give poor correspondence between parameter vectors representing the same instances of a sound. The *nonlinear time normalisation* problem is to compute the warping function $m=g(n)$ which best maps the time axis n of **T** into the time axis m of **R**, by minimising the distance (distortion) along that path [O'Shaughnessy'90]:

$$D = \min_{g(n)} \left[\sum_{n=1}^{t} d(T(n), R(w(n))) \right] \tag{5.20}$$

where d(n,m) is the spectral distortion measure between test frame parameter vector n and reference frame parameter vector $m=g(n)$.
In principle, the optimal path is found by comparing all possible paths in the grid. However, with a *dynamic programming (DP)* technique called *Dynamic Time Warping (DTW)* the best single path is efficiently computed without evaluating all possible paths.

DP is a recursive procedure for solving a problem with interrelated variables. Generally, a single large V variable problem is by DP transformed to V simple problems which are less computation intensive [Cooper'81], [Bertsekas'87].
DP is based on the *principle of optimality* [Bellman'62] stated in [Cooper'81, p.9] as:
"An optimal policy has the property that whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision".
Thus, every optimal policy consists only of optimal subpolicies. If a point in the grid is on the optimal path, the optimal path goes from that point to the start and from the point to the end of the utterance.

With DTW the minimum accumulated distance $D_a(n,m)$ from the start point (1,1) to a point (n,m)

in the grid, is recursively computed as:

$$D_a(n,m) = d(T(n),R(m)) + \min_{k \le m}[D_a(n-1,k)] \quad (5.21)$$

This equation shows the main clue of DP; instead of an exhaustive search through all possible paths, only one path is kept and extended at each node. The DP procedure is optimal in the sense that the accumulated distortion in all discarded paths are always greater than in the path that is propagated, i.e. the *optimal path*.
By saving at each node the information about where the optimal path arrived from, the optimal state sequence is found by *backtracking* the optimal path recursively. See the Viterbi algorithm in section 5.5 below.

In order to optimise the nonlinear time normalisation between the test and reference utterance, the search space must be constrained by (i) limiting the maximum stretching and compression of a utterance (*global constraints*) to eliminate unrealistic paths, and (ii) limiting the set of allowed predecessors, e.g. $k \le m$ in eq. (5.21) (*local constraints*), to avoid that two words with the same phonemes but in different order, are regarded as similar.

5.5 HIDDEN MARKOV MODELLING

In **statistical** or **parametric** methods for ASR, the speech signal is modelled as generated by a stochastic process. One of the most widely used statistical method for ASR and ASS is the *hidden Markov model (HMM)* [Rabiner'89a]. HMMs can be used to model both the language and the speech signal. At the acoustic level HMMs can be regarded as a generalisation of the DP solution to the time normalisation problem, because the time normalisation is often less constrained, the local constraint function can be re-estimated iteratively, and the properties of the network are statistically optimised [Picone'90].

Regarding speech production as transitions between steady states or targets, the hidden Markov modelling is an intuitive representation of speech. An HMM consists of *states* with output probability distributions that model the parametric distribution of speech events, and the states are connected by *transition probabilities* that model the ordering and duration of the events.

If the Markov chain is in state s_i at time t, there is a conditional probability a_{ii}, of staying in that state, and a conditional probability a_{ij} for moving to state s_j, at time t+1. Denoting the state at time *t* by q_t the transition probability for a *first-order* HMM is $a_{ij}=P(q_{t+1}=s_j \mid q_t=s_i)$.

If the signal source produces just discrete symbols $\mathbf{V}=(v_1,v_2,v_3,...,v_M)$ or if the outputs are quantised to the finite symbol set **V**, the observation probability distribution in state s_j is discrete; $\beta_j(m)=P(v_m \text{ at } t \mid q_t=s_j)$, and the corresponding HMM is called *Discrete Density HMM (DDHMM)*.

If the multidimensional parameter vectors, e.g. cepstrum vectors, **O**, are used directly, a continuous multivariate observation probability distribution density (*pdf*) $\beta_j(\mathbf{O})$ in state s_j is estimated. For speech $\beta_j(\mathbf{O})$ is usually selected to be a multivariate Gaussian distribution $N(\mathbf{O}, u_j, \Lambda_j)$ with mean vector u_j and covariance matrix Λ_j. The dimension of the mean vector and covariance matrix equals the parameter vector. Use of Gaussian distributions is convenient

because it leads to simple formulas. Generally, in this *Continuous Density HMM (CDHMM)* the observation *pdf* can be modelled as an arbitrarily complex distribution by representing it as a weighted sum, or *mixture*, of Gaussian distributions:

$$\beta_j(\mathbf{O}) = \sum_{k=1}^{M} w_{jk} N(\mathbf{O},\mu_{jk},\Lambda_{jk}) \quad (5.22)$$

where w_{jk} is the mixture weights and $N(\ )$ is the multivariate Gaussian distribution with mean vector u_{jk} and covariance matrix Λ_{jk} in state s_j and mixture k.

In the above definition of HMM it is assumed that (1) the probability of being in a particular state of the Markov chain at time *t* depends only on the state at *t-1* (the *Markov assumption*), and (2) the outputs are not dependent on any previous outputs, but only on the underlying Markov chain (the *output-independence assumption*). For speech modelling, these assumptions are inappropriate because the realisation of speech sounds depends on the context.

Modelling speech with DDHMM, the multidimensional parameter vectors for each frame are classified by vector quantisation (VQ) [Linde'80] prior to estimation of the output *pdf*. Since no assumption is made about the underlying distribution of the observations, DDHMM can in principle model any distribution. Although VQ introduces quantisation errors, DDHMMs have given good ASR results, e.g. in the *Tangora* [Averbuch'86] and *SPHINX* [K.F.Lee'89] systems.

With CDHMM the VQ step is alleviated and the speech parameters are modelled directly, at the cost of increasing the number of parameters. Increased complexity requires more training and recognition time. A comparison between DDHMM and CDHMM applied for speaker-independent continuous speech recognition over the telephone lines [Fissore'91], concluded that CDHMM gave better results, but required more time. A simplified CDHMM with diagonal covariance matrix and few mixtures has achieved high ASR accuracy, e.g. in the *AT&T* [Rabiner'88] systems.
Since time consumption is not a major issue for automatic speech segmentation, CDHMM is selected in this thesis.

An HMM for a subword unit can be described compactly as $\lambda = \{\mathbf{A},\mathbf{B},\pi\}$, where $\mathbf{A}=\{a_{ij}\}$ is a N_xN matrix containing the state transition probabilities, $\mathbf{B}=\{\beta_j(\mathbf{O})\}$, j=1,N, is the output probability distribution matrix, and $\pi=\{\pi_i\}$, i=1,N, is the ensemble of initial state probabilities. A model λ can be regarded as the generator of the sequence of observations $\mathbf{O}=(O_1, O_2, O_3 ..., O_T)$, where O_t is the parameter vector, e.g. the cepstrum vector, for frame *t*. Then three problems have to be solved:

1) **Evaluation**, i.e. how well does a model λ correspond to an observation sequence **O**?
For instance, in isolated word recognition all word models λ are investigated to find the probability that the model generated the observation. The word model with highest score, i.e. max. $P(\mathbf{O}|\lambda)$, is chosen as the recognised word. Hence, all possible state sequences of length N should be evaluated, requiring in order of N^T computations[4]. Fortunately, the evaluation of HMMs can be solved by a recursive algorithm, the *Forward Procedure*, which requires only in order of N^2T calculations (see e.g. [Rabiner'93, p.335]).

[4] For instance, N=5 states and T=100 observations give more than $5^{100} \approx 10^{72}$ computations.

2) **Decoding**, i.e. what is the most likely state sequence in the model λ that produced the observations **O**? This optimal state sequence is exactly what is searched for in automatic segmentation, and also in ASR of continuous speech.
The *Viterbi algorithm* [Viterbi'67], called the stochastic version of the DP algorithm [Silverman'90, p.11], calculates the best single state sequence (path) efficiently.

3) **Training**, i.e. based on many observations of the same phenomena Ô and a model λ with rather arbitrary parameters, how can the model parameters be tuned to maximise $P(\hat{O}|\lambda)$?
The *Baum-Welsh re-estimation procedure* [Rabiner'89a] computes iteratively the maximum likelihood estimate of the parameters in the model. Alternatively, the *segmental k-means procedure* [Rabiner'86] can be used for training.

The **Viterbi algorithm** applied within a single HMM to calculate the optimal state sequence, $\mathbf{Q}=(q_1,q_2,q_3,...,q_T)$ for a given observation sequence $\mathbf{O}=(O_1,O_2,O_3,...,O_T)$, can be summarized as follows [Rabiner'89a, p.264]:

1) *Initialisation (t=1)*:

$$\begin{aligned} \phi_1(i) &= \pi_i \beta_i(O_1) \quad , \quad 1 \le i \le N \\ \psi_1(i) &= 0 \end{aligned} \tag{5.23}$$

2) *Recursion*:

$$\phi_t(j) = \max_{1 \le i \le N} \left[\phi_{t-1}(i) a_{ij}\right] \beta_j(O_t) \quad , \quad \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{5.24}$$

$$\psi_t(j) = \underset{1 \le i \le N}{argmax} \left[\phi_{t-1}(i) a_{ij}\right] \quad , \quad \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{5.25}$$

3) *Termination*:

$$\begin{aligned} \phi^* &= \max_{1 \le i \le N} \left[\phi_T(i)\right] \\ q_T^* &= \underset{1 \le i \le N}{argmax} \left[\phi_T(i)\right] \end{aligned} \tag{5.26}$$

4) *Path (state sequence) backtracking*:

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \quad , \quad t = T-1, T-2, \cdots, 1 \tag{5.27}$$

where $\varphi_t(i)$ is the best score (highest accumulated probability) for being in state s_i at time t. The array $\psi_t(j)$ stores the argument values which maximizes eq.(5.25) for each t and j, and the optimal state sequence is given by the recursion in eq. (5.27).

The Viterbi algorithm is similar to the deterministic DP procedure except for the multiplications instead of additions in the recursion. Since products can lead to underflow, the logarithm is used in the recursion step, transforming eq. (5.24) and eq. (5.25) to:

$$\phi_t(j) = \max_{1 \le i \le N} \left[\phi_{t-1}(i) + \log(a_{ij})\right] + \log(\beta_j(O_t)) \quad , \quad \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{5.28}$$

$$\psi_t(j) = \underset{1 \le i \le N}{argmax} \left[\phi_{t-1}(i) + \log(a_{ij})\right] \quad , \quad \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{5.29}$$

The output log probability for a single Gaussian distribution is:

$$\log(\beta_j(O)) = \frac{1}{2}(O-\mu_j)\Lambda_j^{-1}(O-\mu_j)^T - \frac{1}{2}(\log(2\pi^p|\Lambda_j|) \tag{5.30}$$

where p is the dimension of the observation vectors and *T* denotes the transposed of a vector (or a matrix). Actually, eq. (5.30) computes the Mahalanobis distance [Duda'73] between two parameter vectors after the input parameter vector has been transformed onto an orthogonal space in which each dimension has equal weight in the distance measure. This transformation decorrelates elements of the parameter vector and equalizes the variances of each dimension, which is important when using different kinds of parameters in a parameter vector (e.g. mixing cepstrum, delta cepstrum and energy) [Picone'90].

The main idea of DP is seen in eq. (5.24), where the only path propagated is the most probable path selected from all possible paths that can make a transition to current state at time t.
In order to compute more efficiently the search space can be reduced by *Viterbi Beam Search* [Picone'90, p.31] or *Level-Building* [Rabiner'89a, p.282] techniques (see also [Rabiner'93]).

5.6 SUMMARY

Frequently used techniques for speech modelling and automatic speech recognition have been described. Cepstral representations of the speech spectrum are convenient because the spectral envelope can be easily separated from the spectral fine structure and the difference between two speech spectra can be calculated simply as the squared difference between the corresponding cepstrum coefficients. Both parametric and non-parametric cepstral representations have been outlined. A cepstral domain filtering technique is proposed and argued for with respect to automatic segmentation. Comparison of cumulative cepstral distortion for ordinary MFCCs and filtered MFCCs indicated that the filtering may enhance the border between acoustically different speech segments.
The dynamic programming technique described will be utilised in the automatic acoustic segmentation, both for segmentation and for evaluation, and the hidden Markov modelling technique and the Viterbi algorithm will be employed for automatic phonemic segmentation.

Chapter 6

AUTOMATIC SEGMENTATION OF SPEECH

This chapter deals with automatic segmentation of speech (ASS). First an overview of different acoustic and phonemic ASS algorithms is given, then our approach to ASS is described in detail, and finally different possible evaluation procedures of ASS methods are discussed.

6.1 AUTOMATIC SEGMENTATION OF SPEECH - AN OVERVIEW

In this overview two main categories of automatic segmentation algorithms will be discussed: (1) **Acoustic** segmentation and (2) **Phonemic** segmentation.

6.1.1 ACOUSTIC SEGMENTATION

Acoustic segmentation is here used as a cover term for ASS algorithms which rely on the acoustics only, i.e. they do not assume any phonological information. The only a priori knowledge provided is the distortion measure for computing the difference between the acoustic parameters. These methods are thus language independent and are often used as front-end for the phonemic segmentation.

The acoustic segmentation maps the continuous speech signal into some discrete entities, such as acoustic segments, acoustic subwords, acoustic-phonetic features, or broad phonetic classes. Since a phonemically defined unit may contain many spectral homogenous or quasi-stable areas, the acoustic segmentation algorithms often provide an *oversegmentation* (o.s), i.e. more segments than phonemic labels.

The acoustic segments can be categorized as transient or steady. A *transient segment* starts when the spectral change exceeds a predefined transient onset threshold and ends where the spectral change drops below a threshold. For certain thresholds the transient segments may correspond to the uncertainty area for manual boundary placements defined in chapter 4 and/or to the transition segments defined in figure 1.2. A *steady segment* is the sequence of frames between the transient segments. The spectral shape will of course change within a steady segment also, but not as much as the relative change between the segments.

By tuning the thresholds properly the spectral variation within the steady segments can be assumed to be so small that each segment can be approximately represented by a constant value, called the *generalised centroid*. The generalised centroid of a vector sequence is the vector which represents the sequence with minimum distortion. The value of the centroid thus depends on the

distortion measure used. With e.g. the squared error measure the centroid is the mean value of the vectors in the segment. In acoustic segmentation algorithms the distance between the centroid of a segment and each frame within the segment is often used as a measure of the intrasegment distortion.

The acoustic segmentation algorithms proposed in the literature can be grouped into three main classes: unconstrained acoustic segmentation, acoustic subword segmentation, and acoustic-phonetic segmentation. In the following subsections these algorithms are discussed.

6.1.1.1 Unconstrained acoustic segmentation

By unconstrained acoustic segmentation is meant that no segment categorization is performed. The main difference between the unconstrained acoustic segmentation methods is whether the calculation of segment boundaries is based on information from the whole sentence, *Global optimisation*, or on information within in a smaller part of the sentence, *local optimisation*.

a) Global optimisation

In global optimisation methods the consecutive segments which yield minimum overall intra-segment distortions are searched for. The obtained segments will exhibit the maximum acoustic homogeneity within their boundaries and the frames within a segment are highly correlated, i.e. steady segments are located.

Optimisation based on the whole utterance may introduce an unacceptable delay if the acoustic segmentation is integrated as a front-end in a real-time continuous speech recogniser. For segmentation of speech databases this delay has no practical consequence.
The number of segments can be predefined as a number of acoustic segments per time unit or the segmentation process can terminate when e.g. the average intra-segment distortion is less than a certain threshold.

Bridle [Bridle'77], recursively computed all possible segment combinations and represented each segment by its centroid, i.e. its mean spectrum, and the segment sequence that minimised the sum of squares of the differences between the spectral frame vectors and the centroid within each segment was found. Bridle also proposed a straight line approximation of the segments. This gave twice as many degrees of freedom for each segment in the optimisation. The boundaries computed by this linear representation of segments were often placed in the middle of the segments approximated by its centroid only.

Bridle's ideas were reformulated in the Constrained Clustering Vector Quantization Segmentation (CCS) [Svendsen'87], where the segmentation problem was viewed as designing a VQ codebook subject to the constraint that all the vectors contained in a cluster had to be contiguous in time.

In a special case of CCS it is assumed that all frames within a steady segment have been generated by the same stochastic model, i.e. an autoregressive (AR) model. The problem now is to estimate in a maximum likelihood statistical framework, the parameters of the source of the input. In this approach, denoted Maximum Likelihood Segmentation (MLS) [Bridle'77], [Svendsen'87], the optimal boundaries are found by minimizing the overall likelihood ratio

distortion (MLR) for all possible segment combinations of the utterance. However, the CCS algorithm can be used with any distortion measure, resulting in a segmentation which is not optimal in the maximum likelihood sense.

The acoustic segmentation may be regarded as a transformation of a continuous signal onto a discrete vector sequence. In [Brown'89] a Karhunen Loeve Transform (KLT) [Jayant'84], based on global signal statistics is used to find the frame sequences with high correlation. The KLT of a segment is the transformation matrix of the eigenvectors of the segment covariance matrix.

b) Local optimisation

By using a distortion measure based on local information only, a segment boundary is marked whenever a spectral change larger than a predefined threshold is detected. Several methods for measuring the spectral change are proposed, e.g. as correlation between consecutive frames [Hemert'87], as "spectral variation contour" [Wilpon'87], as $|O_i-O_{i-1}|^2$ [Poddar'90], or as $(\frac{\partial O_i}{\partial i})^2$ [Svendsen'87], where O_i is the observation vector for frame i.
The segment boundaries are thus obtained almost instantaneously. The motivation for this approach is that a rapid spectral change is believed to correspond to the transition between two phones.

The main problem with the methods based on local optimising is to tune the thresholds and parameters to balance the trade-off between noisy estimates (giving too many distortion peaks), and smoothed estimates (resulting in too few distortion peaks). If too many acoustic segments are obtained a post-processing step which reduces the number of acoustic segments can be introduced. For instance in [Poddar'90] consecutive acoustic segments were merged if the distance between their centroids was less than a predefined threshold, in [Hemert'87] the final boundary was marked when the difference between the centroid and a spectral vector within the preliminary segment changed from negative to positive, and in [Wilpon'87] a matched filter and a smoother were applied.

Another possible approach can be related to the global MLS in that each steady segment is modelled as an AR-process. In [Andre-Obrecht'88] different types of test statistics were used to perform an on-line detection of changes in the parameter vector of the model starting from the location of the previously detected boundary.

In [Algazi'88] maximum likelihood ratio (MLR) tests were applied iteratively to locate spectral changes. For each five-frame group the two first frames were compared with the two last frames. The boundaries were detected at the maxima in this MLR distortion.

In this section we will also briefly describe multilevel segmentation, although it does not fulfil the sequential segmentation definition in this thesis. Instead of choosing one single consecutive segment sequence depending on e.g. the threshold of the distortion peaks, different segmentation alternatives with varying degrees of acoustic similarity are found and stored. These alternatives can be utilized in a front-end for e.g. manual or automatic phonemic segmentation procedures. The multilevel segmentation algorithm proposed in [Glass'88] uses a hierarchical agglomerative clustering procedure [Duda'73, p.229] which is initialised by defining each frame as a segment. Two segments are joined if the actual segment is more similar to the preceding adjacent segment

than the following segment. Each segment is represented by its centroid, and the clustering is repeated until the whole utterance is one segment.

The result of this clustering procedure can be represented by a dendrogram [Duda'73, p.229], where both the boundaries and the distortions at which the segments are merged, are depicted.

6.1.1.2 Acoustic subword segmentation

In order to utilize the acoustic segmentation in speech coding or in ASR, each acoustic segment should be associated with a subword, where the number of acoustic subwords to be found is much less than the number of segments found by the unconstrained acoustic segmentation methods. The subwords can be obtained by post-processing the acoustic segments through a standard clustering method or by imposing the segmentation algorithm to search for acoustic subwords.

a) Clustering the acoustic segments

In [Wilpon'87], 1600 acoustic segments obtained by local optimising were clustered into 25 acoustic subwords using the modified k-means clustering algorithm [Wilpon'85]. Each acoustic subword was represented by the mean value of the segments within the cluster. In a listening test of these acoustic subwords, 17 distinct vowel-like sounds were perceived. The remaining 8 sounds were impossible to characterise. Hence, there may be some underlying structure to the subword segments. By representing the words in terms of concatenated acoustic subwords the ASR was based on level building dynamic programming.

[C.H.Lee'88] and [Svendsen'89] clustered the acoustic segments obtained by global MLS by representing each segment with its centroid and used the LBG-algorithm [Linde'80] to compute a codebook of N codevectors that best represented the centroides. A large segment codebook provides good spectral resolution whereas a small codebook increases the reliability of the statistical segment characterization. In the ASR step an HMM for each acoustic subword was used.

b) Constrained acoustic segmentation

For speech coding at very low bit rates variable-length segment quantization is proposed, e.g. [Roucos'82], [Shiraki'88]. Instead of first segmenting the utterance and then quantizing the segments, a joint segmentation and quantization approach gives less overall quantization error. The joint segmentation and quantization method is similar to the CCS-algorithm except that now both the segment boundaries and a set of acoustic subword models are searched for simultaneously.
Different types of models, such as HMM or stochastic segment models [Roucos'87], [Ostendorf'89] can be used for the subwords.

6.1.1.3 Acoustic-phonetic segmentation

By acoustic-phonetic segmentation is meant a segmentation where the segments are categorised into more or less language independent broad phonetic classes.

In [Leung'84] a sequence of binary classifiers were arranged in a binary decision tree. At the first node each frame was assigned to one of two broad-class labels by k-means clustering on the whole utterance represented by one set of features. Next, each class was divided into two by k-means clustering in a different feature space, where the features were selected so as to maximise the wanted class difference at each node. The best result was obtained when each frame was assigned one of six broad classes; vowel-like sonorant, obstruent, voice-obstruent, silence, nasal and voice-bar, and unlabelled segment. A context-dependent median smoother was used to remove spurious segments.

The segmentation and the broad classification in [Cole'88] were based on empirically derived rules applied to waveform parameters (zerocrossing and peak-to-peak amplitude at different frequency slots) and spectral change parameters (distance between spectra at different time instances). Cole applied a top down parsing strategy where the rules produce a successive refinement of the preceding segmentation.
A general problem for the rule-based systems is that they may be difficult to modify.

Within the SAM project a French method called FR_SALA [Dours'89], and a Danish method, DK_SALA [Dalsgaard'90], were compared with our system, see [Barry'91b]. These systems are two step algorithms. The first step is described here and the second step is described in section 6.1.2.2.
The first step in FR_SALA is a knowledge based vector quantization which assigns a (language-independent) phonetic label to each frame. Consecutive frames with the same label are merged, and too short segments are removed. With 11 phonetic labels, approximately 200% oversegmentation was obtained.
The first step of DK_SALA applies a Kohonen Self-Organising Neural Network [Kohonen'90] to label each frame with continuously valued acoustic-phonetic distinctive features (i.e. language-dependent features).

Trained artificial neural networks (ANNs) have also been used to assign frames to e.g. "quasi-phonemic" labels [Torkkola'88], or broad phonetic classes [Depuydt'91], whereas in e.g. [Poddar'90], the ANN classifies acoustic segments into ten broad phonetic classes, and then merge adjacent segments which have been given the same label.

Instead of segmenting speech as defined in this thesis, the speech can be decomposed in terms of overlapping events characterized by spectral target functions, or segment models, weighted by time-limited interpolation functions, or boundary functions (see e.g. top line in figure 1.2). Atal [Atal'83] suggested a temporal decomposition technique for speech coding, which has got a phonetic interpretation later, e.g. [Marcus'84],[Bimbot'88],[Dijk-Kappers'89].
For a given utterance the optimal boundary functions and their associated targets have to be found simultaneously. A segment may only overlap with the nearest segment on each side, or multiple overlapping can be allowed. For speech coding fixed length overlap has shown to give least spectral distortion [Honda'92].

6.1.2 PHONEMIC SEGMENTATION

By phonemic segmentation we mean a segmentation constrained by a phonemic transcription of the utterance, and the task is to align this transcription with the signal. That is, the number, the identity, and the order of the phonemic units are known a-priori. This can be regarded as a particularization of the constrained acoustic subword segmentation, section 6.2.1.2.b. However, the segment distortion can more easily be computed from its corresponding phonemic model instead of having to search throughout the whole set of models for the model which gives minimum distortion.

The various phonemic segmentation algorithms differ in the choice of phonemic units, how to model the chosen unit, and whether an acoustic segmentation algorithm is utilized (two step algorithms) or not (one step algorithms).
The different subword units were discussed in chapter 3. Here, one and two step algorithms for segmentation into phonemes or phoneme-like units are described.

6.1.2.1 One step algorithms

Phonemic segmentation can be performed within the HMM framework by constraining the grammar network in an HMM phoneme recogniser to only recognise the given label string. The phoneme boundaries are provided as the time instances for the model transitions by tracing back the optimal path found by the Viterbi algorithm. This will be more elaborated in the discussion of our method in section 6.2.2.
The segmental k-means training procedure [Rabiner'86], often used for initial estimates of the model parameters in the HMMs, performs also segmentation via Viterbi decoding.

A more sophisticated method was proposed in [Ljolje'91], employing an HMM structure consisting of three independent models: a trigram phonotactic model, a phone duration model, and an acoustic phone model. These models were interrelated through a structure similar to a second order ergodic continuous variable duration HMM (CVDHMM) [Levinson'86]. Many contexts give the same coarticulation effects on a phone. Such contexts were clustered in order to increase the training material for context dependent acoustic models. Also, quasi-triphonic models exploiting that contextual influences often only span across some part of the phone, were introduced. The most likely phone sequence and its boundaries were found by Viterbi decoding.

An utterance with known transcription can be segmented by a DTW-comparison with synthetic speech based on samples from the actual speech, e.g. [Hemert'87],[Ottesen'91]. If many similar utterances are to be segmented, DTW between a manually segmented reference utterance and another similar utterance can be used for segmentation, e.g. [Rabiner'82], [Haltsonen'85].

6.1.2.2 Two step algorithms

The acoustic and phonemic segmentation approaches are in a sense complementary, as shown in table 6.1 below. It is therefore tempting to combine them by retaining the advantages and avoiding the disadvantages of each of them.

In the second step described here the acoustic segments are aligned with the transcription by merging segments in cases of oversegmentation or by splitting and refining segments when too few segment boundaries are calculated in the first step. The methods differ in whether the acoustic segmentation constrains the phonemic segmentation or the two types of segmentations are performed independently and then combined in a postprocessing stage.

Method	Advantages	Disadvantages
Acoustic segmentation	Language independent No training is needed Accurate determination of boundaries	Wrong numbers of boundaries Phonetic identification unknown
Phonemic segmentation	Correct number of boundaries Phonetic identification known	Language specific Training or rules needed Inaccurate boundaries

Table 6.1 *Some pros and cons for acoustic and phonemic segmentation.*

a) Constraining the phonemic segmentation by the acoustic segment boundaries

When constraining the phonemic segmentation by the acoustic segment boundaries, the final segmentation result cannot be more accurate than the acoustic segmentation. Our algorithm (see section 6.2) is a typical example of this approach. We model each phoneme in a language with an HMM and the phoneme boundaries are obtained by constraining the Viterbi forward recursion by the acoustic segment boundaries and the transcription.

If the first step is acoustic-phonetic, the second step has to align the given transcription with the broad-class label string, often by a DP optimising. In [Leung'84] knowledge-based DP is used for this alignment, i.e. the paths in DP are constrained by rules such as durational rules, phonotactics, and which phonetic label can be mapped to which class label. If one broad class segment is mapped to two or more phonetic events, further segmentation is achieved by applying a set of heuristic rules.
The system in [Torkkola'88] is in principle similar to this approach, whereas FR_SALA [Dours'89], applies automatically derived rules to transform the transcription to broad-class labels prior to DP.

For the multilevel segmentation approach the path through the dendogram which best matches the transcription gives the phonemic segment boundaries [Glass'88].

b) Combining the independently obtained acoustic and phonemic segment boundaries

In [Hemert'87] the acoustic boundaries closest to the phonemic ones are linked with the phonemic boundary so that no linking line crosses. These acoustic boundaries are kept together with the phonemic boundaries not linked to any acoustic boundary, whereas not linked acoustic boundaries are omitted.

In DK_SALA [Dalsgaard'90] the continuously valued acoustic-phonetic distinctive features are further transformed into a smaller set of Principal Components, which are used to model

individual allophones in a multivariate Gaussian probability density function. One HMM for each phoneme based on acoustic-phonetic features is used. The alignment is performed by a Viterbi decoding based on the sequence of feature vectors and one-pass level-building constrained by the label string.

6.2 OUR ALGORITHM FOR AUTOMATIC SEGMENTATION

We aimed at developing an automatic multilingual segmentation algorithm which could place phoneme boundaries with accuracy comparable to the performance of human labellers for any language. For this task an acoustic segmentation algorithm is attractive because it is language-independent and requires no training. Since we assume that the label string is known, the acoustic segmentation algorithm can be forced to segment exactly the wanted number of segments and the labels can be assigned successively to these segments. However, although the CCS-algorithm provides accurate segmentation when oversegmentation is allowed for, the accuracy becomes low when the number of acoustic segments is forced to be equal to the number of phonemes (see experimental results in section 7.2.4).
Thus, knowledge had to be incorporated in the automatic segmentation algorithm. For the given task we found a rule based system which explicitly incorporates language specific phonetic and phonological knowledge too labour intensive in that many rules are needed and each language requires its own set of rules. In addition, the limited set of known rules does not cover all segmentation problems encountered and new rules may be troublesome to incorporate in the system. Therefore, an algorithm which automatically learns the language specific characteristics from a limited annotated speech material was preferred. Since remarkably good ASR performances have been achieved with the HMM technique, we chose the HMM framework for phoneme modelling and phonemic segmentation.

Experiments with pure HMM phonemic segmentation yielded better accuracy than with the acoustic segmentation with no oversegmentation, but it was still too inaccurate. As discussed in section 6.1.2.2, one may utilise the advantages of an acoustic segmentation to refine the segmentation accuracy of the phonemic segmentation. Qualitative analyses of the acoustic segment boundaries calculated by the CCS algorithm (section 7.2.3), show that many boundaries coincide with manual segmentation and the acoustic segments can be given a meaningful interpretation.
We[1] thus proposed *a two step algorithm which uses the language independent acoustic segment boundaries to constrain the statistically based phonemic segmentation.* This approach provides segment boundaries for the phonemic labels as well as a subdivision of these phonemic segments into acoustically stable or quasi-homogenous segments.

The dataflow of our automatic segmentation algorithm is shown in figure 6.1 below. The *transcription files* contain the phoneme label string and the endpoints for each utterance in the *recorded speech* files. The sampled speech waveform and the corresponding label string is fed into the *acoustic segmentation* one sentence at the time. The number of acoustic segments to be computed is decided by the number of phonemes in the sentence and the predefined oversegmentation-factor. The acoustic segmentation performs preprocessing and stores the frame parameter vectors and computes the acoustic segment boundaries. To save computation, the

[1] This ASS algorithm was developed together with Torbjørn Svendsen.

phonemic segmentation can use the frame parameter vectors computed in the acoustic segmentation. (If the phonemic segmentation shall estimate the frame parameter vectors, an arrow should be drawn from the "recorded speech" to the "phonemic segmentation" box in figure 6.1).

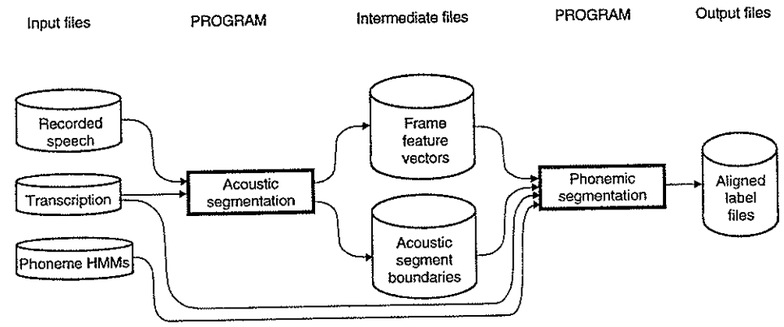


Figure 6.1 *Dataflow in our two step ASS algorithm. See text for details.*

In order to test the different segmentation methods separately, we implemented the acoustic and phonemic segmentation algorithms as two independent blocks. The phonemic segmentation part can thus either perform *pure* HMM segmentation as a separate part, or it can perform *constrained* HMM segmentation where the acoustic segment boundaries constrain the search space, in an integrated system. Acoustic segmentation results with different oversegmentation factors can be stored for later tests with the constrained HMM segmentation.
In the following the acoustic and phonemic segmentation steps are discussed in more detail.

6.2.1 ACOUSTIC SEGMENTATION

Although the main goal is automatic phonemic segmentation, it is important for many research purposes to provide an annotation which is closer to the acoustic-phonetic realisation of the utterance than the phonemic SAMPA labels. Since the acoustic segmentation is based purely on signal processing and the degree of oversegmentation can be more than 100%, the acoustic segmentation can be even more detailed than many narrow phonetic annotations. In chapter 7 a phonetic interpretation of the acoustic segments is given. The acoustic segmentation part is thus by itself interesting for research, see also [Barry'90b], and for constraining the search space in the HMM segmentation.

Since it has been proved difficult to find reliable methods for instantaneous segmentation based on local information, the global optimising CCS approach has been chosen in this thesis. The problem of dividing a speech signal into non-overlapping, consecutive acoustically stable segments can be formally described within the CCS framework as follows:

An utterance of T frames is represented by the spectral parameter vectors $\mathbf{O}=\{O_1,O_2,...,O_T\}$, and is to be segmented into N consecutive segments with boundaries $\{b_0,b_1,...,b_N\}$, where the first, $b_0=0$, and the last, $b_N=T$, is given to the algorithm. The optimal segmentation of this utterance maximizes the acoustic similarity of the speech frames within the segments. We thus seek for the set of boundaries, $\{b_i\}$, i=1,N, which minimizes the minimum accumulated error, $D_a(N,T)$, when using N segments to represent the T frames:

$$D_a(N,T) = \sum_{i=0}^{N-1} \sum_{n=b_i+1}^{b_{i+1}} d(O_n,\xi_i) = \sum_{i=0}^{N-1} D_s(b_i+1,b_{i+1}) \qquad (6.1)$$

where $d(n,i)$ is the spectral distortion measure, ξ_i is the generalized centroid of the ith segment from frame b_i+1 to frame b_{i+1}, and $D_s(b_i+1,\ b_{i+1})$ is the corresponding segment distortion.
With the *cepstral distortion measure* the centroid is defined as the mean vector of the parameter vectors within the segment:

$$\xi_i = \frac{1}{b_{i+1}-b_i} \sum_{n=b_i+1}^{b_{i+1}} O_n \qquad (6.2)$$

and the segment distortion within the i'th segment $D_s(b_i+1,\ b_{i+1})$ can be computed as:

$$D_s(b_i+1,b_{i+1}) = \sum_{n=b_i+1}^{b_{i+1}} [O_n-\xi_i]^T[O_n-\xi_i] = \sum_{n=b_i+1}^{b_{i+1}} O_n^T O_n - 2\xi_i^T \sum_{n=b_i+1}^{b_{i+1}} O_n + \sum_{n=b_i+1}^{b_{i+1}} \xi_i^T \xi_i \qquad (6.3)$$

Substituting ξ_i from eq. (6.2) into eq. (6.3) the segment distortion matrix is efficiently computed as:

$$D_s(b_i+1,b_{i+1}) = \sum_{n=b_i+1}^{b_{i+1}} O_n^T O_n - \frac{1}{b_{i+1}-b_i}\left(\sum_{n=b_i+1}^{b_{i+1}} O_n\right)^T\left(\sum_{n=b_i+1}^{b_{i+1}} O_n\right) \qquad (6.4)$$

The optimal acoustic segmentation is found by evaluating eq. (6.1) for all possible combinations of the segment boundaries. With DP the minimum accumulated distortion $D_a(i+1,\ b_{i+1})$ for the $i+1$ segment solution of the sequence of frames from 1 to b_{i+1} is efficiently calculated as:

$$D_a(i+1,b_{i+1}) = \min_{b_{i+1}} \{ D_a(i,b_i) + D_s(b_i+1,b_{i+1}) \} \qquad (6.5)$$

for all possible b_{i+1}. Thus, the optimum $i+1$ segment solution to frame b_{i+1} is found as the optimum i segment solution to the frame b_i plus the $i+1$'th segment from frame b_i+1 to frame b_{i+1}.

This is done by first computing the segment distortion matrix for all possible segment boundaries,

$$\mathbf{D}_s = \{ D_s(i,j) \} \qquad ; \quad \begin{matrix} 1 \le i \le N \\ 1 \le j \le T \end{matrix} \qquad (6.6)$$

and then initialising the recursive computation of D_a by the single segment solutions,

$$D_a(1,b_1) = D_s(1,b_1) \qquad ; \quad 1 \le b_1 \le T \qquad (6.7)$$

The DP search in eq.(6.5) finds the minimum accumulated distortion for the N-level segmentation as $D_a(N,T)$. The optimal acoustic segmentation boundaries are found by backtracking the DP grid along the optimal path stored in the backpointer array.

Figure 6.2 shows the distortion between each frame parameter vector and the centroid for the segment the frame belongs to. At the bottom row the frame-to-centroid distortions for the 5-segment solution of the word / n}l / (=zero) are depicted, and at the top row the corresponding distortions for the 6-segment solution are shown.

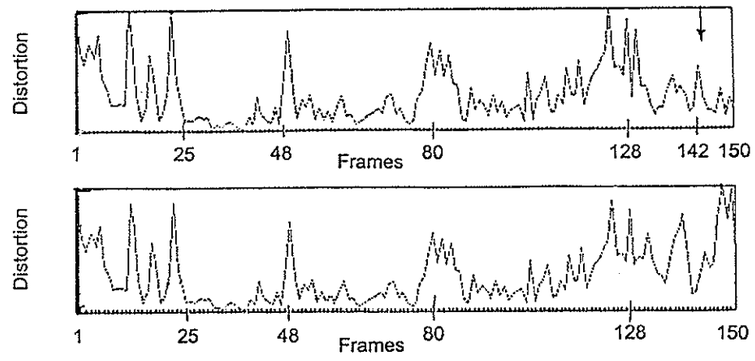


Figure 6.2 *Frame-to-centroid distortions calculated in the acoustic segmentation of the word /n}l/ (=zero), when 5 (bottom row) and 6 (top row) acoustic boundaries are placed. (After [Kvale'87]).*

From figure 6.2 it can be seen that the boundaries are placed at frames with big frame-to-centroid distortion. When increasing the number of segments the segment with most intra-segment distortion is split, whereas the other segment boundaries remain unchanged (see section 7.2.4 for details).

6.2.2 PHONEMIC SEGMENTATION

In this section HMM based phonemic segmentation is described. In order to perform label alignment with HMMs, the grammar has to be the given label string only. The phoneme HMMs were thus concatenated according to the given transcription in order to obtain an HMM for the entire utterance, as shown in figure 6.3 below.
To account for phoneme context a three-state phoneme model was applied. The phoneme context often causes a rather noisy estimate in the first and last state of a phoneme. Since we wanted to use the same configuration for different phonemes, we introduced a skip-transition between the first and the last state in the phoneme model.

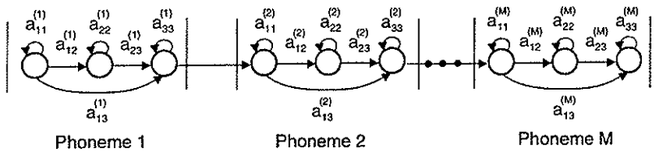


Figure 6.3 *Concatenation of M phoneme HMMs, where each phoneme is modelled by a left-to-right three-state CDHMM with skip-transition from state 1 to 3. The transition probabilities are indicated and the observation probability function in each state was a single continuous Gaussian with a diagonal covariance matrix. Phonemic segmentation based on this model is called pure HMM segmentation.*

As shown in figure 6.3, transitions between phonemes were only allowed from the last state in one phoneme model to the first state in the following phoneme model. The time-instances for jumping from one phoneme-model to the next are the desired phoneme segment boundaries, and these are made explicit by tracing back the optimal path found by the Viterbi algorithm for the whole utterance. This phonemic segmentation will be denoted **pure HMM segmentation** or **pure Viterbi segmentation**, abbreviated **V**.

In order to take advantage of the segmental information provided by the accurate acoustic segmentation, the acoustic segment boundaries were used as constraints of the standard Viterbi algorithm, in that a transition from one phoneme model to the next was only allowed for at the time instances specified by the acoustic segment boundaries $\{b_0, b_1, ..., b_N\}$, as illustrated in figure 6.4. A transition within a single phoneme model was allowed for at any time instant. Thus the forward recursion step of the Viterbi decoding eq. (5.25) for the phoneme model m with $J_{max}(m)$ states became:

$$\phi_t(i,m) = \begin{cases} \max[\phi_{t-1}(1,m)\,a_{11}^{(m)}\ ,\ \phi_{t-1}(J_{max}(m-1),m-1)][\sum_{n=1}^{N-1}\delta(t-b_n)]\beta_i^{(m)}(O_t) & ;\ i=1 \\ \max_{1\le j\le i}\ [\phi_{t-1}(j,m)\,a_{ji}^{(m)}]\beta_i^{(m)}(O_t) & ;\ i=2,J_{max}(m) \end{cases} \tag{6.8}$$

where the delta function is defined as

$$\delta(t-b_n) = \begin{cases} 1 & for\ \ t=b_n \\ 0 & otherwise \end{cases} \tag{6.9}$$

and the other symbols in (6.8) are as defined in chapter 5, augmented by the phoneme model index m, i.e.:
$\varphi_t(i,m)$ is the accumulated probability of being in state i of the phoneme model m at time t
$a_{ij}^{(m)}$ is the probability of making a transition from state i to state j for model m
$\beta_i^{(m)}(O_t)$ is the probability of observing the spectral vector O_t in state i of model m.

That is, with M phonemes in an utterance the optimal path is forced to go through M of the N points in the grid defined by the acoustic segmentation. Between each of these points the stored path is optimal, but the total path found by this approach will deviate from the optimal path computed by Viterbi algorithm without constraints. This phonemic segmentation will be denoted **constrained HMM segmentation** or **constrained Viterbi segmentation**, abbreviated **VC**.

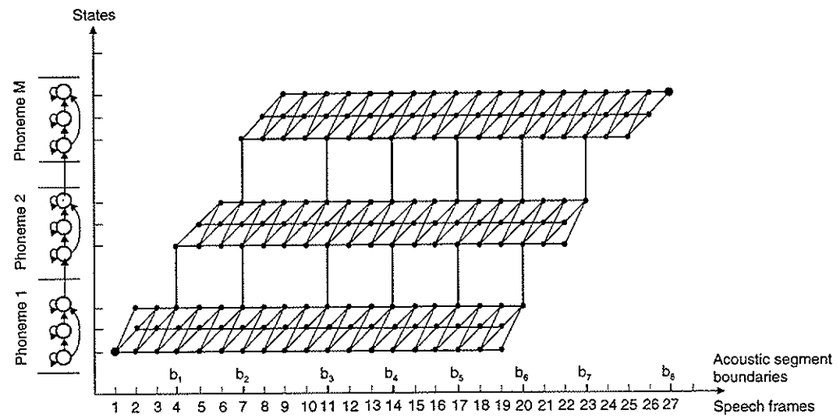


Figure 6.4 *Possible paths in the constrained Viterbi decoding. The utterance consists of M=3 phonemes and 8 acoustic segment boundaries are computed in the acoustic segmentation part.*

6.3 EVALUATION OF AUTOMATIC SEGMENTATION

Due to the lack of an ideal, correct reference it is difficult to quantify the performance of either human or automatic speech annotation. Usually the ASS algorithms are assessed in terms of how they affect the final performance of ASR or TTS or how well the segmentation coincides with some manual annotation of the same speech material.

6.3.1 USE IN REAL APPLICATIONS

For use in speech technology such as for training subword based ASR-systems or generating subword libraries for TTS-systems, the different ASS algorithms can be assessed by evaluating the influence the ASS has on the final performance of the ASR or TTS-systems.
However, when comparing different ASS methods in terms of e.g. intelligibility and naturalness of a TTS-system, one must be aware that other factors such as how the subwords are concatenated and the nature of the sound producing equipment also affect the resulting synthetic speech.
For many ASR systems the initial segmentation of the training material does not have to be very accurate because deviations are corrected through iterative bootstrapping procedures. Absolute performance in terms of coincidence with a manual reference may have little significance for ASR, because recognition rates depend not only of how well templates represent the training material, but also e.g. of how well the training material is representative of the test material.

6.3.2 COMPARISON WITH MANUAL SEGMENTATION

In this thesis work the objective is to annotate continuous speech for multi-purpose use. Calculating meaningful success rate figures for such segmentation is not easy. An evaluation is usually performed in terms of a comparison with a reference annotation produced by a trained phonetician with all knowledge sources available. *An error in automatic segmentation is then defined as any deviation from the manual segmentation.*
As pointed out in chapter 4, such an evaluation is problematic for several reasons. Firstly, there exist very few explicit rules for manual annotation. The rules are not standardized, and even for one person it is difficult to apply the rules consistently. Secondly, the labeller often sub-consciously uses other sources of knowledge than the acoustic signal; e.g. phonemic reconstruction, expectation of what should have been said, and phonological rules. Finally, there is a large uncertainty area for many phoneme transitions.
Hence, such a test of automatic segmentation performance is not an evaluation in the sense of what is correct, but rather a confrontation between the labeller's strategy and automatic segmentation criteria.

Another possibility is to check the ASS results manually, and score whether the boundary placements are acceptable or not. In this way an ASS boundary placed e.g. in the uncertainty area, can be judged as correct even if it does not coincide exactly with the manual segmentation. However, for the development and testing of various ASS algorithms this approach is usually too time-consuming.

Despite the weaknesses of the comparison-test between automatic and manual segmentation, it has been selected because it makes it easy to develop a fully automatic scoring technique.
When the reference is selected the next problem is how to represent the results of such a comparison. For **acoustic segmentation** a comparison based on dynamic programming will be used (see section 7.2.2). For **phonemic segmentation** we distinguish between **fine errors** and **gross errors**, as shown in figure 6.5 below. A *gross error* is detected when the automatically placed boundary is so displaced relative to the manual reference that its position passes beyond the region of the adjacent manually labelled segment(s). Many gross errors may reflect that the ASS algorithm is prone to *ripple errors*[2]. This is a serious miss in that one will not find such a "gross error segment" by direct access in the automatically segmented speech database. Gross errors will thus reduce performance for ASR systems trained on such annotated speech or for TTS based on an automatically generated subword library. It is thus important to keep the number of gross-errors as low as possible.
A *fine error* is detected when the automatic segment boundary is not 100% overlapping the corresponding manually placed segment boundary.
The main objective of segmenting and labelling a speech material is to establish a correspondence between a phonemic symbol and a portion of the signal. The exact placement of the segment boundaries is subject for discussion and hence the fine errors are not as critical as the gross errors.

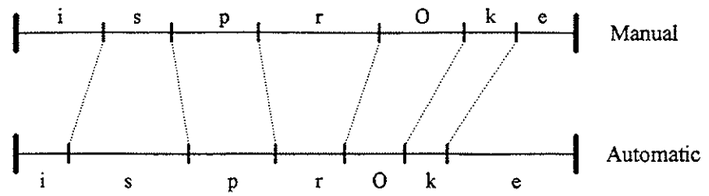


Figure 6.5 *Examples of fine errors and gross errors (see text). The dotted lines indicate the corresponding manually and automatically placed boundaries. The deviation between those corresponding boundaries are called fine errors. The automatically computed /k/-segment is a gross-error because it ends before its corresponding manual /k/-segment starts.*

When the deviation between corresponding manual and automatic segment boundaries is less than a given threshold, we say that these boundaries **coincide. Coincidence rates** are estimated as the number of boundaries that coincide, divided by the total number of segment boundaries computed. Usually the fine errors for all segment boundaries are given as **cumulative coincidence rates**. We may also depict the fine errors in a **deviation histogram** which shows how many occurrences are detected within each deviation threshold. The histogram shows whether the deviation errors are symmetrical around zero or whether there is a systematical bias in the deviation data. The histograms can be made for all the segment boundaries or one may check special phoneme transitions to find which transitions contribute most to the overall deviation results.

---

[2] That is, one large error may cause many subsequent errors.

When evaluating deviation thresholds for coincidence counts, the comparison of the two manual segmentations, such as the one in section 4.2.4, should be accounted for. For instance, at a deviation threshold of ±5ms only 63% coincidence was obtained in section 4.2.4. Also, ±5ms is within the limits of the preprocessing resolution of most ASS-systems.
The figures of coincidence between ASS and manual segmentation within ±10ms deviation is claimed to "represent the realistic portion of automatic boundaries which would not have to be readjusted in the manual correction phase of a database labelling task" [Barry'91b, p.29]. That is, ±10ms is within the limits of accuracy that can be reasonably expected of human labellers with common labelling criteria. In a comparison of different manual segmentations [Erp'88], differences were considered as irrelevant when they were less than 10ms. For research purposes as measurements of phoneme durations, 10ms was considered as the minimum degree of precision required. However, the comparison of two manual segmentations in section 4.2.4, yielded only a coincidence rate of 88% within ±10ms deviation. We should therefore advocate coincidence rates with ±20ms or ±25ms deviation thresholds also[3].

The above proposed evaluation tests in terms of coincidence rates are all **binary** in that they only check whether the automatically generated segment boundaries are within a certain threshold of deviation from the manually placed boundaries. However, from our experience with manual segmentation we know that some phoneme transitions were relatively easy to segment consistently and accurately, whereas other transitions were difficult to segment manually and thus were characterized by large uncertainty areas.
Therefore, ideally the deviation threshold should vary depending of the phoneme transition in question. If such a set of thresholds is used, one may normalise the deviation between manual and automatic boundaries to get a more realistic score function. For instance if the threshold for a given phoneme pair is x ms and the actual deviation is y ms, the normalised deviation could be measured as y/x ms.
In this way we can build in phonetic knowledge in the evaluation of ASS in order to distinguish between *serious fine errors* and *less serious fine errors*. Examples of serious errors could be a large deviation for the segment boundary between a voiced and a voiceless phoneme or labelling a nasal segment as a voiced fricative. And less serious errors would e.g. be a deviation at the vowel to lateral transition or labelling a nasal as a vowel.

In order to evaluate changes made on one ASS-algorithm and to compare different ASS-algorithms, statistical considerations of the coincidence rates are required. For a given test set, one possibility is to estimate an interval in which the true coincidence rate is included with a certain probability. Such **confidence intervals** indicate better how good the evaluation tests are and how much we can rely on the estimated coincidence rates. For instance, a 95% confidence interval for a coincidence rate includes the true coincidence rate with a probability of 0.95. With a given level of confidence the (expected) width of the confidence interval will shrink as the number of observations increases.

For the estimation of confidence intervals in this thesis we will model the automatic segmentation of a given test corpus as a set of N independent experiments. In each of these experiments the automatically placed boundary either coincides with its corresponding manual

[3] For instance, in the SAM-project the ±20ms deviation threshold was selected as basic reference threshold because this was the minimal common number for the preprocessing resolution of the methods evaluated in [Barry'91b].

boundary, or it does not coincide (within the given deviation threshold). If we then assume that each such experiment has the same probability of obtaining coincidence, the number of coincidences is **binomially distributed** [Høyland'83].
In the assessment of ASS algorithms, the number of phonemes in the test-sets is usually in order of thousands. The number of coincidences can then be approximated by a normal distribution [Høyland'83, p.165], and a confidence interval can be conveniently estimated according to eq. (17.16) and (17.18) in [Høyland'83].
In this thesis a difference in coincidence rate is regarded as **significant** if the corresponding 95% confidence intervals do not overlap.

However, the assumptions of N independent experiments required for the binomial distribution, is obviously not correct for automatic segmentation. But since the dependency is difficult to model and it probably differs among ASS algorithms, we simply used the independency assumption[4]. With the binomial distribution the coincidence rate estimator defined above is both *unbiased* and *consistent.*

If the test database is not phonemically balanced, statistics of deviation errors may give a wrong assessment of the ASS performance, e.g. if an ASS algorithm has problems with a certain phoneme pair which is over-represented in the test speech. Therefore some qualitative results should also be included in a proper evaluation of ASS.
The individual phonemes can be evaluated with respect to all its right and left contexts, as e.g. in [Schmidt'91]. Large amounts of data will be produced if all phonemes are to be thoroughly investigated.

For practical use, e.g. for shared use across laboratories, the ASS methods should also be evaluated with respect to other criteria, such as *user friendliness* (e.g. easy installation and use, documentation, and flexibility for different test sets), *hardware requirements* (e.g. computer system and memory size), and the amount of manually annotated speech material required for training the ASS software.

[4] If the dependency had been modelled, the final confidence intervals would probably be broader, i.e. the confidence will be overestimated.

6.4 SUMMARY

In this chapter several approaches for automatic segmentation of speech (ASS) have been described and a taxonomy of the ASS methods is proposed. The two main classes were the acoustic segmentation algorithms and the phonemic segmentation algorithms. Since the methods were complementary we developed a two-step algorithm which utilised the advantages of both acoustic and phonemic segmentation algorithms. Our algorithm was implemented in a modular manner so that each part of the system could be tested alone and as an integrated part of the total system.

Problems concerning assessment of ASS algorithms have been discussed. In this thesis the performance of the automatic segmentation will be measured as the coincidence with the manual annotation in terms of fine errors and gross errors.

Chapter 7

EXPERIMENTS

In this chapter several experiments with our automatic segmentation algorithm are evaluated. The acoustic segmentation part is assessed independently and with respect to its affects on the constrained HMM segmentation. For the phonemic segmentation the performances of the pure HMM segmentation and the constrained HMM segmentation will be compared for almost all experiments throughout the chapter.
Some approaches for utilising existing annotated speech corpora for enlarging the phoneme training sets are suggested and the phonemic segmentation algorithm is attempted made more independent of language and recording conditions by cepstral domain filtering.
Finally we have investigated how optimisations with respect to segmentation accuracy influenced the performance of the corresponding HMM-based phoneme recogniser.

7.1 PREPROCESSING

Speech material

The automatic segmentation experiments were performed on continuous passage recordings of English, Italian, Danish, and Norwegian, referred to as the *EUROM0 recordings* in this thesis (see section 4.1.1). In the English EUROM0 recording the mean value of the DC-component had a linear PCM value -4878. This bias was subtracted from each sample value prior to the analysis. The other recordings were used without changes. (For more details regarding recording protocols, see Appendix B for the Norwegian recording and e.g. [Grice'89] for the other EUROM0 recordings).

A shorter English *Test Passage* was provided in the SAM-project as an independent test-set. However, one of the female English EUROM0 speakers, EAE, also read the Test Passage. In order to make the speaker "unknown" for the testing on the Test Passage we had to use the phoneme HMMs trained on speaker JHE, JWE, and MBE only.

Label files

In order to examine automatic segmentation properly for all phoneme transitions and to obtain reliable estimates of the HMM parameters, the recordings for each language should have been phonemically balanced and only the phoneme inventory as defined for each language in Appendix A should have been applied in the manual annotation. However, the EUROM0 recordings were not phonemically balanced (see e.g. Appendix B), and some new symbols were introduced in the manual labelling (see Appendix C). Thus, a few editions of the label files had to be done prior to the automatic segmentation.

In the *Danish* label file 158 "impossible transitions" were marked off with asterisk, *, i.e. no boundary was set between two successive phonemes (see analysis of the Danish annotation in Appendix C). Most often only two labels were grouped into one segment, but in the 45th sentence of speaker JDD, five successive phonemes were not segmented; i.e. the words "han ligeså" ('he likewise') were segmented into three segments labelled /h*a*n*l*i/, /s/, and /Q/. In the preprocessing all such segments were divided into equal duration segments, one for each phoneme, and the asterisks were removed.
Many (not all), of the Danish plosives were segmented into two segments; a closure part, e.g. /p0/, and a burst part, e.g. /p/. These were merged in the preprocessing. The Danish rising and falling diphthongs, and the /j/ and the "stød" /?/, did not occur in the Danish EUROM0 label files.

In the experiments we did not distinguish between various extralinguistic factors which could have been marked off in the manual label files. Thus, the segment labels 'breath' and 'lipsmack' in the *English* EUROM0 label file were converted to the silence symbol '...'. If the altered label file contained consecutive segments with the silence label, these segments were merged to one segment with one silence label. The phoneme classes were used as defined in table A.3 in Appendix A, except for the *Sonorant* class which was divided into *Nasals* /n m N/, *Liquids* /l r/, and *Glides* /w j/.

For *Italian* some of the geminate consonants, /bb dd gg ff vv/, did not occur at all, whereas some, /pp, kk, SS, LL/, occurred so rarely that they were merged with their single counterparts for training phoneme HMMs. A few phonemes were grouped into one symbol, so that the label /tS/ covered both /tS/ and /ts/, /dZ/ covered /dZ/ and /dz/, and /L/ covered /L/, /LL/, and /JJ/. With this merging 37 symbols were used for Italian.
When an unvoiced plosive initialised an utterance, the preceding silence and the plosive were joined as one segment, labelled e.g. /#k/, in the manual annotation of Italian. The preprocessing divided such a segment into a silence segment and a plosive segment by placing the start boundary of the plosive 2400 samples (=150ms) before its end boundary. If the pause was shorter than 150ms, the start of the plosive was placed 10 samples after the end of the preceding phoneme segment.

For *Norwegian* some of the phonemes, /C {i A} rn rl/, occurred too rarely to provide phoneme HMM estimates. Thus, in the preprocessing /C/ was grouped with the /S/-segments, /{i/ with the /i:/ segments, /A}/ with the /A/ segments, /rn/ with the /n/ segments, and /rl/ with the /l/-segments.

Signal processing

As pointed out in section 5.1, the spectrum at the glottal source of voiced sounds decreases with increasing frequency. The high frequency formants thus have lower amplitude than the formants at lower frequencies. To model e.g. F_3 and F_4 as well as F_1 the speech signal was pre-emphasised prior to the LPC analysis. The dynamic range of the digitized speech signal was compressed by pre-emphasising with a first order high-pass filter $T(z)=1-\alpha z^{-1}$. The coefficient α is usually chosen close to 1 and was in this work selected to be *0.95*.

As argued for in chapter 5 the LPC-coefficients were estimated by the *autocorrelation method*. To reduce the edge effects due to limited analysis window, a *Hamming window w(n)* of length

N was applied. The selection of analysis window duration N was a compromise between two requirements: (i) the window should be *short enough* to ensure that the speech spectrum was relatively constant within the window, and (ii) the window should be *long enough* to ensure reliable estimates of the autocorrelation function. The Hamming window allows for using longer window duration than normally allowed by the stationarity constraint because it emphasizes the middle of the window.
Since we aimed at automatic segmentation of speech, which often implies marking boundaries at abrupt changes in the spectrum, we stressed the importance of a short window more than is usually done in e.g. ASR systems. Hence, a Hamming window of length N=240 samples corresponding to 15ms at 16kHz sampling frequency, was applied.

The Hamming window was centred around every 5ms speech frame t to be analyzed:

$$\tilde{s}_t(n) = s_t(n)\cdot w(n) \tag{7.1}$$

The maximum accuracy of the ASS system was thus constrained to the 5ms analysis window shift rate.

The autocorrelation coefficients for each frame t were estimated from the windowed data according to eq. (5.7):

$$\hat{R}_t(m) = \sum_{n=0}^{N-1-m} \tilde{s}_t(n)\cdot\tilde{s}_t(n+m) \qquad m=0,1,...,p \tag{7.2}$$

The parameter choices were based on knowledge of speech, speech modelling, proposals for ASR in the literature, and what was practical for the implementation on a PC with MS-DOS v.3.1.
In the first experiments, a *p=15* order LPC-vector was estimated from eq. (5.8) by Levinson-Durbin's recursive algorithm. The choice of LPC-analysis order was a trade-off between reliable estimate of the acf lags, spectral accuracy, and computation time and memory.

Although the LPC-cepstrum coefficients were derived directly from the LPC-coefficients, eq. (5.11), experiments with ASR have shown that employing more cepstral coefficients than the LPC analysis order improves the ASR accuracy, e.g. [Husøy'91]. Thus, *nc=18* LPC-cepstrum coefficients were computed by eq. (5.11), and the distortion between a reference frame R and a test frame T was estimated by the *Euclidean cepstral distortion measure*:

$$d_{cep} = \sum_{n=1}^{nc} [c_R(n)-c_T(n)]^2 \tag{7.3}$$

The energy contour may also convey information that can be utilized for automatic segmentation of speech. To reduce the dependency on the absolute energy level, which depends on e.g. the loudness of the different speakers and the recording conditions, the energy in frame t within an utterance was normalised with respect to the peak frame energy for that utterance:

$$E(t) = \log\left[\frac{\hat{R}_t(0)}{\hat{R}_{max}(0)}\right] \tag{7.4}$$

The first order time derivatives of the normalised log energy values, called the *delta energy*, were

approximated for a frame t as:

$$\Delta E(t) = E(t) - E(t-1) \tag{7.5}$$

The use of CDHMMs allowed for easy expansion of the frame parameter vector for phonemic segmentation. The basic context-independent phoneme models were defined with 3 states, with skip-transition between first and last, 1 mixture per state, no explicit duration modelling, and diagonal covariance matrix.

Due to practical problems only the cepstrum representation was used in the acoustic segmentation experiments.

7.2 ACOUSTIC SEGMENTATION

7.2.1 CONSTRAINTS

Some sentences (utterances) in the EUROM0 recordings had a duration of 6 to 7 seconds, or 1200 to 1400 frames. The segment distortion matrix, $\mathbf{D}_s$, thus became large and the computation could be rather extensive. To reduce the memory requirements the segment distortions were not computed and stored prior to the DP search, but computed when needed in the DP calculation. Since many frames belonged to the same acoustic segment, succeeding frames in the backpointer array have the same backpointer value. Thus, a backpointer value was only recorded when the value changed, and a maximum of 50 backpointer values were kept at each segmentation level. The modification had minimal effect on the segmentation accuracy, and gave only a slight deterioration within the small deviation thresholds.

The maximum duration of an acoustic segment was constrained to 250ms, or 50 frames. For the Norwegian EUROM0 the average phoneme duration was 96ms. With 100% acoustic *oversegmentation* (o.s.) the average duration of the acoustic segments was 50ms and with 150% o.s. the average duration of the acoustic segments was 40ms. That is, the constraint of acoustic segment duration of 250ms should not influence the segmentation performance. Actually, although constraining the segment duration is sub-optimal, the constrained version gave slightly better results at small deviations, (i.e. within ±5ms and ±10ms), from the manual segment boundaries.

7.2.2 QUANTITATIVE EVALUATION

Due to the acoustic oversegmentation we cannot compare the acoustic segment boundaries directly with the manually placed phonemic boundaries. With M phonemes in an utterance, only the M best out of the N (N = oversegmentation factor * M) acoustic boundaries should be used as basis for the performance results. However, if we simply count the M acoustic boundaries with least deviation to any phoneme boundary, an acoustic segmentation which places all boundaries at one single time instance near one phoneme boundary, would give 100% coincidence with the manual segmentation as result. Thus, the acoustic segment boundaries should be compared with the manual broad phonetic segmentation boundaries by using the *Dynamic Time Warping (DTW)* technique, where *each phonemic boundary is compared with its closest acoustic boundary only.*

Interpreting T(n) as the position of the n'th acoustic segment boundary and R(m) as the position of the m'th manually placed phonemic segment boundary, the DTW-formula in eq. (5.21) was used to compute recursively the minimum accumulated difference $D_s(n,m)$ from the start point (1,1) to a boundary pair (T(n), R(m)) as:

$$D_a(n,m) = d(T(n),R(m)) + \min_{m-1 \le k \le n-1} \left[D_a(k,m-1) \right] \quad , \text{for } n \in \{m, N-M+m\} \tag{7.6}$$

where the constraints ensured that only one acoustic boundary was "paired" with a phonemic boundary. The local distortion measure used was simply the absolute value of the difference between these two time instances, i.e.: $d(T(n),R(m)) = | T(n)-R(m) |$.

If two acoustic boundaries were at equal distance to one phonemic boundary, (i.e. they were placed symmetrically on each side of the phonemic boundary), the first appearing acoustic boundary was chosen as the "correct" one.
The best match of pairs of acoustic and phonemic segment boundaries were found by backtracking the optimal path.

This DTW-comparison was run for each deviation threshold, ±5ms, ±10ms, ±15ms, ±20ms and ±25ms. If the local distortion was bigger than the actual threshold it was penalised by adding 50ms to the distortion value in order to avoid that such points were included in the optimal path.

As an example, the DTW-coincidences for acoustic segmentation with 100% oversegmentation (o.s.) for each speaker in the Norwegian EUROM0 recordings are shown in table 7.1. These results will be further discussed in section 7.2.4.

	Deviation from manual broad phonetic segmentation				
Speakers	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	53.29	76.35	87.63	92.84	94.80
SHN	57.46	80.65	89.12	94.03	96.82
TBN	59.08	82.80	92.99	96.45	97.89
TGN	60.87	84.57	92.47	95.86	97.37
Average	57.68	81.10	90.55	94.79	96.72

Table 7.1 *DTW-coincidence rates in percent for the acoustic segmentation with 100% oversegmentation for each speaker of the Norwegian EUROM0 recordings. (Totally 4158 boundaries in the coincidence calculation, see section 7.3.1).*

7.2.3 QUALITATIVE EVALUATION

The automatic quantitative evaluation of the acoustic segmentation in terms of DTW-comparisons with manually placed phonemic boundaries may be difficult to interpret. In order to gain more insight into how the acoustic segmentation algorithm performed, *qualitative analyses* were required.
The purposes of the qualitative analyses of the acoustic segmentation were threefold. Firstly, we wanted to investigate whether the acoustic segments could be given reasonable *phonetic interpretations*; e.g. whether the plosives were segmented into a closure, burst and aspiration part. If the acoustic segmentation algorithm was able to isolate identifiable sub-phonemic segments consequently, it could be useful for e.g. speech analyses and ASR based on acoustic subwords. Secondly, referring to the discussion in chapter 4.1, it was interesting to find out whether the acoustic segmentation was accurate and consistent enough to be used as a *pre-segmenter for a standard multi-lingual manual segmentation*.
Thirdly, the qualitative analyses of the acoustic segmentation will *disclose at which instances* the acoustic segmentation boundaries were displaced compared to the manually placed boundaries.

The qualitative analyses were carried out on the Norwegian EUROM0 recording. As an example,

the qualitative analysis is first carried out on the acoustic segment boundaries computed with 100% o.s. of the **first sentence of speaker AFN**, shown in figure 7.1 below. General findings of the analyses are described at the end of this section.

The starting and ending points of the utterance were given as input to the algorithm, so the beginning of /i/ is correctly marked. Within the /i/-segment the amplitude increases evenly to a maximum and then it decreases. After the top in amplitude the high F_2 of /i/ moves downwards due to the following alveolar fricative. The algorithm detects such spectral changes and an acoustic boundary is placed where the spectral change starts, (which happens to be at the amplitude top in the waveform)[1].

The voiced to voiceless transition /i/-/s/ was shown in section 4.1.2 to be segmented differently by different human labellers, because the frication of [s] superimposed the voiced waveform. Two acoustic boundaries are placed in this transition: one similar to the Norwegian manual segmentation and the second is placed even later. At the end of /s/ also two acoustic boundaries are marked: (1) when an abrupt intensity change is seen in the spectrogram and (2) when no activity is seen in the spectrogram (the manual boundary decision is based on the waveform and is set between the acoustic boundaries).

The small, voiceless [p]-burst is not detected and the boundary is placed where the voicing begins, i.e. the /p/-/r/ boundary coincides with the manually placed boundary. When /r/ follows a plosive, a voiced segment, an epenthetic schwa sound [@], appears before the apical tap closure. This [@] is in the manual segmentation included in the /r/ segment (see chapter 4.2.2), whereas the acoustic segmentation separates these two parts of /r/.

In the manual segmentation one extra pitch period on each side of the /r/-closure was included in the /r/-segment. For the /r/-/O:/ transition here the first acoustic boundary, which marks the end of the closure phase of /r/, coincides with the manual segmentation. The following acoustic boundaries are marked off because of the huge formant-transitions (especially in F_2 and F_3) in the beginning of /O:/.
The following acoustic segment may represent the "stable" or "homogenous" portion of the vowel. The intensity of /O:/ gradually decreases towards the closure of the following /k/. In this gradual decrease one acoustic boundary is placed when the intensity reduction in the higher frequencies begins, the next is placed when almost no intensity is seen in the spectrogram and no amplitude appears in the waveform, and the last one is placed when no intensity is detected in the spectrogram.

The abrupt burst onset in /k/ is detected and in addition, the aspiration phase is assigned a separate segment. At the /k/-/e/ transition the algorithm does not notice the voicing onset (intensity seen in F_1), but places the boundary at the onset of the other formants.

The /e/-/k/ transition represents a word boundary, but otherwise the transition should be similar to the /O:/-/k/ described above. However, the intensity is here turned off more instantaneously at all frequencies so that the second acoustic boundary (of the three boundaries in the /O:/-/k/ transition) is not marked.

---

[1] In the acoustic segmentation algorithm analyzed here, only spectral information, represented by 18 cepstrum coefficients, was used. That is, no energy or voice information was utilised.

The onset of the /k/ burst is detected and the burst and aspiration phase are here grouped into one common segment.
The /k/-/A/ boundary coincides with the manual boundary. The realisation of the vowel /A/ is similar to /i/ described above and also segmented similarly. The /A/-/n/ transition is segmented correctly with one extra boundary within /n/ because the F_2 and F_3 start moving to positions for the following /i/.

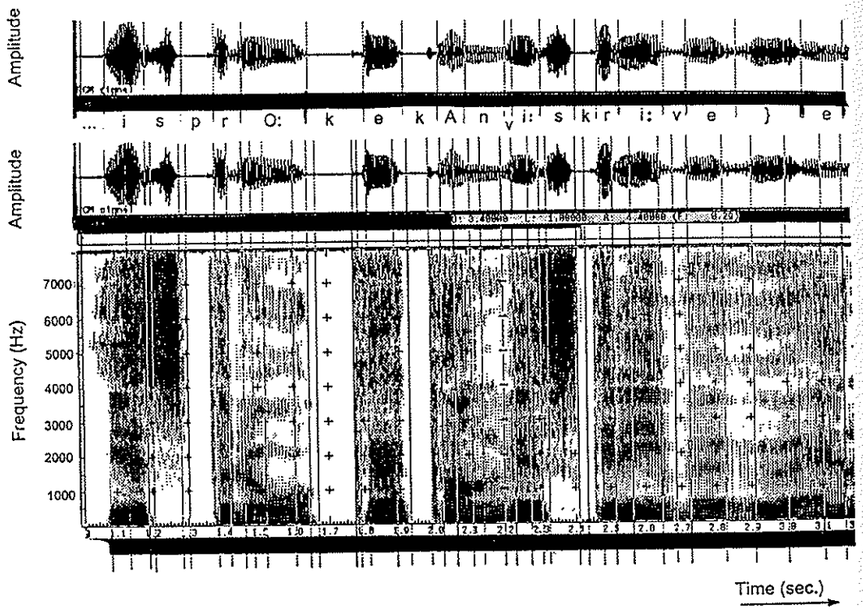


Figure 7.1 *The sentence "i språket kan vi skrive ue(ndelig)" (=in language we can write inf(initly)) is manually segmented and labelled with SAMPA symbols: /i sprO:ke kAnvi:skrive}e/ shown under the waveform in the top row of the figure. In the waveform below and in the broad band spectrogram the acoustic segmentation boundaries with 100% o.s. are shown. The + signs in the spectrogram are separated by 0.1 sec. horizontally and 1000 Hz vertically.*

The /n/-/v/-/i/ sequence is rather difficult to segment and in the manual segmentation a short /v/ segment is squeezed in between /n/ and /i/. This transition-segment is reasonably well detected acoustically because the acoustic segmentation algorithm tends to place a couple of boundaries in gradual spectral transitions.
The waveform pattern within the /i/-segment is similar to that of the other vowels and at the /i/-/s/ transition one acoustic boundary is placed on each side of the manual one. The /s/-/k/ boundary coincides with the manually placed boundary, but in the closure phase of /k/ an extra

boundary is marked.
As for /s/-/p/ above, a voiceless plosive preceded by a /s/ is unaspirated and the weak burst is not detected by the acoustic segmentation. But the /k/-/r/ boundary coincides with the manual one. Within /r/ the schwa sound is delimited as a separate segment. The /r/-/i/ is difficult for the acoustic segmentation because the r-closure is not a segment with less intensity, but only a short dip followed by an evenly increasing amplitude. Manually this boundary is thus placed rather arbitrarily. The next acoustic boundary is the one that splits the vowel in two parts (as for the other vowels described above).
For /i/-/v/ the acoustic boundary is marked at the intensity drop in the spectrogram, a little too early compared to the manually placed boundary. Then one extra acoustic boundary is placed in the middle of /v/ segment. The /v/-/e/ boundary coincides with the manual segmentation.

As discussed in chapter 4, vowel-sequences are often realised with a short period of creaky voice, especially at word and morpheme boundaries as exemplified here by the /e/-/}/ transition. Manually the boundary between the vowels is placed in the middle of this period of creaky voice. In the automatic segmentation /e/, the creaky voice period, and /}/ are segmented into separate segments. These acoustic boundaries are thus much displaced compared with the manual segmentation and will result in big deviations in the quantitative analyses.
-The fundamental question is then: *Is the acoustic segmentation wrong in this case*?

General findings:

Based on similar qualitative analyses of the acoustic segmentation of a larger speech material some general trends were found:

-*Plosives* were at least segmented into a closure part and a burst part by the acoustic segmentation algorithm. If the closure contained some voicing, this was also separated as one segment. Often some alternatives for the beginning of the closure and the end of the burst were made. If the plosive release contained both a burst and an aspiration part, these were marked off as two separate segments.

-Segments containing *extralinguistics* (e.g. creaky voice, epenthetic silence, epenthetic sound, breath and lipsmack) were marked off separately.

-*Vowels* realised with an amplitude that increased evenly to a maximum value and then decreased towards the next phoneme often contained formant-transitions which were detected by the acoustic segmentation, and an acoustic segment boundary was placed near the amplitude top. (For these examples the acoustic segmentation could be used for marking the "centre" of the phonemes as done manually for the Dutch EUROM0 recording [Erp'88]).

-In the transition from *vowel to silence* the acoustic segmentation algorithm calculated two or three boundaries. The first one was placed where the intensity reduction began in the higher frequencies, the second (optional) one was placed where almost no intensity was registered in the spectrogram, and the third one was placed where no intensity at all was detected in the spectrogram.

-When the /r/ was realised with an epenthetic schwa, as in /sprO:/ and /skriv/ in figure 7.1, the [@] and [r]-closure became separate acoustic segments.

More generally we may conclude that with 100% oversegmentation most of the acoustic segments could be given a phonetic interpretation and that the acoustic segmentation was consistent in that similar instances of the speech spectrum were segmented similarly.
The issue concerning acoustic segmentation used as a pre-segmenter for manual segmentation will be discussed after the effects of different oversegmentation factors are investigated.

7.2.4 ACOUSTIC SEGMENTATION AS A FUNCTION OF OVERSEGMENTATION

The accuracy of the acoustic segmentation in terms of DTW-coincidence with manual segmentation will increase with increasing o.s.-factor. However, too many acoustic boundaries will make it difficult for the succeeding constrained HMM phoneme segmentation to chose the correct ones. It is thus preferable to keep the o.s.-factor as low as possible while still achieving high DTW-coincidence with manual segmentation. In this section we investigate the performance of the acoustic segmentation as a function of oversegmentation.

7.2.4.1 Quantitative analyses

In table 7.1 the DTW-coincidence for acoustic segmentation with 100% oversegmentation (o.s.) was shown for Norwegian. Here, these results will be compared DTW-coincidences obtained with other o.s. factors. For instance with 150% o.s. the acoustic segmentation accuracy for the four Norwegian EUROM0 speakers were:

	Deviation from manual broad phonetic segmentation				
Speakers	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	58.59	81.35	91.17	95.39	97.25
SHN	64.68	88.64	95.38	97.98	99.13
TBN	64.46	88.47	96.06	98.46	99.14
TGN	67.07	89.93	96.05	97.65	99.06
Average	63.70	87.10	94.67	97.37	98.65

Table 7.2 *DTW-coincidence rates in percent for the acoustic segment boundaries within different deviation thresholds from the manually placed phoneme boundaries, using 150% o.s. on Norwegian EUROM0.*

For acoustic segmentation with 150% o.s. of speaker AFN the DTW-coincidence is similar to the coincidence between two manual segmentations of AFN for deviations bigger than ±15ms, but significantly worse for small deviations (cf. figure 4.11). On average for the four speakers the DTW-coincidence for the acoustic segmentation is approximately equal to that of the two manual segmentations.

The acoustic segmentation with 150% o.s. obtained highest DTW-coincidences for speaker TGN who had the fastest speaking rate and thereby the shortest average phoneme duration (see table B1 and B2 in Appendix B). Since speaker AFN was the easiest and TGN was the most difficult

to segment manually, the reason for this high accuracy of the acoustic segmentation of TGN is probably the shorter phoneme segments.

The performance of the acoustic segmentation of the female speakers, SHN and TBN, was between that of AFN and TGN for DTW-coincidence rates within the ±5ms deviation threshold for all oversegmentation factors.

Decreasing the o.s.-factor from 150% to 100% the acoustic segmentation performance was reduced most at the ±5ms and ±10ms deviation thresholds (compare table 7.1 and table 7.2). This indicates that some of the most "correctly" placed boundaries obtained with 150% o.s. were not calculated with 100% o.s, but because the average acoustic segment duration with 100% o.s. is still relatively short compared to the average phoneme duration, most of the acoustic boundaries were still within the ±25ms deviation threshold.

Detailed results for the acoustic segmentation of Norwegian EUROM0 using other oversegmentation factors are provided in Appendix D, table D.1-D.7. Figure 7.2 shows the acoustic segmentation DTW-coincidence rates within ±20ms deviation as a function of the o.s.-factor. For high o.s.-factors, acoustic segmentation of speaker AFN obtained least accuracy. However, the acoustic segmentation accuracy of speaker AFN dropped less than for the other speakers, and with no oversegmentation at all (i.e. the number of acoustic segments were equal to the number of phonemes), best accuracy was obtained for speaker AFN. These results confirm the impression that speaker AFN had the most distinct articulation.

The results in table 7.1, table 7.2, tables D.1-D.7, and in figure 7.2, suggest that an oversegmentation of at least 75% is needed to obtain adequate DTW-coincidences with the manual segmentation.

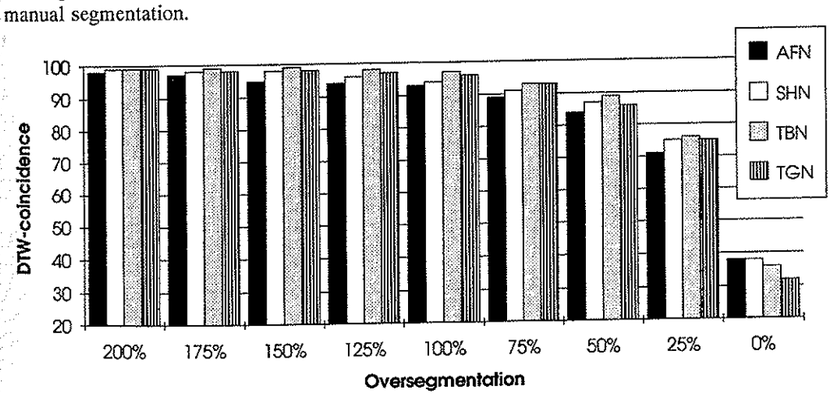


Figure 7.2 *DTW-coincidence within ±20ms deviation as a function of oversegmentation for the four Norwegian EUROM0 speakers.*

A minimum requirement for the acoustic segmentation algorithm is that it must perform better than *uniform segmentation*. By uniform segmentation a sentence is divided into equal length segments. The DTW-coincidences of the uniform segmentation actually indicates the difference

in articulation rates. For instance, since highest DTW-coincidence with uniform segmentation was obtained for speaker TGN, (table D.8 and D.9), speaker TGN had the highest articulation rate (cf. Appendix B).

Comparing the acoustic and uniform segmentation at 100% o.s. with respect to average DTW-coincidences for the four speakers, figure 7.3, the acoustic segmentation performed significantly better than the uniform segmentation for deviations less than ±20ms, showing that the acoustic segments actually detects stable parts of the signal. Within the ±25ms deviation threshold the DTW-coincidence rates were more similar. This is probably due to the short average acoustic segment duration, which is 40 ms with 150% o.s.

With no oversegmentation (0% o.s.) the acoustic segmentation obtained significantly higher coincidence rates than the uniform segmentation at all deviation thresholds.

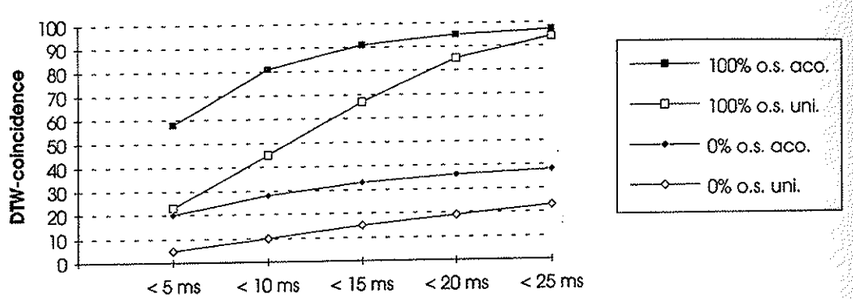


Figure 7.3 *Average DTW-coincidence rates for acoustic (aco.) and uniform (uni.) segmentation with 100% and with no oversegmentation for Norwegian EUROM0.*

7.2.4.2 Qualitative analyses

The quantitative analyses showed that the DTW-coincidence between acoustic boundaries and manually placed boundaries increased with increasing o.s.-factor. For o.s.-factors higher than 150% the improvement in performance was minimal. However, the quantitative analyses did not reveal which boundaries were changed or at which instances new boundaries were placed when gradually increasing the o.s.-factor. To pursue these issues further and to supplement the discussion in section 7.2.3, a qualitative analysis was needed.

As in section 7.2.3, the analysis is first exemplified by the first sentence of speaker AFN. Figure 7.4 shows the (first part of the) first sentence by speaker AFN segmented by different o.s.-factors in the acoustic segmentation. With **no oversegmentation** (21 boundaries) the acoustic segment boundaries obtained good coincidence with the manual segmentation, expect for: The first /i/ is divided into two parts, the first /r/ is divided into two parts: [@] and r-closure, the start of the first /v/-segment is left out, there are two boundaries in the second /i:/-/s/ transition, the end boundary of the second /r/ and the end boundary of the second /v/ are missing, and the period of creaky voice is marked as one segment.

With **25% o.s.** (26 boundaries) there are unnecessary new boundaries at /s/-/p/, /O:/-/k/, and /s/-/k/. The second /r/ is divided in two parts: a [@] and the r-closure.

At **50% o.s.** (31 boundaries), a correct end-boundary at the last /k/ is marked, but there are unnecessary new boundaries at /e/-/k/, /k/-/A/, /i:/-/s/, and /s/-/k/.

At **75% o.s.** (35 boundaries) new boundaries are placed at /O:/-/k/, /k/-/e/ (the aspiration for /k/ was isolated), /A/-/n/, and /i/-/v/ (more correct /v/-boundary).

The **100% o.s.** (41 boundaries) is analyzed more thoroughly in section 7.2.3, but we notice the new boundaries at /i/-/s/, /r/-/O:/, /k/-/e/ (the /k/-burst became isolated), /A/-/n/, /n/-/v/, /v/-/i/ (more correct /v/-boundary), and in the middle of /ive/.

At **125% o.s.** (47 boundaries) the new boundaries are placed at /k/-/e/ and /k/-/n/ (due to the steep transitions), at /i/-/s/, /k/-/r/, /e/-/}/, and at /}/-/e/.

At **150% o.s.** (50 boundaries) the vowels which have a big amplitude in the middle are divided, i.e. new boundaries within /O:/, /A/, and /i/.

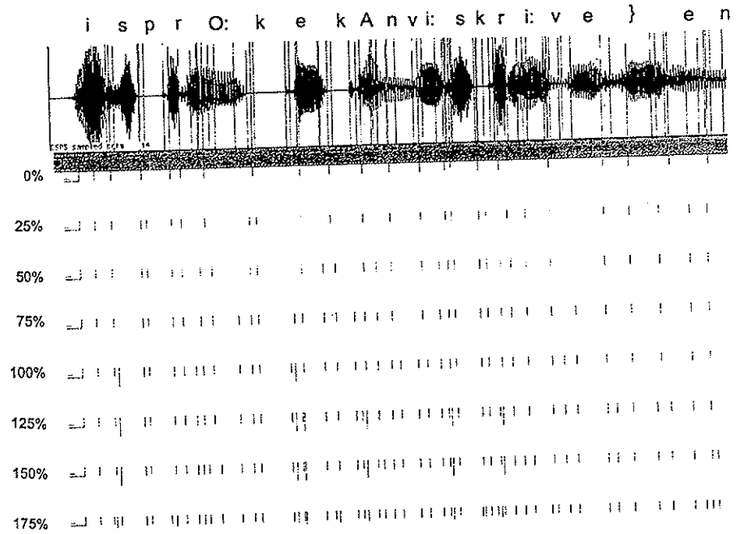


Figure 7.4 *The first sentence by speaker AFN segmented by different o.s.-factors in the acoustic segmentation.*

The qualitative analyses of the acoustic segmentation for a larger speech material showed some **general trends**:

1. Boundaries computed with a lower o.s.-factor remained fixed when increasing the o.s.-factor. That is, the effect of increasing the o.s.-factor was to split the segment(s) with highest intra-segmental distortion (see also figure 6.2).
If a boundary computed with a lower o.s.-factor is altered in a computation with a higher o.s.-factor, it is almost always changed into two boundaries symmetrically placed around the previous one, see e.g. the /O:/-/k/ transition with 25% o.s. in figure 7.4.

2. The acoustic segmentation algorithm searched for stable segments, and the boundaries were placed in transient areas because vectors from these areas increase the intra-segmental distortion. As the o.s.-factor increased, transition areas could be segmented into several short acoustic segments, providing several alternative boundaries for the succeeding constrained HMM phoneme segmentation.
(This reflects the ever returning, fundamental segmentation problem of placing a boundary between two sounds at one, single, "correct" time instant).

3. As a consequence of point 1 and 2, relatively stable portions of the speech signal were isolated and were not divided, even with a high degree of oversegmentation.

4. With more than 75% o.s., the acoustic segmentation obtained high coincidence in the DTW comparison with the manually placed boundaries (section 7.2.4.1). The few deviations from manual segmentation were mainly due to:

a) Some *half-way*, *mid-point*, or *symmetric* conventions used in the manual segmentation, e.g. our convention of placing the boundary in the middle of the creaky voice area between two vowels (instead of at the end of the segment where abrupt changes often occur). If this area was spectrally stable, the acoustic segmentation assigned boundaries at the ends of it.
Thus, when the following phonemic segmentation step is constrained by the acoustic segment boundaries it is forced to include such segments in one of the neighbouring phonemes.

b) "*Impossible cases*", where no boundary cue was seen in the waveform or spectrogram, and the human labeller has placed the boundary rather arbitrarily or based the decision on listening only.

c) "*Squeezed in segments*", i.e. phoneme which is perceived when listening to it in context but which is without any corresponding visible acoustic cues in the waveform or spectrogram, was often squeezed in as a very short segment between the phonemes with clear acoustic cues by e.g. the human labeller of Norwegian (see chapter 4).
(However, if the squeezed-in segment contained big formant transitions it would through the acoustic segmentation be divided into many short segments and some of these boundaries would therefore coincide with the manual boundary alternative).

Based on the qualitative and quantitative assessments of the acoustic segmentation we recommend the acoustic segmentation used as a pre-segmenter tool for manual segmentation. When this tool is accompanied with conventions for which boundaries to select for the various phoneme transitions, it will reduce the randomness in manual segmentation.
The acoustic segmentation may also be employed for more detailed, sub-phonemic segmentation.

7.2.5 COMPARING LANGUAGES

The acoustic segmentation algorithm is language independent, and since no training is needed it is also independent of recording conditions. Thus, very similar acoustic segmentation performances were obtained for the four languages. As an example the acoustic segmentation with 100% o.s. obtained the following DTW-coincidences within ±20ms deviation on average for the four EUROM0 speakers of each language:

Norwegian: 94.79%, **English:** 94.24%, **Italian:** 94.49%, and **Danish:** 94.33%.

Actually, within the ±15ms, ±20ms, and ±25ms deviation thresholds the average DTW-coincidence rates were very similar for the four languages, as shown in figure 7.5 (details are provided in table 7.1, D.10, D.11, and D.12).
Within smaller deviation thresholds highest DTW-coincidences were obtained for English, and lowest DTW-coincidences for Italian.

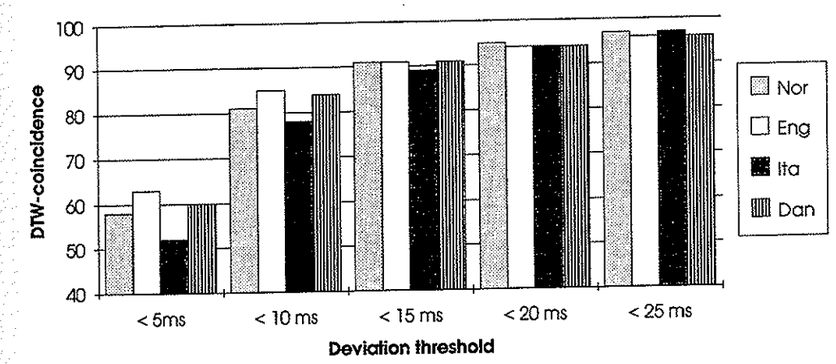


Figure 7.5 *Average DTW-coincidence in percent for acoustic segmentation with 100% o.s. of the Norwegian, English, Italian, and Danish EUROM0 recordings.*

To check whether the acoustic segmentation performed especially bad or good for some particular speakers, we compared the acoustic segmentation accuracy with 150% o.s. for each speaker, shown in table 7.2. for Norwegian, table 7.3 for English, table 7.4 for Italian, and table 7.5 for Danish.

The variation in acoustic segmentation accuracy for the different speakers was smaller for the English speakers than for the Norwegian speakers. One reason for this may be that the English speakers had approximately the same articulation rate (see Appendix C), whereas the articulation rates for the Norwegian speakers varied (see Appendix B). However, there was no one-to-one correspondence between high articulation rates and good acoustic segmentation performances.

English Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
EAE	69.59	89.93	95.52	97.85	98.69
JHE	69.06	89.21	94.70	96.78	98.20
JWE	68.30	89.19	95.66	97.60	98.43
MBE	68.77	89.15	95.28	97.36	97.92
Average	68.93	89.37	95.29	97.40	98.31

Table 7.3 *DTW-coincidence in percent for the acoustic segment boundaries with 150% o.s. on the English EUROM0 recordings.*

The acoustic segmentation performance for Italian was significantly poorer than for Norwegian, English, and Danish at ±5ms and ±10ms deviation thresholds. Particulary bad results were reached for speaker LCI and GCI. When higher deviation is allowed for, the acoustic segmentation accuracy for the four languages was similar.

Italian Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
GCI	55.70	81.08	93.29	96.93	98.46
LCI	53.69	81.10	93.03	96.27	97.97
LVI	63.86	88.04	94.99	98.06	99.19
PUI	61.63	87.35	95.92	98.53	99.18
Average	58.72	84.39	94.31	97.45	98.70

Table 7.4 *DTW-coincidence in percent for the acoustic segmentation with 150% o.s. for each speaker in the Italian EUROM0.*

Danish Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
BLD	65.26	88.55	94.59	96.98	98.09
CHD	68.16	90.19	95.68	98.00	98.67
JDD	69.07	89.08	93.28	95.73	97.07
LFD	63.65	87.20	93.81	96.58	98.45
Average	64.54	88.76	94.34	96.82	98.07

Table 7.5 *DTW-coincidence in percent for the acoustic segmentation with 150% o.s. for each speaker in the Danish EUROM0.*

7.3 PHONEMIC SEGMENTATION

7.3.1 TEST AND EVALUATION CONDITIONS

In contrast to the acoustic segmentation, the phonemic segmentation algorithm needs to be trained on a language specific speech corpus. With the manually annotated EUROM0 recordings available, two independent tests can be performed for each language:

A. 4 speaker independent (SI) sub-tests, where the ASS algorithm is trained on the whole passage for three speakers and tested on the whole text recorded by the remaining speaker; a so-called jack-knife experimental design. These tests show the effect of including different speakers in the training set, i.e. they show the ASS algorithm's speaker dependency.
B. 2 text-independent (TI) sub-tests, where the ASS algorithm is trained on all four speakers using half the text, and tested on the four speakers' recording of the other half of the text. These tests show the sensitivity of the system to different text materials.

Thus, for the SI-tests 75% of the available speech corpus is used as training set, whereas for the TI-tests only 50% of the available speech corpus can be used for training. However, the SI-tests may still be more difficult because for the SI-tests the phoneme HMMs are trained on only one speaker with the same sex and on two speakers with opposite sex of the test speaker.

With the **Test Passage** the phoneme HMMs are trained for the speaker independent sub-test for the English EUROM0 speaker EAE.

In order to achieve the highest possible phonemic segmentation performance for a particular speech corpus and parameter setting, the ASS algorithm can be trained and tested on the whole available speech corpus. Such a testing on the training set will be referred to as **Full HMM** tests.

For evaluation, corresponding phoneme boundaries provided by automatic and manual segmentation were compared by the *ELSA (ESPRIT Labelling System Assessment) v2.3* software[2] [Bourjot'91]. ELSA provided both *gross errors* and *fine error deviation histograms*.
However, the ELSA program produced too good coincidence rates in that the given endpoints for each sentence were included in the counts of coincidence between manual and automatic segment boundaries. For instance, if a label file consists of N segments including P pauses there are N-1 segment boundaries, but only N-1-(2*P) *unknown* phoneme-to-phoneme transitions, or boundaries, to be found by the automatic segmentation algorithm. The ELSA program counted in this case N-1 coincidences.
As a concrete example, the English EUROM0 recordings contained N=4655 phonemes including P=190 pauses (cf. table C.4). The total number of transitions for the 4 speakers was then 4651, but since the endpoints of each sentence were given together with the label string to the ASS algorithm, only 4271 boundaries were unknown to the ASS algorithm.
Of the boundaries calculated in the English SI-test (see section 7.3.2.2), ELSA counted 3812

[2] ELSA (ESPRIT Labelling System Assessment software) was designed according to our discussion at the ETR meeting in Toulouse, 18-20 Jan. 1990, to provide a common tool to assess automatic segmentation methods.

coincidences within ±20ms deviation from the manual segmentation, and the coincidence rate was estimated as $^{3812}/_{4651}$=81.96%. A more correct figure of the performance is obtained by subtracting the given pause-boundaries from both the automatic and manual segmentation, giving, in this case: $^{(3812-380)}/_{(4651-380)} = {}^{3432}/_{4271}$ = 80.36% coincidence within ±20ms deviation from the manual segmentation.

In all the assessments of automatic segmentation performances presented in this thesis the transitions to and from pauses are excluded, providing coincidence rates for the unknown boundaries only.

In table 7.6 below the numbers of transitions to and from silence for the English EUROM0 recordings are presented with the percentage of the total number of occurrences in parentheses:

	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids
Silence/Class	51 (6.5%)	12 (24.5%)	32 (3.6%)	57 (3.4%)	12 (1.7%)	26 (19.5%)	0
Class/Silence	24 (3.1%)	3 (6.1%)	53 (6%)	46 (2.8%)	59 (8.4%)	0	5 (1.9%)

Table 7.6 *The numbers of transitions for the different phoneme classes to and from silence. The percentage of the total number of the given phoneme class transitions is placed in parentheses.*

For the speech corpora and label files used in this thesis, (cf. table B.1, C.1, C.4, and C.7), the Italian recordings contained 4973 boundaries which were unknown to the ASS algorithm, the Danish recordings contained 4400, the Norwegian recordings had 4158, and English had 4271 unknown boundaries.
For the gross-errors there were correspondingly 5125 unknown segments for Italian, 4582 unknown segments for Danish, 4500 unknown segments for Norwegian, and 4465 unknown segments for English.

For a first assessment of the phonemic segmentation algorithm performances, we employed the frame parameter vectors estimated in the acoustic segmentation stage, i.e. the 18 LPC-cepstrum coefficients for each 5ms frame, augmented by normalised energy and delta-energy parameters. The phoneme CDHMM defined in section 7.1 was estimated as described in chapter 5.
The acoustic segment boundaries used to constrain the Viterbi recursion were calculated with 150% oversegmentation.

In order to compare the performance of the phonemic segmentation algorithm with other automatic segmentation algorithms, we first used the English EUROM0 recordings because this has been used as reference material for several evaluation tests of both manual and automatic segmentation accuracy, e.g. [Erp'89a], [Erp'89b], [Dalsgaard'90], and [Barry'91b]. With the HMMs trained for the English SI-test available it was simple to test the algorithm on the Test Passage as well.

7.3.2 ASSESSMENT ON ENGLISH EUROM0 AND THE TEST PASSAGE

In this section the performance of the constrained HMM segmentation is first compared with the pure HMM segmentation performance for the English SI-test. Then the constrained phonemic segmentation algorithm is assessed in more detail for different tests on the English EUROM0 recordings and the Test Passage. Finally these performances are compared with those of other algorithms.

7.3.2.1 Pure versus constrained HMM segmentation

As mentioned in section 6.2.2, the phonemic segmentation could be performed either by *pure HMM segmentation* (V) or by *constrained HMM segmentation* (VC). By pure HMM segmentation the optimal phoneme segment boundaries were found as the optimal path computed by the Viterbi algorithm. In constrained HMM segmentation the transitions between phoneme models in the forward recursion of the Viterbi algorithm were only allowed for at time instances specified by the acoustic segment boundaries (see figure 6.4).

The investigations in section 7.2 showed that the acoustic segmentation coincided (in DTW sense) often with manual segmentation and that the acoustic segment boundaries could be given reasonable phonetic interpretations. *Incorporating this segmental information in the phonemic segmentation stage as constraints in the Viterbi forward recursion was thus believed to provide anchor points for the phonemic segmentation and thereby increase the phonemic segmentation accuracy over that of pure HMM segmentation.*

Pure and constrained HMM segmentation will be compared in several experiments throughout this chapter. As a first example, the performances of the two phonemic segmentation algorithms were compared on the SI-test on the English EUROM0 recordings. The acoustic segment boundaries used to constrain the Viterbi recursion were computed with 150% oversegmentation.

The segmentation results shown in table 7.7 are given as an average of the four speakers.

	Gross errors	**Deviation from manual broad phonetic segmentation**			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
English SI-test, V	3.0 - 4.1%	39.7 - 42.6%	59.3 - 62.2%	77.4 - 79.8%	82.0 - 84.3%
English SI-test, VC	1.8 - 2.7%	48.0 - 51.0%	63.4 - 66.3%	79.1 - 81.5%	83.3 - 85.5%

Table 7.7 *Gross errors and fine errors for the SI-test on the English EUROM0 for the pure HMM segmentation (V) and constrained HMM segmentation (VC).*
The gross error results are given as 95% confidence intervals for the number of gross errors, whereas the fine errors are given as 95% confidence intervals for the coincidence rates within each deviation threshold.

Table 7.7 shows that constraining the Viterbi recursion (VC) with acoustic segment boundaries implied *significant decrease in the number of gross errors* and a *significant increase in the coincidence rates* at ±5ms and ±10ms deviation thresholds. The coincidence rates at higher

deviation thresholds increased also, but these improvements were not significant (with the given level of confidence).

When the pure HMM segmentation calculated a gross error segment, some of the directly following segments may also be largely displaced through a ripple effect. Constraining the Viterbi recursion with acoustic segment boundaries reduced the number of possible instances for phoneme transitions. This reduced the ripple effects and thereby reduced the number of gross errors.

The constrained Viterbi algorithm selected the acoustic segment boundaries closest to the ones that would have been computed by the pure Viterbi algorithm. Thus, at phoneme transitions where the pure HMM segmentation obtained a reasonable coincidence rate with manual segmentation (e.g. within ±25ms deviation), the constraints with an accurate acoustic segmentation will probably force some of the phoneme boundaries to a more "correct" placement (e.g. within ±5ms or ±10ms deviation from the manual segmentation). On the other hand, if the acoustic segment boundaries did not coincide well with the manual segmentation, the constrained Viterbi algorithm is forced to select boundaries that are more displaced than that calculated by the pure Viterbi algorithm (cf. qualitative analyses of the acoustic segmentation in section 7.2.3 and 7.2.4.2).

Since the constrained HMM segmentation was most accurate for the English SI-test, several tests of the phonemic segmentation constrained by the acoustic segmentation at 150% o.s. were analyzed in more detail in the following subsections.

7.3.2.2 The English SI-test

For the constrained HMM segmentation (constrained by acoustic segment boundaries calculated with 150% o.s.) on the English SI-test, the **fine errors** for the different deviation thresholds were (given as average coincidence of the four speakers):

Absolute deviation from manual segmentation	Number of coincidences within the given threshold	Coincidence in percent	95% confidence interval for the coincidence rates
< 5 ms	2116	49.54%	48.0% - 51.0%
< 10 ms	2770	64.86%	63.4% - 66.3%
< 15 ms	3176	74.36%	73.0% - 75.7%
< 20 ms	3432	80.35%	79.1% - 81.5%
< 25 ms	3607	84.45%	83.3% - 85.5%
< 40 ms	3877	90.77%	89.9% - 91.6%

Table 7.8 *Cumulative coincidence rates for the English SI-test of the constrained HMM segmentation with acoustic segments calculated with 150% o.s.. The coincidences are given as absolute numbers, percentages, and 95% confidence intervals within the given deviation thresholds.*

For a more detailed assessment of the constrained HMM segmentation table 7.9 shows how many boundaries that were placed within ±20ms deviation from the manually segmented boundaries for each phoneme class transition. E.g. for the transitions from plosives to fricatives, 65 out of 86 possible boundaries were within the ±20ms deviation threshold.

	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids	All
Plos	24/33	0/0	65/86	471/525	13/24	14/28	36/61	623/757
Affr	1/2	0/0	2/3	39/41	0/0	0/0	0/0	42/46
Fric	137/165	1/1	33/51	428/465	69/81	19/30	25/38	712/831
Vow	271/313	17/20	496/534	14/30	424/566	10/17	48/133	1280/1613
Nas	171/201	11/12	144/149	173/221	1/1	25/28	17/25	542/637
Glid	0/0	0/0	1/1	77/131	1/1	0/0	0/0	79/133
Liq	12/16	3/4	27/28	99/185	9/15	3/4	1/2	154/254
All	616/730	32/37	768/852	1301/1598	517/688	71/107	127/259	3432/4271

Table 7.9 *Number of automatic phoneme boundary placements within ± 20 ms deviation from manual segmentation in relation to number of occurrences in the material for each phoneme class transition for the English SI-test.*

The corresponding 95% confidence intervals for the coincidence rates within ±20ms deviation for the individual phoneme class transitions are shown in table 7.10:

	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids
All/Class	82-87%	72-94%	88-92%	79-83%	72-78%	57-74%	43-55%
Class/All	79-85%	80-96%	83-88%	77-81%	81-88%	51-67%	55-66%

Table 7.10 *Coincidence within ±20ms deviation given as 95% confidence intervals for phoneme class transitions for the English SI-test.*

The best segmentation accuracy was obtained for *fricatives*, *plosives*, and *affricates*. In the transition into or out of these phonemes sharp acoustic changes were observed in the waveform and spectrogram. Therefore, these were the most easy ones to segment consistently manually and were obviously easy to locate for the automatic segmentation procedure as well. For these phoneme classes the most difficult transitions were the transitions to phonemes within the same class.

Also, transitions from nasals (*nasals/all*) were accurately segmented. Actually, the transitions *from* nasals were significantly better segmented than the transitions *into* nasals (all/nasal). The main reason for this is probably that the speech material was not phonemically balanced, so that there were more transitions that were easy to locate, as nasal/plosive and nasal/fricative, than visa versa. For transitions to nasals the vowels dominated, and a vowel succeeded by a nasal may be nasalised and thus difficult to segment accurately.

Segmentation of glides and liquids achieved the worst results. Glides are by definition transitional in their structure and in this speech material almost all glide transitions were to or from vowels which made it even more difficult both for manual and automatic segmentation. In addition, the half-way point definition for segmenting /w/, /j/ and some realisations of /r/ [Barry'90b] may be difficult to follow consistently in the manual segmentation. Such mid-point segmentations were also difficult to locate by means of the automatic segmentation (cf. section 7.2.3).

7.3.2.3 The English TI-test

The division of the speech corpus for the English TI-test is listed in table C.6. The results will be presented as an average of these two parts since the results were similar for each of them. For the constrained HMM segmentation (constrained by acoustic segment boundaries calculated with 150% o.s.) on the English TI-test, the **fine errors** for the different deviation thresholds were:

Absolute deviation from manual segmentation	Number of coincidences within the given threshold	Coincidence in percent	95% confidence interval for the coincidence rates
< 5 ms	2165	50.69%	49.2% - 52.2%
< 10 ms	2844	66.59%	65.2% - 68.0%
< 15 ms	3275	76.68%	75.4% - 77.9%
< 20 ms	3538	82.84%	81.7% - 83.9%
< 25 ms	3711	86.89%	85.8% - 87.9%
< 40 ms	3980	93.19%	92.4% - 93.9%

Table 7.11 *Cumulative coincidence rates for the English TI-test of the constrained HMM segmentation with acoustic segments calculated with 150% o.s..*
The coincidences are given as absolute numbers, percentages, and 95% confidence intervals within the given deviation thresholds.

For more detailed assessment table 7.12 shows how many boundaries that were placed within ±20ms deviation from the manual segment boundaries for each phoneme class transition.

	Plosives	Affri.	Fricatives	Vowels	Nasals	Glides	Liquids	All
Plos	23/33	0/0	69/86	474/525	15/24	10/28	40/61	631/757
Affr	1/2	0/0	2/3	39/41	0/0	0/0	0/0	42/46
Fric	133/165	1/1	32/51	426/465	70/81	18/30	23/38	703/831
Vow	260/313	15/20	511/534	15/30	499/566	9/17	50/133	1359/1613
Nas	174/201	11/12	145/149	187/221	1/1	26/28	19/25	563/637
Glid	0/0	0/0	1/1	79/131	1/1	0/0	0/0	81/133
Liq	14/16	4/4	27/28	98/185	12/15	4/4	0/2	159/254
All	605/730	31/37	787/852	1318/1598	598/688	67/107	132/259	3538/4271

Table 7.12 *Number of automatic phoneme boundary placements within ±20 ms deviation from the manually placed boundaries in relation to the number of occurrences in the material for the English TI-test.*

The corresponding 95% confidence intervals for the coincidence rates within ±20ms deviation for the individual phoneme class transitions are shown in table 7.13:

	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids
All/Class	80-85%	69-92%	93-96%	80-84%	84-89%	42-59%	46-58%
Class/All	81-86%	80-96%	82-87%	82-86%	86-91%	52-69%	57-68%

Table 7.13 *Coincidence rates within ±20ms deviation given as 95% confidence interval for phoneme class transitions for the English TI-test.*

Although a smaller training set was used for the TI-test than for the SI-test, the TI-test obtained higher coincidence rates than the SI-test within all deviation thresholds. Actually, when deviation thresholds bigger than ±15ms were allowed for, the TI-test performed *significantly* better.

As for the SI-test, the best results were obtained for transitions to and from fricatives, affricates, and plosives. Due to the limited speech material there were few occurrences for some phoneme-classes, implying rather broad confidence intervals. It may thus be difficult to compare the performances in the TI and SI tests for each phoneme-class. However, for transitions to fricatives and nasals and for transitions from vowels, the TI-test performed significantly better than the SI-test.

7.3.2.4 The Test Passage

The Test Passage was a short continuous passage, with a speaking time of 19.45 sec. distributed over 9 sentences. The 237 phonemes had an average duration per phoneme as in English EUROM0 recordings, i.e. about 82 ms. Since the HMMs were trained on the three EUROM0 speakers, JHE, JWE, and MBE, and the recording environments differed for the Test Passage and

the EUROM0, the phonemic segmentation performance obtained in the test on the Test Passage could not be better than for speaker EAE in the English SI-test.
For the constrained HMM segmentation (constrained by acoustic segment boundaries calculated with 150% o.s.) on the Test Passage the **fine errors** for the different deviation thresholds were:

Absolute deviation from manual segmentation	Number of coincidences within the given threshold	Coincidence in percent	95% confidence interval for the coincidence rates
< 5 ms	90	39.5%	33.4% - 45.9%
< 10 ms	117	51.3%	44.9% - 57.7%
< 15 ms	141	61.8%	55.4% - 67.9%
< 20 ms	156	68.4%	62.1% - 74.1%
< 25 ms	165	72.4%	66.3% - 77.7%
< 40 ms	184	80.7%	75.1% - 85.3%

Table 7.14 *Cumulative coincidences for the Test Passage test of the constrained HMM segmentation with acoustic segments calculated with 150% o.s..*
The coincidences are given as absolute numbers, percentages, and 95% confidence intervals within the given deviation thresholds.

These results are significantly worse than the average results for the SI-test (table 7.8) and also for speaker EAE in the SI-test.

For more detailed assessment table 7.15 shows how many boundaries that were placed within ±20ms deviation from the manually segmented boundaries.

	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids	Sum
Plosives	2/5	0/1	6/7	23/26	0/0	0/0	4/7	35/46
Affr	0/0	0/0	0/0	2/2	1/1	0/0	0/0	3/3
Fricatives	9/12	0/0	4/7	22/26	3/3	1/1	0/1	39/50
Vowels	14/22	1/1	19/24	2/5	16/25	0/0	1/5	53/82
Nasals	4/5	1/1	8/10	5/12	0/2	0/0	1/1	19/31
Glides	0/0	0/0	0/0	0/3	0/0	0/0	0/0	0/3
Liquids	1/1	0/0	0/0	5/10	1/2	0/0	0/0	7/13
Sum	30/45	2/3	37/48	59/84	21/33	1/1	6/14	156/228

Table 7.15 *Number of automatic phoneme boundary placements within ±20 ms deviation from the manually placed boundaries in relation to the number of occurrences in the material for the Test Passage using 150% o.s. in the acoustic segmentation step.*

7.3.2.5 Comparison of tests

Table 7.16 summarizes the some most important results obtained by the constrained phonemic segmentation for the tests of English:

	EUROM0, Full HMM	EUROM0 - SI	EUROM0 - TI	Test-passage
Gross errors	50/4465 = 1.1%	99/4465 = 2.2%	71/4465 = 1.6%	14/237 = 5.9%
Fine errors (±20ms)	3706/4271 = 86.7%	3432/4271 = 80.4%	3530/4271 = 82.7%	156/228 = 68.4%

Table 7.16 *Fine errors (provided as coincidence rates within ±20ms deviation thresholds) and gross errors for the English EUROM0 recordings and the Test Passage.*

Table 7.17 provides the corresponding 95% confidence intervals:

	EUROM0, Full HMM	EUROM0 - SI	EUROM0 - TI	Test-passage
Gross errors	0.9% - 1.5%	1.8% - 2.7%	1.3% - 2.0%	3.57% - 9.65%
Fine errors (±20ms)	85.7% - 87.8%	79.1% - 81.5%	81.5% - 83.8%	62.14% - 74.09%

Table 7.17 *95% confidence intervals for the fine errors (given as coincidence rates within ±20ms deviation thresholds) and gross errors for tests on the English EUROM0 recordings and the Test Passage.*

The best results were obtained when testing on the training set, i.e. the full HMM test. The fine error performance was significantly best for the Full HMM test, whereas the 95% confidence intervals for the number of gross errors overlapped for the SI, TI, and Full HMM tests on English EUROM0.
Although the confidence intervals for the fine errors for the TI and SI-test overlapped somewhat, the phonemic segmentation algorithm seemed less text dependent than speaker dependent.
For instance, within ±20ms deviation there were detected 106 more boundaries in the TI-test than in the SI-test. This difference was mainly due to more accurate segmentation of the transitions between the phoneme classes: vowel/nasal (+75), nasal/vowel (+14) and vowel/fricative (+15). For the other transitions the TI and SI test performed very similarly.
Also, fewer gross-errors were detected in the TI-test (71) than in the SI-test (99), but this difference was not significant. Especially, more liquids and nasals were displaced as gross errors in the SI-test.

The tests on the Test Passage obtained significantly worse coincidence rates than on the SI and TI tests on the English EUROM0 recordings. In addition there were detected significantly more gross errors for the Test Passage than for any other test. (However, half the gross errors in the Test Passage were very small, i.e. 5 gross errors were only 1ms displaced and 2 gross errors were 2ms displaced).

In the ESPRIT-SAM project three ASS-algorithms were tested and compared on the same speech

material by several laboratories[3] independently [Barry'91b]. The three methods were the Danish DK_SALA [Dalsgaard'90] and the French FR_SALA [Dours'89] described in section 6.1.1.3, and our algorithm called ELABSEG.
The tests were evaluated by the ELSA v.2.3 [Bourjot'91] program (i.e. the pauses were included in the score results).

When comparing the segmentation algorithms care must be taken. For instance, whenever the criteria for boundary placement were not satisfied in FR_SALA, it did not mark any boundary at all but marked this uncertainty with an asterisk.
Another point is that the reference sets, i.e. the manual labelling files, were not identical in the different tests. E.g. for DK_SALA the given manual segmentation and labelling had be to altered manually prior to the automatic segmentation, (the plosives were divided into a closure and a burst part, the affricates were divided into a closure and a fricative part, and all /@/ labels were substituted with /3/).
ELABSEG used the given SAMPA label file as input.

The detailed segmentation performances were compiled in [Barry'91b, pp.1-32], where the main conclusion was that the FR_SALA obtained very poor results compared with ELABSEG and DK_SALA. E.g. for the test on the Test Passage over 50% of the boundaries were either gross-errors or asterisk ones. In table 7.18 the results for the SI-tests with DK_SALA and ELABSEG in [Barry'91b] are summarized.

	Fine errors, ±20ms deviation		*Gross errors*	
Speech corpus	**ELABSEG**	**DK_SALA**	**ELABSEG**	**DK_SALA**
English EUROM0	4013/4651 (85-87%)	3949/5373 (72-75%)	62/4655 (1-2%)	589/5373 (10-12%)
Test Passage	174/245 (65-76%)	221/280 (74-83%)	14/246 (3-9%)	21/281 (5-11%)
Italian EUROM0	3752/5269 (71-73%)	3007/5277 (56-58%)	362/5273 (6-8%)	897/5281 (16-18%)

Table 7.18 *Some results from comparison of the SI-tests for ELABSEG and DK_SALA as given in [Barry'91b]. The fine errors are given as the number of coincidences within ±20ms deviation divided by the total number of boundaries (pauses included), and with the corresponding 95% confidence intervals in parentheses. Similarly the gross errors are counted and divided by the total number of segments (pauses included), and with the corresponding 95% confidence intervals in parentheses.*

From the SI-test results in [Barry'91b] the following conclusions can be drawn:

1) ELABSEG performed *significantly better* than DK_SALA both with respect to fine errors and gross errors on the English and Italian EUROM0 recordings.

2) For the Test Passage DK_SALA obtained highest coincidence rates, but ELABSEG still got the least number of gross-errors. However, since the Test Passage was so short the differences

[3] The universities referred to in this section is UCL - University College London, ICP -Institute de la Communication Parlee, Grenoble, CSRF(CNR)-University of Padova, and CRIN-INERIA, Nancy.

in performance were not statistically significant (the broad confidence intervals overlapped) for this test material.

3) For both algorithms the segmentation performances varied for the different test sets. ELABSEG performed best on the English EUROM0, but had a significant decrease in performance for the Test Passage and Italian EUROM0 recordings.
The DK_SALA performed similarly on the English EUROM0 and the Test Passage, but obtained significantly worse results for Italian EUROM0 recordings.

In order to compare the test-results in [Barry'91b] with the tests in this thesis, *the pauses had to be removed and new 95% confidence intervals computed*, as shown in table 7.19 below.

	Fine errors, ±20ms deviation		*Gross errors*	
Speech corpus	**ELABSEG**	**DK_SALA**	**ELABSEG**	**DK_SALA**
English EUROM0	3633/4271 (84-86%)	3569/4993 (70-73%)	62/4465 (1-2%)	589/5183 (11-12%)
Test Passage	158/229 (63-75%)	205/264 (72-82%)	14/238 (4-10%)	21/273 (5-11%)
Italian EUROM0	3489/4973 (69-71%)	2711/4981 (53-56%)	362/5125 (6-8%)	897/5133 (16-19%)

Table 7.19 *Some results from comparison of the SI-tests for ELABSEG and DK_SALA taken from [Barry'91b], but with the pauses excluded. The fine errors are given as the number of coincidences within ±20ms deviation divided by the total number of (unknown) boundaries, and with the corresponding 95% confidence intervals in parentheses. Similarly the gross errors are counted and divided by the total number of (unknown) segments, and with the corresponding 95% confidence intervals in parentheses.*

The new results in table 7.19 did not change the conclusions above. We note that the results in table 7.19 are better than those described in the previous sub-sections of our automatic segmentation algorithm on the English EUROM0 recordings. The discrepancy may be due to printing errors, or the UCL's SI-tests and the CSRF's TI-test may have included parts of the training set in their test sets, providing too good results for that tests. However, this does not change the conclusions above either.

7.3.3 SENSITIVITY OF AUTOMATIC SEGMENTATION RESULTS TO VARIATIONS IN MANUAL ANNOTATION

The comparison of tests in the previous section showed that the automatic segmentation performance of one algorithm varied significantly when testing on different test sets. One possible reason for these variations is the different recording conditions. This motivates for incorporation of a speaker or microphone normalisation in the preprocessing step of the ASS algorithm. Another reason may be the difference in manual annotation. E.g. the DK_SALA algorithm obtained similar results on the English EUROM0 and the Test Passage (which had different recording conditions but similar manual annotation), but obtained significantly worse results for the Italian EUROM0 recording (with different recording conditions and annotation strategy).

Due to lack of standards for manual annotation of speech, a speech recording will be annotated differently by different human labellers, and the same labeller may not be able to annotate consistently. Since a manual annotation nonetheless is used as reference for assessing automatic segmentation, the effect that different manual annotations have on the automatic segmentation performance results should be investigated.

As an example, the word (and sentence) "but" /bVt/ excerpted from the Test Passage is depicted in figure 7.6 with the automatically (acoustic and phonemic) and manually placed segment boundaries indicated. The acoustic segment boundaries calculated with 150% o.s. were used to constrain the phonemic segmentation. Here, the /t/ is preceded and accompanied by a glottal stop, where the glottal and alveolar closure are probably released simultaneously. The period of creaky voice after the vowel is assigned to the closure part of the plosive by the human labeller. However, since the excitation pulses are voiced and have a vowel quality, it is also "correct" to include the creaky area in the vowel segment, as argued for in [Kvale'91]. This is exactly what is done by the phonemic segmentation pass of our ASS procedure, as shown in figure 7.6.

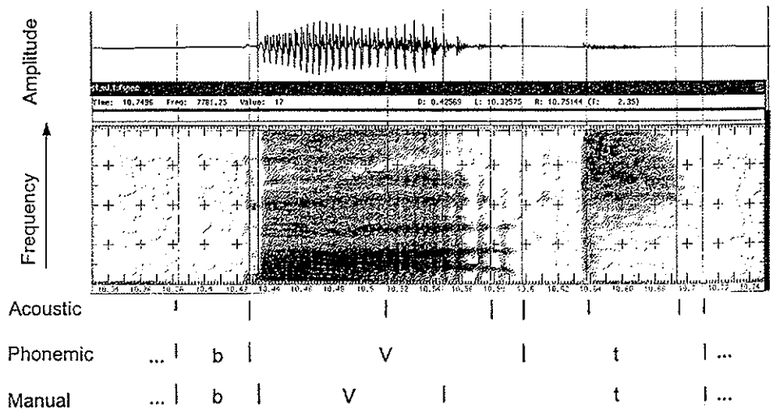


Figure 7.6 *The speech waveform and the broadband spectrogram of the word "but" /bVt/ excerpted from the Test Passage. The acoustic and phonemic segmentation are shown together with the manual annotation. The + signs in the spectrogram are separated by 0.02 sec. horizontally and 2000 Hz vertically.*

As a pilot study, the annotation of the Test Passage was further analyzed. In Appendix C.3 we have argued for several possible alternative annotations. We gradually incorporated these alternatives in the label file of the Test Passage. With altered label file, a new segmentation was run and new assessments were performed.

Details of the different changes are shown in Appendix C.3. The *deviation plot* in figure 7.7 below exemplifies the effect of using another label file for the third sentence of the Test Passage by comparing the deviations with the results when using the original label file.
(The deviation plot shows the difference between corresponding manually and automatically placed segmentation boundaries. A *negative deviation* means that the automatic boundary is marked off *later* than the corresponding manual one).

In the second correction alternative of sentence three, three changes were made (see Appendix C.3), at the boundaries indicated with arrows in the left part of figure 7.7. The end boundary of /r/ was moved into the formant-transition to fulfil the half-way convention [Barry'90b], the end boundary of /k/ was moved to incorporate a burst in the plosive segment, and the /t/ was removed. The deviation plot to the right in figure 7.7 shows that the first two changes reduced the deviation between manual and automatic segmentation, whereas removing /t/ increased the deviation at the succeeding transitions /e/-/n/ and /n/-/@/.

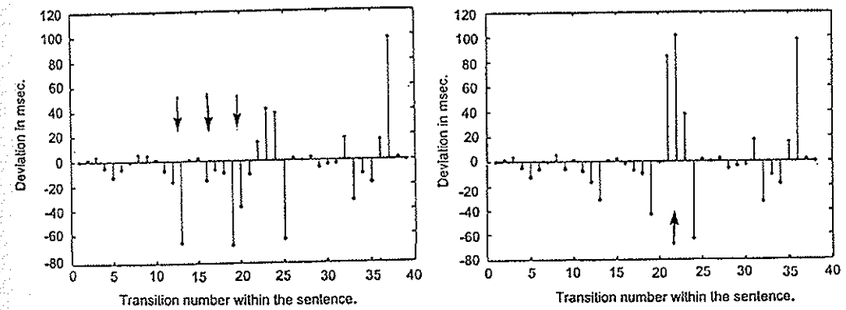


Figure 7.7 *Deviation plots for sentence 3 of the Test passage, showing the effect of alternative annotation (results when comparing with the original label file to the left, and with the altered label file to the right).*

Incorporating all changes implied 4 fewer gross errors (i.e. 29% less) and 15 boundaries more were detected within ±20ms deviation, (i.e. 9% increase).

The Test Passage is too short to establish general conclusions upon. However, this investigation may indicate the uncertainty in the comparison between manual and automatic segmentation. *That is, about 90-95% coincidence with manual segmentation within ±20ms deviation may be an upper limit for automatic segmentation performance.*

7.4 IMPROVING THE PHONEMIC SEGMENTATION

The experiments in the previous sections showed that the acoustic segmentation with higher oversegmentation factor than 75% obtained high DTW-coincidence with manual segmentation whereas the phonemic segmentation was significantly less accurate. E.g. with 150% oversegmentation, 97.4% of the acoustic segment boundaries DTW-coincided within ±20ms deviation from the manual segmentation for the English EUROM0 recordings. However, the corresponding SI-tests with phoneme HMM segmentation constrained by these acoustic segment boundaries obtained only 80.4% coincidence rate within the ±20ms deviation threshold, and the pure HMM segmentation performed even worse.

Hence, the phonemic segmentation part should be improved. This section describes the HTK-software (HMM Tool Kit) [Young'92] to optimise the parametrisations and HMM topologies for phonemic segmentation. The HTK is a flexible toolkit for building subword based CDHMM recognisers. Since it is much more flexible and easy to use for generating HMMs than our own software, we decided to modify it and use it for further improvements of the phonemic segmentation.

Phonemic segmentation, or label alignment, can be regarded as phoneme recognition with a very restricted grammar; namely the phoneme string itself. The wanted outputs from the Viterbi algorithm are then only the time instances for the changes from the last state in one phoneme model to the first state in the following phoneme model. Thus, the only modification needed to the HTK-software was to incorporate the acoustic segment boundaries' constraint on the Viterbi forward recursion.

In HTK there were some slight differences from our original parametrisation:
i) The LPC-cepstrum was computed from the LPC representation by the recursion in eq. (5.11) *but the LPC-analysis order was in HTK required to be equal to or bigger than the number of cepstral coefficients.*
ii) The delta cepstrum coefficients were computed according to eq. (5.17), and *delta energy was computed by the same formula.* We used K=2 frames on each side of the actual frame for delta-calculations. At the edges of the data file, simple first order differences, as in eq.(7.5), were used.
iii) The HTK did not offer the possibility to compute delta energy without computing delta cepstrum as well.
iv) An initial estimate of means and variances of each phoneme HMM (the transition probabilities remain unchanged) was obtained by a segmental k-means procedure, i.e. first each phoneme was uniformly segmented and then nine iterations of Viterbi alignment were performed. Then the models were trained by means of the Baum-Welsh reestimation procedure.
However, these changes had only minor effects on the phonemic segmentation accuracy.

7.4.1 OPTIMISING PARAMETERS

In this section the effect of different parameter sets and training strategies are shown with respect to phonemic segmentation. Experiments with the **English SI-test** showed that the phonemic segmentation performance was improved by increasing the number of cepstrum coefficients up to about 18 coefficients, proving that our initial choice of cepstral order was reasonable. However, the phonemic segmentation accuracy was relatively insensitive to the number of

coefficients in the cepstral vector. For instance, the coincidence rate within ±20ms deviation varied with less than 1% when varying the number of cepstral coefficients in the frame vector between 12 and 18.

The phonemic segmentation accuracy **improved** by adding successively *absolute normalised energy*, E, *delta energy*, ΔE, *delta cepstrum*, Δcep, and *sine liftering*.
Adding more states and transitions in the phoneme CDHMMs changed the phonemic segmentation accuracy only slightly.

On the other hand, *increasing from one to two mixtures* for modelling the output probabilities in the HMM states, **decreased** the phonemic segmentation performance. This may be due to that twice as many parameters had to be estimated from the same small training set, and these new parameters did not add new information. (This is in contrast to the introduction of delta-cepstrum which also doubled the number of parameters but which added new information about the temporal evolution of speech and thus improved the segmentation performance).
Embedded re-estimation also **reduced** the phonemic segmentation accuracy; especially the coincidence rate at ±5ms deviation. This is natural since embedded re-estimation only uses the label string as input for the training, i.e. not the manually placed segment boundaries, and will thus not learn the details in the manual segmentation.

If the bias was not removed from the English EUROM0 recordings prior to the training of the phoneme HMMs (see section 7.1), the segmentation accuracy became only slightly deteriorated (about 1% less coincidence rate at all deviation thresholds) for tests on the same recording. However, when models trained on the recording with bias were used for segmenting the Test Passage (without bias), the segmentation performance was seriously deteriorated - especially with MFCC-parametrisation.

We also compared the effect of LPC-cepstra (LPCCEP) versus mel-scale cepstra (MFCC) with the parametrisation: *pre-emphasis 0.95, 12cep., 12Δcep., E, ΔE, and sine liftering (L=22)*, and with the simple phoneme CDHMM defined in section 7.1. The segmentation results for the **English SI-test** shown in table 7.20, are given as an average of the four speakers.

	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
LPCCEP, V	1.6 - 2.4%	43.5 - 46.5%	65.1 - 67.9%	83.1 - 85.3%	86.6 - 88.6%
LPCCEP, VC_150	1.3 - 2.1%	49.8 - 52.7%	66.8 - 69.6%	83.0 - 85.2%	86.6 - 88.6%
MFCC, V	1.5 - 2.3%	46.2 - 49.2%	66.3 - 69.1%	84.9 - 87.0%	88.0 - 89.9%
MFCC, VC_150	1.1 - 1.8%	50.6 - 53.6%	67.8 - 70.6%	83.8 - 86.0%	87.3 - 89.3%

Table 7.20 *Gross errors and fine errors for the SI-test on the English EUROM0 for the pure HMM segmentation (V) and constrained HMM segmentation (VC) (150% o.s.), using either LPC-derived (LPCCEP) or mel-scale based (MFCC) cepstrum coefficients. The gross error results are given as 95% confidence intervals for the number of gross errors, whereas the fine errors are given as 95% confidence intervals for the coincidence rates within each deviation threshold.*

Compared to the phonemic segmentation accuracy obtained with the parametrisation used in section 7.3, i.e. with 18LPCCEP + E + ΔE, (table 7.7), the performances with LPCCEP in table 7.20 are significantly improved for the pure HMM segmentation (V). The constrained HMM segmentation (VC) performed also better, but for coincidence rates within ±5ms deviation and for the gross errors, the improvements were not significant.

In table 7.7 the constrained HMM segmentation (VC) performed significantly better than the pure HMM segmentation (V) for gross errors and for coincidence rates within ±5ms and ±10ms deviation. The constraints also improved the performances at the other deviation thresholds.
With optimal parameter setting, table 7.20, the VC performed significantly better than V for the coincidence rate within ±5ms deviation only. The VC in table 7.20 performed also better than V at ±10ms deviation and reduced the number of gross errors, but these improvements were not significant. When bigger deviations were allowed for, constraining with acoustic segment boundaries did not influence the coincidence rates.

When it comes to the effect of LPCCEP versus MFCC parametrisation, the phonemic segmentation based on MFCC obtained highest coincidence rates and least number of gross errors. Thus, the property that LPC-cepstra fit the spectral peaks best (chapter 5.3), was not enough to perform better than MFCC for automatic segmentation. Although the difference in segmentation performance between the two parameter sets was not significant, the MFCC was selected for the rest of the experiments in this thesis.
Thus, the following parameter set will be used: *pre-emphasis 0.95, 12MFCCs, 12ΔMFCCs, absolute normalised energy, E and ΔE, and sine liftering (L=22), and the CDHMM configuration defined in section 7.1.*

Comparison of the SI, TI, and Full HMM tests with the optimal parameter set, table D.13 and D.14, showed no significant difference between the SI and TI test, whereas the Full HMM test performed significantly better than the SI test.

7.4.2 PHONEME-CLASSES

The EUROM0 recordings were not phonemically balanced. Too few training tokens for some phonemes implied unreliable model estimates. For instance the English EUROM0 recordings contained only 4 segments labelled /e@/ and 5 segments labelled /Z/.
There exist no general rule of how many training token that are optimal for HMM based phonemic segmentation, and probably will the optimal number of training tokens depend on the given test corpus. One study, [Schmidt'91], indicated that the segmentation performance increased with the number of phonemes used in training the phoneme HMMs. However, increasing the number of training tokens above 100 yielded only marginal improvements in segmentation accuracy for the given test corpus.

Thus, our training set for each language should be enlarged. If no other annotated speech material for a language is available, one may utilise the existing by *grouping the phonemes into phoneme-classes* and thereby increasing the number of occurrences for each class. Training and testing with phoneme-classes cause no problem for the automatic segmentation because the label string is known to the segmentation.

As an example, 11 phoneme classes were defined for **English** as:
bilabial plosives /p b/, *alveolar plosives* /t d/, *velar plosives* /k g/, *fricatives and affricates* /f s S z Z T h tS dZ/, *approximants* /v D/, *front close vowels* /I i: i/, *central and mid-close front vowels* /e { @ 3:/, *back vowels* /Q V U u u: A: O:/, *diphthongs* /eI aI aU OI @U I@ e@ U@/, *nasals* /m n N/, and *semivowels* /w j l r/.
With this phoneme grouping no class occurred less than 140 times and the maximum number of occurrences for a class was 700 for the English EUROM0 recordings. Thus, for the SI-test at least 105 training tokens were available for each class.

Since phonemes occurring more than 100 times in the given speech corpus may provide enough training material for reliable model estimates, these phonemes may not be grouped into classes. Thus in another experiment the phonemes /p t k b v D m r l @/ were not grouped. To ensure larger training material for the other phonemes, they were grouped on a phonetically basis as:
/d g/, /tS dZ/, /n N/, /w j/, /I i: i/, /e { 3:/, /f T h/, /s S/, /z Z/, /O: Q/, /V A:/, /u u: U/, /@U U@/, /I@ e@/, and /eI aI OI/.
One of these classes had only 30 training tokens, the others included more than 50.

However, *using the phoneme-classes defined above for training and segmentation decreased the coincidence rates significantly* (by about 4% at all deviation thresholds).
One reason for the reduced segmentation accuracy may be that there were too many transitions between two phonemes belonging to the same class which in most cases were more difficult to segment (see e.g. table 7.9). Another serious problem is that although the grouping of phonemes into classes increased the training material for each label, the *variance within the classes also increased.*

In order to reduce the variances within the models while maintaining the effect of many training tokens, the phoneme class HMMs were only used for initialising the individual phoneme HMMs. That is, the initial phoneme HMMs were a copy of the class HMM they belonged to. Then the *phoneme models* were trained with Baum-Welsh reestimation.
This approach improved the segmentation performance over that of using phoneme class models, but the results were still slightly (not significantly) worse than those in table 7.20. (The coincidence rates were about 1% less at all deviation thresholds).

That is, grouping into phoneme classes in the training process destroys phoneme-specific information valuable for accurate phoneme segmentation. In section 7.4.6, the possibility enlarging the phoneme training set by using phonemes from other languages is investigated.

7.4.3 COMPARING THE SI-TESTS ACROSS LANGUAGES

Using the optimal MFCC based parameter set defined in section 7.4.1, the following segmentation performances were obtained for the SI-tests on the Norwegian, Danish, Italian, and English EUROM0 recordings (detailed results are provided in Appendix D, section D.4, D.5, D.6, and D.7).

	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
English, SI, V	1.8 %	47.7 %	67.7 %	86.0 %	89.0 %
English, SI, VC	1.4 %	52.1 %	69.2 %	84.9 %	88.3 %
Italian, SI, V	3.2 %	43.5 %	62.0 %	80.6 %	84.9 %
Italian, SI, VC	2.0 %	44.7 %	62.2 %	81.9 %	86.0 %
Norwegian, SI, V	1.6 %	48.5 %	68.1 %	85.4 %	88.3 %
Norwegian, SI, VC	1.4 %	47.5 %	66.1 %	85.3 %	88.3 %
Danish, SI, V	3.4 %	44.2 %	62.9 %	79.5 %	82.8 %
Danish, SI, VC	2.9 %	46.8 %	64.7 %	80.0 %	83.5 %

Table 7.21 *Coincidence rates within different deviation thresholds and percent gross errors for the English, Norwegian, Italian, and Danish SI-tests, with pure HMM segmentation (V) and constrained HMM segmentation (VC) with acoustic segmentation at 150% o.s.*

Compared to the phonemic segmentation performances on the English and Italian SI-tests obtained with the simple parameter setting (see section 7.3.2.5, table 7.17 and 7.19, and [Svendsen'90]) *the optimal parameter set defined in section 7.4.1 clearly yielded improvements.* For instance, with the simple parameter set, 362 gross errors occurred in the constrained HMM segmentation of Italian (table 7.19), whereas for the optimal parameter set only 104 gross errors occurred, i.e. an improvement of 71%. For fine errors for Italian, e.g. the number of coincidences within ±20ms deviation increased by 585 boundaries, or 16.5%.

The results for Danish EUROM0 in table 7.21 were significantly better than the results obtained by the Danish algorithm DK_SALA tested in [Barry'91b]. For the coincidence rate within ±20ms, DK_SALA obtained a 95% confidence interval of 55-58%, whereas our algorithm obtained a corresponding interval of 79-81%. DK_SALA got 18-20% gross errors (with 95% confidence), whereas our algorithm achieved 2-3% as 95% confidence interval for the gross errors.

Constraining the Viterbi forward recursion with acoustic segment boundaries *reduced the number of gross errors* compared to pure HMM segmentation for all languages. For Italian, (and Danish with 100% o.s.), the reductions in gross errors were significant, see table D.28 and D.25 respectively. In addition, the largest fine error deviation for each speaker (except JHE) was

reduced with constrained HMM segmentation.

With respect to fine errors the constrained Viterbi (VC) increased the number of boundaries placed within ±5ms deviation significantly for English, but for bigger deviations the performances with VC were slightly worse than with pure HMM segmentation. For Norwegian the constraining of the Viterbi recursion had almost no influence. For Italian and Danish the constrained HMM segmentation increased the coincidence rates (not significantly) at all deviation thresholds. For Danish, constraining with acoustic boundaries at 100% o.s. further improved the phonemic segmentation performance both with respect to fine and gross errors.

Difference between speakers

The segmentation performance for the *Italian* speakers varied. For speaker GCI there were significantly more gross-errors than for the other speakers. Also significantly fewer coincidences within ±20ms and ±25ms deviations were obtained for speaker GCI. The best performances were obtained for LVI and PUI. See table D.29. Constrained HMM segmentation improved the performance over that of pure HMM segmentation with respect to both gross errors and fine errors for all speakers.

For the *Danish* speakers the segmentation accuracy with respect to fine errors was worse than average for speaker BLD, around average for CHD and LFD, and above average for speaker JDD. The constrained HMM reduced the number of gross errors for all speakers and improved the segmentation performance with respect to fine errors for speaker BLD and JDD, whereas for speaker CHD and LFD the constraints with acoustic segment boundaries had no effect. See table D.26.

The segmentation performance was similar for the English speakers, both with pure and constrained HMM segmentation. See table D.15. The segmentation performances for the individual speakers in the *Norwegian* EUROM0 will be discussed in section 7.4.4.1.

Difference between languages

The SI-tests for English and Norwegian EUROM0 gave very similar results, whereas for Italian and Danish the performance was slightly worse. The robustness of the segmentation algorithm to different languages may be measured with the standard deviation of the average number of gross errors or coincidence rates for the different languages. With this measure the results summarised in table 7.21 show that with constrained HMM segmentation the variation in segmentation results for different languages was less than the corresponding variation with pure HMM segmentation. E.g. the standard deviation for the coincidence rates within ±20ms deviation threshold for the four languages was 2.86 with pure HMM segmentation and 2.18 with constrained HMM segmentation. The standard deviation for the gross errors was correspondingly reduced from 2.08 to 1.42 with constrained HMM segmentation.

We may also apply a so-called ANOVA (ANalysis Of VAriance) test [Johnson'88], to measure the statistical significance of the robustness measure. The ANOVA test will indicate whether the difference with respect to fine errors and gross errors between the languages is larger than the variation in segmentation performance on the different speakers within a single language.
For this test we regard the number of gross-errors for speaker i, in language j, as a stochastic

variable X_{ij}, and assume that X_{ij}, i=1,4, j∈(Norwegian, English, Italian, Danish) is Gaussian distributed $N(\mu_j, \sigma^2)$, where μ_j is the mean number of gross errors for langauge j and σ is the standard deviation for the gross errors. Here σ was assumed to be equal for all languages. With the ANOVA we tested the null hypotheses that the phonemic segmentation for the different languages had equal mean value for the gross errors, H_0: $\mu_1=\mu_2=\mu_3=\mu_4$ against the alternative hypotheses H_A: $\mu_j \neq \mu_k$ for at least one language j≠k.

The ANOVA test indicated that pure HMM segmentation of the different languages obtained different performance with respect to gross errors with a significance level of 14% (p=0.14), whereas with the constrained HMM segmentation the number gross errors for the different languages did not differ significantly (p=0.43).

Using the same approach for the fine errors, e.g. for coincidence rates within the ±20ms deviation threshold, the ANOVA test indicated that the coincidence rates for the different languages differed significantly (p=0.06) with pure HMM segmentation but differed less (p=0.15) with the constrained HMM segmentation.

Based on only 4 speakers for each language it is difficult to make a general statement about the difference in segmentation performance between the languages. However, if we assume that the speakers are representative of their respective languages, we may conclude that the performance of pure HMM segmentation differed significantly for the different languages, whereas constraining the Viterbi recursion reduced the differences in segmentation performance between the languages. In addition, the overall standard deviations for the gross errors and fine errors for the languages were reduced by constraining the Viterbi recursion[4].
The difference in segmentation performance for the different speech corpora is probably due to the different recording conditions, the different manual segmentation strategies, and the different phoneme inventories.

[4] For instance, for gross errors σ was 0.61 with pure HMM segmentation and 0.53 with constrained HMM segmentation, whereas for coincidence rates within the ±20ms deviation threshold σ was 1.84 with pure HMM segmentation and 1.58 with constrained HMM segmentation.

7.4.4 TESTS ON NORWEGIAN

The Norwegian EUROM0 recordings have been carefully analyzed both with respect to manual annotation conventions (chapter 4) and with respect to the performance of the acoustic segmentation algorithm (section 7.2). It was therefore natural to examine how the phonemic segmentation with optimal parameter set performed on this speech material. First the quantitative and qualitative analyses are briefly discussed. Then the effect acoustic segmentation had to the constrained HMM segmentation is described.

7.4.4.1 Quantitative analyses

The phonemic segmentation performances for the SI, TI, and Full HMM tests for the Norwegian EUROM0 recordings are shown in table D.16 and D.17. The division of the speech corpus for the Norwegian TI-test was as defined in Appendix B, and the results for the TI-tests are presented as an average of the two parts since the results were similar for each of them.

Length is a phonemic feature for the Norwegian vowels. Since many phonemically long vowels were acoustically short and phonemically short vowels were acoustically long (see figure B.1 in Appendix B), we though that joining corresponding short and long vowels, e.g. /A/ and /A:/, would increase the training material and thereby improve the segmentation accuracy for vowels. Table D.17 shows that the grouping gave no significant changes in the segmentation performance, but the number of gross-errors decreased slightly.

The segmentation performance on the different speakers in the Norwegian SI-test varied, see table D.18 and D.19. E.g. for speaker TBN there were significantly more gross errors than for the other speakers. Constraining with acoustic segment boundaries (with 150% o.s.) reduced the number of gross errors for speaker TBN and SHN, but for speaker AFN there was no difference, and for speaker TGN the number of gross-errors increased. For speaker AFN the acoustic segmentation with 150% o.s. performed worse than for the other speakers (table 7.2), and constraining the Viterbi forward recursion with these boundaries reduced the coincidence rate slightly compared to pure HMM segmentation. For speaker TBN the constrained Viterbi increased the coincidence rates within ±5ms, ±20ms, and ±25ms deviation thresholds, whereas for speaker TGN the coincidence rates decreased. However, none of the changes were significant.

Comparing the segmentation performances for the different tests with the corresponding tests for English, we notice that constraining the Viterbi recursion increased the coincidence rate within ±5ms for all English tests but had no influence on the Norwegian tests. E.g. for the SI-test the pure HMM segmentation performed similarly on English and Norwegian, whereas the constrained HMM segmentation obtained significantly higher coincidence rates within ±5ms and ±10ms deviation thresholds for English.
For the TI-tests significantly better results where obtained on the Norwegian recordings, both with respect to gross errors and fine errors (expect for the coincidence rate within ±5ms and ±10ms deviation thresholds for the constrained HMM segmentation).
With Full HMM the algorithm performed similarly on Norwegian and English (expect for the coincidence rate within ±5ms deviation thresholds for the constrained HMM segmentation).

7.4.4.2 Qualitative analyses

In figure 7.8 below the first sentence by speaker AFN is segmented by six different means:

(1) "full HMMs" constrained with acoustic segmentation at 100% o.s.,
(2) "full HMMs" for the pure HMM segmentation ,
(3) acoustic segmentation with 100% o.s.,
(4) SI-test constrained with acoustic segmentation at 100% o.s.,
(5) SI-test for the pure HMM segmentation,
(6) the manual segmentation.

Most of the automatically placed boundaries coincided with the corresponding manually placed boundaries. Phonemes that were realised with two or more acoustically "stable areas", as plosives and some realisations of /r/, got all the parts into the same phoneme segment. However, in the example below only the *problems* are discussed in more detail:

A. The /r/-/O:/ boundary was placed too late by the pure HMM segmentation both for the SI-test and the full HMM test, (5) and (2). The constrained Viterbi segmentation obviously selected the acoustic boundary closest to that of the pure HMM segmentation, so that (1) and (4) also segmented too late even if the acoustic segmentation provided a more correct boundary alternative.

B. The /n/-/v/ transition was difficult to segment. In (5) the /n/-segment was only 10ms, imposing a very long /v/ segment. In the constrained HMM segmentation (4), the acoustic boundary forced the /n/ to be 30ms. Note that although the automatic segmentation missed that boundary, the end of /A/ and beginning of /i/ were "correct".
With full HMM the segmentation was reasonable although the end of /n/ did not coincide exactly with the manual segmentation.

C. The /r/-/i:/ transition was smooth. The SI-test, (4) and (5), performed well. For full HMM the constraining with acoustic segmentation made the Viterbi decoding select the acoustic boundary which marked the end of the [@] part of the /r/-segment.

D. The /v/-/e/-/}/-/e/ sequence was difficult for the automatic segmentation. The /v/ and /e/ was included in one long /v/-segment. For pure HMM segmentation, (2) and (5), the /e/-/}/ boundary was placed in the middle of the area of creak (as in manual segmentation). The acoustic segmentation gave no boundary alternative in the area of creak, so the /e/-/}/ boundary was placed at the end of this area with the constrained HMM segmentation, (1) and (4).
Also for the /}/-/e/ transition the pure HMM segmentation computed boundaries closer to the manual ones than the constrained phonemic segmentation did. In this example the constraining amplified the error of the pure HMM segmentation: The boundary in (5) was a little too late, in (4) it became even later. The boundary in (2) was a little too early, in (1) it was even earlier.

The following /e/-/n/ transition (not shown in figure 7.8) was correctly segmented. This indicates that the phonemic segmentation was not prone to ripple errors.

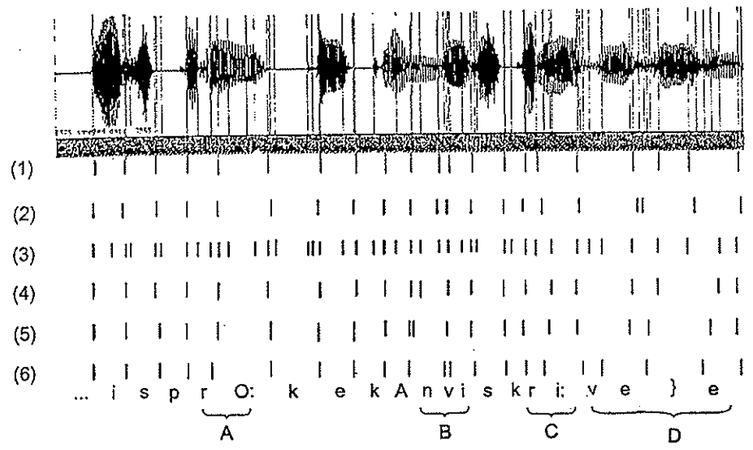


Figure 7.8 *Comparison of six different segmentations of the first sentence of speaker AFN. The numbers correspond to segmentation method defined in the text.*

Based on similar qualitative analysis of the phonemic segmentation of a larger portion of the EUROM0 recordings some **general trends** were:

-The automatic segmentation seldom placed the boundary in the middle of an area of creaky voice.
-Constrained HMM segmentation selected the acoustic boundaries closest to the ones computed by pure HMM segmentation. So, when the pure HMM segmentation performed well, the acoustic segmentation helped to tune the boundaries.
-The constrained HMM segmentation algorithm was very robust against ripple effects. That is, one huge deviation from the manual segmentation did rarely cause many errors later in the utterance. The algorithm seems to manage to get "into phase" again by locking to a plosive burst or a transition between voiceless and voiced sounds.
-When errors were caused by a ripple effect, the computed boundaries as such were often correctly placed to the speech signal, but the corresponding labels were placed at one segment too early or too late compared with the manual ones.
-The phonemic segmentation algorithms performed equally well on long and short utterances. Thus, dividing long sentences into smaller ones by artificial silence symbols of no duration, will probably not improve the segmentation accuracy.
-The closure part of a plosive was often included in the preceding phoneme segment. Hence, the plosive segment often contained the burst only. However, since the bursts usually have short durations, this error will not be detected by the quantitative analyses.

7.4.4.3 Acoustic segmentation effects on the constrained segmentation accuracy

Using the acoustic segmentation boundaries to constrain the following phonemic segmentation pass, the final phonemic segmentation accuracy cannot be better than the acoustic segmentation accuracy. In this section we analyze how variations in acoustic oversegmentation-factor influenced the constrained HMM phonemic segmentation accuracy.

Table 7.22 below compares the number of coincidences and gross errors for pure HMM segmentation and constrained HMM segmentation with varying degree of oversegmentation. The corresponding numbers and 95% confidence intervals are shown in table D.20 and D.21 respectively.

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	1.6 %	48.5 %	68.1 %	85.4 %	88.3 %
SI, VC, 225% o.s.	1.2 %	47.2 %	65.9 %	85.0 %	88.1 %
SI, VC, 200% o.s.	1.2 %	47.4 %	66.0 %	85.1 %	88.3 %
SI, VC, 175% o.s.	1.0 %	47.5 %	66.1 %	85.3 %	88.3 %
SI, VC, 150% o.s.	1.4 %	47.2 %	65.7 %	84.5 %	88.1 %
SI, VC, 125% o.s.	1.1 %	47.6 %	66.3 %	84.5 %	88.0 %
SI, VC, 100% o.s.	1.4 %	46.8 %	65.4 %	84.3 %	87.9 %
SI, VC, 75% o.s.	1.6 %	45.5 %	64.4 %	83.1 %	86.7 %
SI, VC, 50% o.s.	1.6 %	43.3 %	62.2 %	80.4 %	84.1 %
SI, VC, 25% o.s.	5.6 %	36.7 %	53.0 %	70.0 %	73.9 %
SI, VC, 0% o.s.	28.0 %	19.6 %	28.2 %	35.8 %	38.2 %

Table 7.22 *Coincidence rates within different deviation thresholds and percentage gross errors for the SI-test on the Norwegian EUROM0 for the pure HMM segmentation (V) and constrained HMM segmentation (VC), using different oversegmentation factors in the acoustic segmentation.*

The constrained HMM segmentation was rather insensitive to variations in o.s.-factors higher than 75%. However, when the o.s.-factor was reduced below 75%, the segmentation accuracy decreased severely, as it did for the acoustic segmentation analyzed in section 7.2.

With respect to *gross errors* the constrained HMM segmentation with o.s.-factors bigger than 50% performed **better** than the pure HMM segmentation. Least gross errors were obtained with 175% o.s., which got 37.5% less gross errors than with pure HMM segmentation. However, with 95% confidence, the improvements were not significant, see table D.21.

With respect to *fine errors* the constrained HMM segmentation for all oversegmentation factors

performed slightly worse than pure HMM segmentation. However, the performance of the constrained HMM segmentation was not significantly worse for o.s.-factors bigger than 75%.

We conclude that for the constrained HMM segmentation the optimum number of acoustic segments is between 1.75 and 3 times the number of phonemes. This is also intuitively a reasonable number of acoustic segments because many phonemes are made up of two or more spectrally stable areas (see also qualitative analyses of acoustic segmentation in section 7.2.3 and 7.2.4.2).

Although the acoustic segmentation obtained high DTW-coincidence rates with manual segmentation, it will sometimes provide a bad input to the constrained HMM segmentation pass, and the phonemic segmentation might have calculated a correct boundary if the acoustic segmentation did not force it to select the wrong one. E.g. for the period of creaky voice between two vowels, figure 7.1 and 7.8, the human labeller placed the boundary in the middle of the creaky area, whereas the acoustic segmentation marked a boundary on each side of the period of creak. With a corrected acoustic segmentation the constrained HMM segmentation will not be forced to select such "wrong" acoustic boundaries.

In order to examine how many of the phonemic segmentation errors that originated from bad acoustic segmentation, we substituted the acoustic boundaries closest to the "correct" manually placed phoneme boundaries by the manual boundaries, denoted *corrected* acoustic segmentation in table 7.23 below. (That is, the corrected acoustic segmentation obtained 100% DTW-coincidence with the manual segmentation).

If the constrained HMM segmentation now simply selected the same acoustic boundary sequence as before the correction, the performance would improve most at the ±5ms and ±10ms deviation thresholds. Table 7.23 shows that correcting the acoustic segmentation improved the constrained HMM segmentation performance for the Norwegian SI-test, especially at small deviation thresholds. The improvements in coincidence rates were significant (see table D.23), whereas the number of gross errors decreased only slightly.

However, the phonemic segmentation was still far from 100% "correct" even when the acoustic segmentation yielded 100% DTW-coincidence.

SI, VC, 100% o.s.	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Ordinary	1.4 %	46.8 %	65.4 %	84.3 %	87.9 %
Corrected	1.3 %	80.6 %	83.7 %	89.4 %	92.4 %
Uniform	0.5 %	19.1 %	36.0 %	69.6 %	80.0 %

Table 7.23 *Coincidence rates within different deviation thresholds and percentage gross for the constrained HMM segmentation (VC) of the SI-test on the Norwegian EUROM0, using different types of acoustic segmentation at 100% o.s.. (See text).*

Constraining the Viterbi recursion with the *uniform* segmentation boundaries, (see section 7.2.4.1), reduced the number of gross errors significantly. Because the "acoustic" boundaries were equally spaced, ripple errors were totally avoided.
However, the coincidence rates were significantly lower than when using "ordinary" acoustic segmentation as constraints, and particularly at ±5ms and ±10ms deviation thresholds the coincidence rates became very low with the uniform segmentation.

In section 7.2.4.1 the uniform segmentation performed comparable with the acoustic segmentation within the ±25ms deviation threshold. However, when the uniform segmentation boundaries were used to constrain the Viterbi recursion, the phonemic segmentation become much worse than when constraining with acoustic segment boundaries. Thus, when optimising the acoustic segmentation, it is *not* sufficient only to optimise the DTW-coincidence.

7.4.5 CEPSTRAL DOMAIN FILTERING

For the SI-tests on the EUROM0 recordings, the phoneme HMMs were trained on three speakers and tested on the fourth. It is thus important to suppress the individual speakers characteristics, such as the spectral tilt, prior to the training of the phoneme HMMs. Speaker and environment normalisation is also necessary when HMMs trained on speech recorded in one recording environment are used for a test on speech recorded under different recording conditions.

Several methods for inter-speaker normalisation and environmental adaption have been proposed for improving automatic speech recognition, such as in [Matsumoto'79], [Shiraki'90], [Zahorian'91], [Junqua'93], [Hermansky'92], and [Hanson'93]. In this section we will investigate the effects *cepstral domain filtering* (described in section 5.3.3) have on the phonemic segmentation performance.
For these experiments we applied the RASTA filter, eq. (5.18), with k=0.1 and N=5, to the MFCCs (optimal set defined in section 7.4.1).

The main conclusion was that the cepstral domain filtering *improved* the phonemic segmentation performance for all languages, and the segmentation results were rather *insensitive* to different pole-values in the RASTA-filter. For instance for the Norwegian SI-test, the segmentation performance was improved by the RASTA processed MFCCs for all pole values between ρ=0.60 to 0.99, with an optimum for ρ=0.88 (see figure 7.12). For the Italian SI-test all pole-placements improved the segmentation performance, and even delta-cepstrum alone increased the segmentation (see tables D.32, D.33, D.42, and D.43).
Thus the time constant of the RASTA-filter (figure 5.3) did not influence the segmentation performance significantly.

Similarly we found the pole value that yielded best phonemic segmentation performance for the other languages, to be ρ=0.96 for English, ρ=0.90 for Italian, and ρ=0.80 for Danish.
Table 7.24 shows the SI-test results for the different languages when using RASTA-filter with optimal pole on the MFCCs (complementary results are provided in section D.8 in Appendix D).

For all languages the number of gross errors was *significantly reduced* for the pure HMM segmentation, (cf. table 7.21). With pure HMM segmentation also the coincidence rates within deviation thresholds larger than ±5ms *increased significantly* for all languages except English. The RASTA-filtering also improved the constrained HMM segmentation performances, but only the coincidence rates within ±25ms deviation for Italian and Danish were improved significantly.

With the robustness measure defined in section 7.4.3 we found that the segmentation performances on the different languages differed less with the RASTA-processed MFCCs than with the ordinary MFCCs (table 7.21). E.g. the standard deviation for the coincidence rates within ±20ms deviation threshold for the four languages was now 2.08 with pure HMM segmentation and 1.42 with constrained HMM segmentation. The standard deviation for the gross errors was correspondingly 0.60 with pure HMM segmentation and 0.52 with constrained HMM segmentation.

RASTA with optimal pole	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Eng., ρ=0.96, SI, V	1.0 %	48.2 %	68.7 %	87.3 %	90.3 %
Eng., ρ=0.96, SI, VC	1.1 %	52.8 %	70.3 %	86.1 %	89.7 %
Ita., ρ=0.90, SI, V	1.7 %	45.2 %	66.5 %	85.9 %	89.8 %
Ita., ρ=0.90, SI, VC	1.5 %	46.2 %	64.2 %	84.5 %	88.6 %
Dan., ρ=0.80, SI, V	2.4 %	46.3 %	66.1 %	82.8 %	86.8 %
Dan., ρ=0.80, SI, VC	2.2 %	48.4 %	66.2 %	82.3 %	85.8 %
Nor., ρ=0.88, SI, V	0.9 %	49.8 %	71.1 %	88.6 %	91.4 %
Nor., ρ=0.88, SI, VC	0.8 %	48.9 %	68.1 %	86.4 %	89.9 %

Table 7.24 *Coincidence rates and percentage gross errors for the SI-tests with RASTA processed MFCCs for the English, Italian, Danish, and Norwegian EUROM0 using pure HMM segmentation (V) and constrained HMM segmentation (VC) with acoustic segmentation at 150% o.s..*

The optimal RASTA-filtering improved the accuracy for all Italian speakers, especially for GCI and LCI. For instance, with pure HMM segmentation of GCI the number of gross errors was reduced from 73 to 20.

For Danish also, the average performance for pure HMM segmentation improved more than for constrained HMM segmentation when applying cepstral filtering. However, the constrained HMM segmentation obtained still the lowest number of gross errors and the highest coincidence rate within the ±5ms and ±10ms deviation thresholds. The cepstral filtering improved the segmentation performance for all four Danish speakers, especially for speaker JDD where e.g. the number of gross errors was halved.

For the Norwegian speakers the cepstral filtering improved most the segmentation of the two female speakers, SHN and TBN, see table D.30 and D.31. E.g. for speaker SHN, cepstral filtering reduced the gross errors to one third of that without filtering. In addition to significant reduction of the gross errors for SHN and TBN, the coincidence rates increased significantly for deviation thresholds larger than ±10ms. For the two Norwegian male speakers, the segmentation performances were either unchanged or slightly reduced (compare table D.19 with table D.31).

The effect of cepstral filtering was less for English than for the other languages. For speaker JWE and MBE the segmentation performances were improved, but for speaker EAE and JHE the RASTA filtering made no change to the segmentation performance. Surprisingly, for the Test Passage (speaker EAE in another recording environment) the cepstral filtering did not affect the segmentation performance.

Thus, with the cepstral filtering with optimal pole value the segmentation performance on the different speakers for each language become more similar. This was also confirmed by the

ANOVA test introduced in section 7.4.3, where the overall standard deviation for both gross errors and fine errors was halved[5]. Thus, the ANOVA test indicated with higher significance level ($p < 0.02$ for all tests) that the segmentation performances differed for the different languages (although the variance in segmentation performances for the different languages *was reduced*).

The tables D.34 to D.41 in Appendix D provide the coincidences within ±20ms deviation for the individual phoneme class transitions for the constrained HMM segmentation when using optimal pole in the RASTA-filter. A comparison of the English SI-test results in table D.34 and D.35 with the corresponding results obtained with the simple parameter set, table 7.9 and 7.10, shows that especially the transitions to and from vowels have been improved by applying optimal parameter set and cepstral domain filtering. Significant higher coincidence rates were also obtained for the transitions to nasals and from liquids. Thus, particularly the transitions between vowels, nasals, and liquids were segmented better than in table 7.9 and 7.10. In addition, the transitions from nasals and from fricatives were more accurately segmented. For the other phoneme class transitions only minor improvements were achieved, and for transitions to plosives slightly fewer boundaries coincided within the ±20ms deviation threshold.

For all languages the automatic segmentation of plosives, affricates, and fricatives obtained highest coincidence rates with manual segmentation. For Norwegian, the segmentation of the retroflex consonants and diphthongs also obtained high coincidence rates because in the Norwegian EUROM0 most of the retroflex consonants were plosives and most of the diphthongs were surrounded by plosives. Vowels and nasals were also accurately segmented in all languages, whereas the automatic segmentation of glides, liquids, and approximants obtained a lower coincidence rate with the manual segmentation. Particularly difficult transitions to segment proved to be the transitions to and from the Danish approximants and to the English liquids. Based on the discussion of manual segmentation in this thesis, those results were as expected.

The gross errors were equally distributed among the different phoneme-classes. Almost all gross errors occurred separately, i.e. the gross errors were seldom due to ripple errors. The displaced "gross errors segments" were often close to the manually marked segment, e.g. about 25% of the gross errors were less than 5ms displaced and about 40% of the gross errors were less than 10ms displaced from the corresponding "correct" segment.

[5] For instance, for gross errors σ was 0.32 with pure HMM segmentation and 0.25 with constrained HMM segmentation, whereas for coincidence rates within the ±20ms deviation threshold σ was 0.84 with pure HMM segmentation and 0.84 with constrained HMM segmentation.

In order to show the range of segmentation performances using RASTA processed MFCCs, the pole of the filter could be set to zero. In this case the RASTA-filtering on the MFCCs simply computed the delta-cepstrum. Segmentation results using delta-cepstrum are summarized in table D.42 and D.43, and in figure 7.9 these results are compared with ordinary MFCCs (table 7.21) and RASTA processed MFCCs where the pole value was optimal for each language (table 7.24).

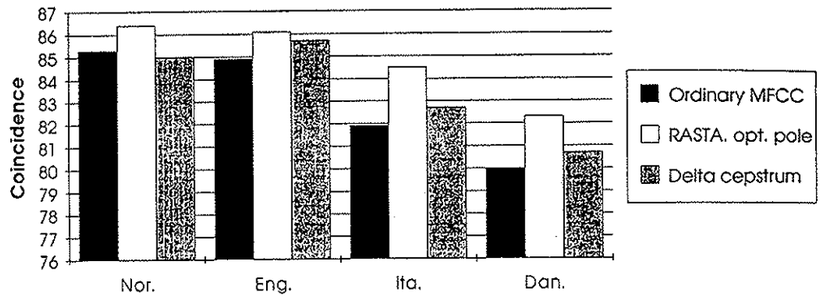


Figure 7.9 *Comparison of coincidence rates within the ±20ms deviation threshold for the SI,VC (150% o.s.) for the Norwegian, English, Italian, and Danish EUROM0 recordings when using three different parametrisations; ordinary MFCCs and RASTA-filtered MFCCs (with optimal pole value and without pole in the RASTA-filter).*

7.4.6 ENLARGING THE PHONEME TRAINING SET ACROSS LANGUAGES

So far, the phoneme HMMs for each language have been trained on very small and phonemically unbalanced training sets for the given language. One obvious method for making more adequate and reliable phoneme models is to expand the training sets. However, for some languages new recordings and manual annotations are then needed. To alleviate this labour demanding and expensive work, an intuitive strategy is to use the annotated speech material available across languages. For the EUROM0 recordings of the four languages used in this thesis, the training set for a SI-test would then increase from 3 speakers (or approximately 6 minutes of speech) to 15 speakers (or approximately 30 minutes of speech). More speakers and different recording conditions may provide more robust models. However, the multi-lingual training set will contain a greater variation in phoneme realisations than a language dependent set. *It will thus be a trade-off between the advantages of augmented training sets and the disadvantages of increased variability in phoneme models.*

Based on the EUROM0 recordings, annotations of different languages have been compared (see chapter 4 and Appendix C). To enlarge the training set for the Norwegian phoneme HMMs without making new recordings and annotations, we investigated the relationship between the phoneme symbols and the sounds in the Norwegian, Danish, Swedish, English, and Italian EUROM0 recordings. The phonemes were categorised as *poly-phonemes* and *mono-phonemes*, where *poly-phonemes* are phonemes which are realised roughly similarly in different languages, and *mono-phonemes* are phonemes which have to be kept separate for the specific languages [Dalsgaard'92].
The selection of phonetically similar phonemes was mainly auditorily based but was also based on comparisons of waveforms and spectrograms.

Similarity as defined above is a highly subjective term. For instance the poly-phonemes were defined similar enough to be equated for the given purpose. Also, experiments have shown that there is no single "correct" phonetic labelling of an utterance[6] [Cole'92], so that our grouping of phonemes may not be "worse" or "better" than any other labelling of speech.
However, the cross-language grouping of phonemes into poly-phonemes can, and should, be automatized by some decision-tree [Riley'92] or Vector Quantization technique.

The **consonant** symbols covered roughly the same acoustic realisations in the different languages, with some few exceptions, e.g the /r/ and /j/ in *RP English*, which we defined as mono-phonemes. Also the successive plosives without release of the first one in *RP English* should not be used to enlarge the Norwegian phoneme training material (see analyses in Appendix C).

Of the *Danish* consonants, /s/ is often shorter and with an unusual waveform (i.e. the frication is superimposed on a slowly varying wave, see e.g. figure C.1). The /n/ represents both [n] and some instances of [N]. The plosives have to be constructed, e.g. [p0]+[p] = /p/. Some cases of successive plosives without release, plosives initialising a sentence, and segments labelled with

[6] For instance, labelling experiments on the phonetically annotated TIMIT database have shown that different labellers on average agreed with TIMIT vowel labels in only 54.8% of the cases when the vowel segments where played with no context [Cole'92]. By including the adjacent segment before and after, the average agreement increased to 65.9%. Listener label agreement for stops and affricates in CV-syllables from TIMIT were 85-95% on average.

two symbols (e.g. /f*R/) were not included in the Norwegian training set.

The *Swedish* consonants were realised rather similarly to the Norwegian ones, but when plosives start a sentence they were segmented differently from our approach (see section 4.1.2). Note also that both /rs/ and /S/ are used for the Swedish [S]-sound to indicate its orthographic origin.

To conclude, the plosives in Norwegian, Swedish, English and Danish may be regarded as poly-phonemes. Since Italian have no aspirated plosives, the Italian /p,t,k/ may be equated with the Norwegian /b,d,g/. The /l/ and all the nasals (except the Danish /n/) are poly-phonemes.
For the **fricatives** we got the following table:

Nor	f	s	S	C	h	j	v
Swe	f	s	S rs	C	h	j	v
Ita	f	s ss	S		h	j	v
Eng	f	s	S		h		v
Dan	f	(s)			h		

Table 7.25 *Poly-phoneme fricative symbols that can be used for training the Norwegian statistical models.*

The **vowels** were rather roughly assigned to their Norwegian counterparts as shown in table 7.26. For instance the Danish and English /{/ can be realised quite differently. Also within one language the differences were noticeable, e.g. the Swedish speaker GW pronounced /no:gra fo:/, with the first /o:/ sounding like [u] and the second /o:/ sounding [O].

Nor	i	i:	e	e:	{	{:	u	u:	O	O:	A	A:	y	2	}	}:
Swe	I	i:	e	e:	E {	E: {:	U	u: o:	O		a	A:	Y		u0	
Ita	i		e E				u		O o		a					
Eng	I	i:	e		{		U		O	Q	A	A:		3		
Dan					E {		o u		O					2	u	

Table 7.26 *Poly-phoneme vowel symbols that can be used for training the Norwegian statistical models.*

In addition to the poly-phonemes suggested for vowels in table 7.26, the RP English /u/ may be equated with the Norwegian /}/, and /V/ (and /Q/) with /O/. The Danish /Q/ may be equated the Norwegian /O/ and /9/ can be equated with the Norwegian /2/.

As regards the r-sounds, the *Swedish* /r/ may not have the same allophones as the Norwegian /r/. /R/ in *Danish, German,* and *French* and /r/ in *English* are realised differently from /r/ in the Norwegian database which only exemplifies apical /r/ pronunciation while the most common /r/ realisation in *British English* is a (post-) alveolar approximant with glide-like features, i.e. quite different from the Norwegian /r/. (Norwegian dorsal /r/ dialect speakers have /r/ pronunciations similar to /r/ in German and French, see figure 4.8).
The *Italian* /r/ however, is realised as an apical tap and the geminate /rr/ as a trill. The

pronunciation is similar to Norwegian /r/ pronunciation intervocalically and when preceded by a fricative and voiceless plosive. But /r/ in sentence initial position differs from Norwegian in the same context. The Italian segmentation of /r/ is on the whole consistent with the Norwegian approach and the /r/ realisations for the two languages can thus be combined to enlarge the language specific training material for our statistically based automatic segmentation method. For training of the Norwegian retroflexed consonants the Swedish ones can be used.

Results for the Norwegian SI-test

Combining phonemically unbalanced training sets does not necessarily yield a balanced training set. For instance with the training set suggested above, the were 1582 /e/-segments and 1489 /s/-segments, but only 21 tokens of /rt/ and 24 tokens of /oy/. Thus the new training set became even more unbalanced than the Norwegian EUROM0 training set only.

In table 7.27 segmentation results for the Norwegian SI-test are compared with respect to different training sets. The results should be compared with the results for the Norwegian SI-test using "normal" training conditions (table 7.22 and D.17), or the results obtained by using cepstral filtering with pole ρ=0.88 (table 7.24 and D.31). In table 7.27 below training condition *Combined* means that the HMMs were trained on the four speakers in English, Danish, and Italian[7] in addition to the three Norwegian speakers (SI-test training), according to the poly-phoneme definitions above. This training set did not improve the segmentation performances. Compared to "normal" training conditions the performance of the pure HMM segmentation (V) was most deteriorated, e.g. the coincidence rates at the ±5 ms and ±10 ms deviation thresholds were significantly reduced. With the constrained HMM segmentation (VC) the new training set had almost no influence on the segmentation performance.
The coincidence rates for the transitions from vowels to nasals were reduced by 22% and for liquids the coincidence rates were reduced by 10%. Especially the transitions between liquids and vowels were segmented less accurate with the enlarged training set. However, at some transitions the segmentation accuracy increased for both algorithms, particulary at the vowels/vowels[8] and nasals/fricatives transitions.

Compared with the *Combined* training condition, RASTA-filtering with optimal pole value for each language prior to training, (*Combined* + *RASTA*), reduced the number of gross errors significantly and increased the coincidence rates within ±25ms deviation significantly for both algorithms. Especially the segmentation of the transitions from nasals to vowels were improved. Actually, with this training material the segmentation performance was slightly better with respect to gross errors and coincidence rates within ±20ms and ±25ms deviation than when training on Norwegian only. These results clearly show the speaker and environment normalisation effect of the cepstral domain filtering.
Compared with "normal" training conditions the (*Combined* + *RASTA*) reduced the number of gross errors (significantly for the constrained HMM segmentation) and increased the coincidence rates within ±20ms and ±25ms deviation slightly for both algorithms. However, the performances with (*Combined* + *RASTA*) were not as good as when training on Norwegian alone with RASTA

[7] Due to practical data problems, the Swedish EUROM0 recording was not employed in these tests.

[8] This was remarkable since the vowels were expected to vary most across the languages.

filtered MFCCs (table 7.24).

Enlarging the training set across languages introduces more variability. In order to model the variety in the data better it was reasonable to increase the number of mixtures from one to two in the phoneme HMMs, (*Combined + RASTA + 2 mix.*). The extra mixture reduced the number of gross errors for pure HMM segmentation and increased the coincidence rates within the ±20ms and ±25ms deviations slightly for both algorithms. This shows that with a larger training set the mixture parameters were better estimated than when training on one language only (cf. section 7.4.1). At small deviation thresholds poorer performance was obtained with 2 mixtures. For the rest of the experiments only one mixture was used.

Training conditions	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Combined, V	2.0 %	45.0 %	64.7 %	83.9 %	87.2 %
Combined, VC	1.2 %	46.7 %	65.4 %	84.5 %	87.8 %
Comb. + RASTA, V	1.2 %	44.0 %	67.4 %	86.3 %	89.6 %
Comb. + RASTA, VC	0.6 %	46.6 %	66.2 %	86.1 %	89.7 %
Comb. + RASTA + 2 mix, V	0.8 %	42.4 %	65.8 %	87.1 %	90.5 %
Comb. + RASTA + 2 mix, VC	0.6 %	46.1 %	65.6 %	86.1 %	89.9 %
Comb. + sex dependent + RASTA, V	1.7 %	40.5 %	63.8 %	83.8 %	87.1 %
Comb. + sex dependent + RASTA, VC	1.1 %	44.6 %	63.8 %	83.6 %	87.3 %
Eng. + Nor. + RASTA, V	1.1 %	49.7 %	70.5 %	86.7 %	89.9 %
Eng. + Nor. + RASTA, VC	0.8 %	49.1 %	68.3 %	86.4 %	89.8 %
Ita. + Nor. + RASTA, V	1.0 %	42.6 %	66.1 %	86.6 %	90.2 %
Ita. + Nor. + RASTA, VC	0.6 %	46.3 %	65.9 %	85.4 %	89.2 %
Dan. + Nor. + RASTA, V	1.0 %	42.8 %	66.5 %	86.4 %	90.0 %
Dan. + Nor. + RASTA, VC	0.4 %	46.4 %	66.0 %	85.3 %	89.1 %

Table 7.27 *Coincidence rates and percentage gross errors for the Norwegian SI-test with phoneme HMMs trained on the English, Italian, Danish, and Norwegian EUROM0 using pure HMM segmentation (V) and constrained HMM segmentation (VC) with acoustic segmentation at 150% o.s..*
Detailed results are provided in table D.44 and D.45.

HMM based ASR systems have achieved improved recognition accuracy by training and testing on speakers with similar sex only, e.g. [Huang'93]. However, using sex dependent HMMs in the segmentation experiment (*Combined + sex dependent + RASTA*), reduced the segmentation performance compared with (*Combined + RASTA*). The reductions were significant for

coincidence rates within ±10ms, ±20ms, and ±25ms deviation thresholds for pure HMM segmentation, and for coincidence rates within ±25ms deviation threshold for constrained HMM segmentation.

Since enlarging the training set across all languages did not improve the segmentation performance, we investigated the addition of one language at the time to find whether the languages had a different effect on the segmentation performance of Norwegian.
For pure HMM segmentation a training set based on Norwegian and English (*Eng.+ Nor.+ RASTA*), yielded significantly higher coincidence at ±5ms and ±10ms deviation thresholds than the training sets based on Norwegian and Italian (*Ita.+ Nor.+ RASTA*) or Norwegian and Danish (*Dan.+ Nor.+ RASTA*). This may indicate that the manual segmentation strategies applied for Norwegian and English were quite similar. Again, the constrained HMM segmentation was less sensitive to the different training conditions.
For these three enlarged training sets, the transitions vowels/vowels, liquids/nasals, liquids/fricatives, and nasals/fricatives got particularly reduced segmentation accuracy.
For the training sets augmented by Italian (*Ita.+ Nor.+ RASTA*) or Danish (*Dan.+ Nor.+ RASTA*), the transitions from fricatives and liquids to plosives, fricatives, nasals, and liquids got most decreased performance. For the individual speakers, the segmentation performance of speaker AFN was least sensitive to different training sets.

The results with enlarged training set across languages show that the augmented training set did not weigh up for the increased variability in phoneme realisations due to differences in languages, manual annotation strategies, and recording conditions.
On the other hand, the performances were not much worse than when training on Norwegian only, (except from the deterioration in coincidence rate at ±5ms deviation). This shows that the new multi-lingually trained models have preserved most of their discriminative cues. In addition, it is reasonable to expect that the phoneme models based on an enlarged training set are more robust, in the sense that when employed for segmenting other recordings of Norwegian they may achieve better results than the models trained on the Norwegian EUROM0 recording only.

For some languages some speech recordings are made but not annotated. Instead of laborious manual annotation, poly-phonemes from other languages may be used for an initial segmentation[9]. Here this possibility is investigated for Norwegian.
In table 7.28 below, *others* denotes that the phoneme HMMs were trained on the other languages, i.e. English, Danish, and Italian EUROM0 recordings, for the poly-phonemes defined above, and that the SI-test training is used for the Norwegian mono-phonemes; /A A: ae ae: O: e: oe u: y ou ou: oy rt rd/.
Others + RASTA denotes the corresponding test with RASTA processed MFCCs, where optimal pole value is used in the RASTA-filter for each language.

The segmentation performances were significantly worse than when training on Norwegian only. However, with RASTA-processed MFCCs and constrained HMM segmentation the performances were comparable with those obtained when training on Norwegian only, e.g. the 95% confidence intervals for the gross errors overlap (compare table D.17 with table D.46).

[9] As a minimum requirement models for the mono-phonemes in the given language have to be provided and a label string has to be given as input to the automatic segmentation algorithm.

We also employed the phoneme HMMs estimated above as an initial estimate for the Norwegian phoneme HMMs and then ran the Baum-Welsh reestimation on the Norwegian EUROM0 recording only. These experiments are denoted (*others, init.*) and (*others + RASTA, init*), respectively in table 7.28 below.
Compared with training on Norwegian only, the new initialisations of phoneme HMMs did not influence the segmentation performance significantly.

Training conditions	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Others, V	4.1 %	35.1 %	54.2 %	75.8 %	80.3 %
Others, VC	2.2 %	41.5 %	58.8 %	78.7 %	82.6 %
Others + RASTA, V	3.1 %	36.5 %	58.9 %	80.0 %	84.1 %
Others + RASTA, VC	1.8 %	42.6 %	61.1 %	81.0 %	85.4 %
Others, init., V	1.8 %	48.4 %	67.7 %	84.8 %	87.6 %
Others, init., VC	1.2 %	48.4 %	66.7 %	85.2 %	88.3 %
Others + RASTA, init., V	1.0 %	48.8 %	70.3 %	88.5 %	91.3 %
Others + RASTA, init., VC	0.4 %	48.8 %	68.2 %	86.7 %	90.4 %

Table 7.28 *Coincidence rates and percentage gross errors for the Norwegian SI-test with HMMs trained on the English, Italian, and Danish EUROM0 using pure HMM segmentation (V) and constrained HMM segmentation (VC) with acoustic segmentation at 150% o.s..*
The corresponding 95% confidence intervals are provided in table D.46.

7.4.7 ALTERNATIVE PHONEME-STRINGS

To further reduce the human activity in annotating speech databases, the input label string to the automatic segmentation algorithm can be produced by a text-to-phoneme (TTP) converter. As discussed in section 4.2.5, such a phonotypical label string will not represent the actual speech very accurately, i.e. insertions, deletions, and alternative pronunciations occur.

To get a first indication of how well the phonemic segmentation algorithm will handle TTP-generated phoneme-strings, alternative paths in the label network provided by manual annotation were allowed for. For instance, in the Test Passage the symbol /z/ was used for both voiced [z] and voiceless [s] realisations of /z/ (see Appendix C). With alternative pronunciations allowed for, both the transcription [zirU] and [sirU] were accepted for the word "zero".

In the input label file for English we replaced every /z/ by /s/ and /z/ in parallel, and every /n/, /m/ and /N/ was replaced by /n m N/ in parallel. Similar alternatives were made for every label in /i i: I I@/, /e {/, /e@ 3: @/, /u u: U U@ @U/, /V O:/, /A: Q/, and /eI aI OI/.
The label file with alternative phoneme-strings was then used as input network to the phonemic segmentation and the Viterbi decoding selected the model for each alternative path with the highest probability.

Although this task was more difficult than the ordinary label alignment, in that many alternatives were to be chosen among (i.e. had to be recognised), the phonemic segmentation accuracy was not significantly altered.
Thus, alternative pronunciations can, and should, be allowed for when TTP-generated label-strings are used as input to the phonemic segmentation algorithm.

7.5 COMPARISON WITH PHONEME RECOGNITION

Automatic segmentation, or label alignment, can be regarded as a kind of phoneme recognition where the ASR system is forced to recognise a given label string. Thus, in the development of our automatic phonemic segmentation software, most of the parameter choices were based on reported experiments with ASR. In this section the impact the *different parameter sets*, the *constrained Viterbi recursion*, and the *cepstral filtering* had to ASR, are compared with the effects on phonemic segmentation. The same experimental setup was used for the ASR as for the ASS experiments and the HMMs used were the same context independent phoneme models as used for phonemic segmentation.

The recognised phoneme string was evaluated by an optimal string match, i.e. by *dynamic programming*, between the recognised and the "correct" phoneme string. Denoting the number of phonemes N, the number of deletions D, substitutions S, insertions I, and H=N-D-S, we defined (according to [Young'92]):

$$\%\ Correct = {}^{H}/_{N} * 100 \quad \text{and} \quad \%\ Accuracy = {}^{H-I}/_{N} * 100.$$

In the tests below we assumed that one algorithm obtained *significantly* different accuracy rate than another on the same test set if the 95% confidence intervals for the accuracies did not overlap. In order to estimate the confidence intervals the correctness and accuracies were modeled with *binomial distributions*. That is, we assumed that there were N independent tests where the recognised phoneme was either correct or wrong, where the success of one phoneme recognition was independent of all the other phonemes, and where the probability for a correct recognition was equal for all tests.

7.5.1 PARAMETER SET

Optimising the parameter set with respect to the English EUROM0 recording did not always lead to the same conclusions for recognition as for segmentation. For instance, *embedded reestimation* and *increasing the number of mixtures* improved the accuracy and correctness rate of recognition, but reduced the segmentation accuracy. Also the *grouping of phonemes* into classes for the initial training and then *splitting the class-models* for further training on individual phoneme models, improved the ASR performance but reduced the segmentation accuracy. Adding *more states and transitions* in the HMMs decreased the recognition accuracy, whereas the segmentation performance was insensitive to these changes.
On the other hand, liftered cepstrum and successively increasing the parameter vector by delta-cepstrum, energy, and delta-energy, improved the recognition as well as the segmentation performance. Also, applying MFCC achieved better recognition accuracy than when applying LPCCEP, as it did for segmentation. Table 7.29 below shows the ASR performances for the SI-test using pure HMM recognition on the English EUROM0 and the Test Passage with the same parameter set as used in table 7.20. For the English EUROM0 the correctness and accuracy rates obtained with MFCC were significantly higher than those obtained with LPCCEP. For the Test Passage only the accuracy rate was significantly higher with the MFCC.

In the following the optimal parameter set for segmentation (section 7.4.1) was used for the ASR experiments.

SI,V - test, ASR	% Correct	% Accuracy
MFCC, English EUROM0	57.83%	24.97%
LPCCEP, English EUROM0	52.45%	12.14%
MFCC, TP	48.52%	13.50%
LPCCEP, TP	41.35%	1.27%

Table 7.29 *ASR performance for SI,V -test on English EUROM0 and Test Passage, using the same setup as for the SI,V tests in ASS.*

7.5.2 CONSTRAINED HMM RECOGNITION

For most of the segmentation tests, constraining the Viterbi search with acoustic segment boundaries improved the performance by decreasing the number of gross errors and increasing the coincidence rates within small deviation thresholds[10]. In this section it is shown that *constraining the Viterbi search with acoustic segment boundaries also improved the ASR accuracy of the corresponding HMM phoneme recogniser considerably.*

Constraining the possible time-instances for transitions between phoneme models in the Viterbi decoding by acoustic segment-boundaries reduces the search space. Thus, intuitively we expected that the number of insertions would decrease and the number of deletions increase.
The results in table 7.30 show that the major effect of constraining the Viterbi recursion with acoustic boundaries with 150% o.s., was a considerable reduction of insertions, whereas the number of deletions increased comparably less, and the number of substitutions was relatively unchanged. Therefore, *the recognition accuracy increased significantly by constraining the Viterbi decoding with acoustic segment boundaries, while the correctness rate did not decrease significantly.*

The ASR results differed from the segmentation results in these respects:

i) For recognition, constraining the Viterbi forward recursion with acoustic segment boundaries improved the accuracy for all languages and all speakers. For segmentation, constraining the Viterbi forward recursion with acoustic boundaries improved the accuracy for only some speakers.

ii) The segmentation accuracy for Norwegian was similar with pure HMM and constrained HMM, but the pure HMM *recogniser* achieved very low accuracy for Norwegian.

iii) Constraining Viterbi recursion had similar positive effect for all four languages.

[10] However, with the best parameter set pure HMM segmentation achieved best performance for some speakers.

SI-test, ASR	Del.	Sub.	Ins.	H=N-D-S	%Correct	%Accuracy
English, V	206	1677	1467	2582	57.83	24.97
Norwegian, V	173	1839	1928	2488	55.29	12.44
Italian, V	375	1986	1515	2735	53.67	23.94
Danish, V	338	1868	1218	2376	51.86	25.27
English, VC	299	1650	966	2516	56.35	34.71
Norwegian, VC	232	1869	1047	2399	53.31	30.04
Italian, VC	579	1894	894	2623	51.47	33.93
Danish, VC	454	1818	793	2310	50.41	33.11

Table 7.30 *Recognition results for the SI-tests for the English, Norwegian, Italian, and Danish EUROM0, with pure HMM recognition (V) and constrained HMM recognition (VC) with acoustic segmentation at 150% o.s. (No grammar).*

In order to investigate the differences between the individual speakers, the recognition performance as a function of oversegmentation-factor, and the effect of grammar, the ASR experiments on the **Norwegian** EUROM0 are discussed in more detail. A *grammar* assigns a probability to the possible phoneme successors for each phoneme. Without a grammar, any phoneme sequence is allowed to occur. The grammar used here was the *diphone statistics* of the whole EUROM0 speech material for the given language. Since many diphones never occurred, the grammar restricted the branching factor in the phoneme network considerably. Table D.47 shows that the grammar reduced the number of possible phoneme sequences so the number of insertions decreased which in turn increased the accuracy rate significantly.

Constraining the Viterbi forward recursion with the acoustic segmentation improved the ASR accuracy significantly, as shown in table D.47. For instance for the SI-test the introduction of constrained HMM recognition increased the accuracy by 41% compared with pure HMM recognition.
The SI-tests with pure HMM and grammar had the same rate of accuracy as the SI-test with constrained HMM and no grammar. The effects of the grammar and the constrained HMM were *additive*, so the accuracy increased by 250% for the SI-test from pure HMM recognition to using constrained HMM recognition and grammar.
By grouping short and long vowels the ASR accuracy increased from 12.4% to 18.8% for the SI-test with pure HMM recognition without grammar (in section 7.4.4 this grouping of vowels decreased the segmentation performance).

The SI-test recognition performances for the individual Norwegian EUROM0 speakers are shown in table D.48. For pure HMM recognition, the readers with slowest and the fastest articulation rate, AFN and TGN, got the lowest accuracy. Especially for AFN there were many insertions, maybe due to the extremely low articulation rate. For the other experiments, the accuracy was more similar for the different speakers.

Constraining the Viterbi recursion by acoustic boundaries almost halved the number of insertions for AFN while keeping the other factors almost the same, and thereby increasing the accuracy

from 7.5% to 31.7%. An additional constraint by the grammar further reduced the number of insertions for AFN with 46.5%, giving the best ASR accuracy of the Norwegian speakers; 48.34%.
(For segmentation, there were individual differences where constrained HMM led to a reduction in accuracy for AFN, made no difference for SHN, and improved the performance of TBN).

For comparison with the upper limit for the chosen models and speech material, the full HMM test with constrained HMM and no grammar, obtained 66.52% accuracy for AFN, 63.44% for SHN, 57.01% for TBN, and 61.59% accuracy for TGN.

Constrained HMM phoneme recognition as a function of oversegmentation factor

The performances of the constrained HMM segmentation decreased with decreasing oversegmentation factor, but was relatively constant for o.s.-factors between 75%-175%.
For ASR, the recognition accuracy *increased* when decreasing the o.s.-factor, as shown in figure 7.10 (see table D.49 for details). The recognition accuracy increased because the number of insertions was reduced more than the increase in deletions, and the number of substitutions was relatively constant or decreased. Thus, the correctness rate decreased relatively less than the accuracy rate increased.

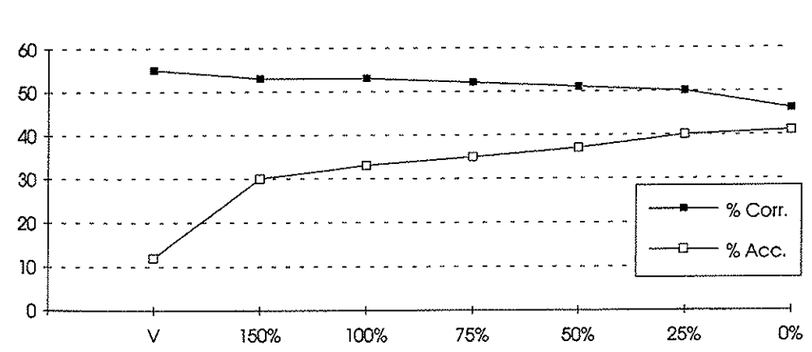


Figure 7.10 *Comparison of phoneme recognition performance for the Norwegian SI-test with the pure HMM recognition (V) and the constrained HMM recognition with constrained Viterbi for different oversegmentation factors.*

Even with only 25% oversegmentation, more than 70% of the acoustic boundaries DTW-coincided within ±20ms deviation from manually placed segment boundaries (see table D.6). Constraining the Viterbi recursion with these boundaries improved the accuracy from 12.4% (for pure HMM recognition) to 40.0%, whereas the corresponding correctness rate decreased from 58.3% to 49.5%. This tendency was the same for all speakers, except for AFN, where both the correctness and accuracy rate was higher with 0% o.s. than with 25% o.s..

However, in the experiments with constrained HMM segmentation the number of acoustic segments for an utterance were determined by the number of phonemes in the utterance. In a practical ASR system, the number of phonemes is not known a priori. Thus, the acoustic segmentation should terminate when e.g. the average distortion for the sentence becomes less than a predefined threshold (see e.g. [Kvale'87] for experiments with different stop criteria).

7.5.3 CEPSTRAL DOMAIN FILTERING

As for phonemic segmentation, RASTA-filtering of the MFCCs improved the ASR performance of all four EUROM0 recordings. The ASR accuracy rate was improved more with the RASTA-filtering than the correctness rate. In figure 7.11 the ASR accuracies for English, Norwegian, Italian, and Danish SI-tests with constrained HMM recognition, are shown for optimal values of the pole in the RASTA-filter.

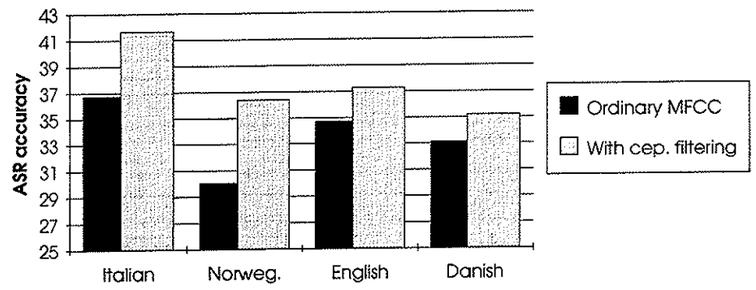


Figure 7.11 *Improvement in accuracy rate with RASTA-processed MFCC with optimal pole placement for English, Norwegian, Italian, and Danish SI-test with constrained HMM recognition with acoustic segment boundaries computed with 150% o.s.*

The speakers that got most improvement with optimal RASTA-filtering for segmentation did not necessarily get most improvement with respect to ASR accuracy. For instance in Italian, the ASR performance was improved least for speaker GCI and most for speaker LCI, and for Norwegian most ASR improvement was noted for speaker TGN.

The ASR-system may recognise the "uncertain areas" between two phonemes (see chapter 4) as a phoneme. With RASTA-filtering the frame vectors were smoothed and the number of insertions was reduced. For segmentation the coincidence rates were less sensitive for such filtering because the ASS had to align the given labels (i.e. insertions were impossible).

The ASR performances were much more sensitive to variations of the pole value in the RASTA-filter, than the segmentation results. For segmentation, RASTA-filtering with almost all pole values, for some speakers even ρ=0, improved the performance, with an optimum pole value around 0.88 for most speakers. For ASR, the optimum pole value was close to 1, e.g. ρ=0.99. Decreasing the pole value decreased the ASR performance considerably, to a level below that obtained without RASTA-filtering, as shown in figure 7.12 for Norwegian.

Figure 7.12 shows the relative performance of the constrained HMM recogniser and the constrained HMM segmenter as a function of the pole in the RASTA-filter for the Norwegian SI-test. For **segmentation** the coincidence rates within ±20ms deviation are plotted relative to 85.3% (table 7.21) which was obtained with the optimal MFCC-based parameter set defined in section 7.4.1. That is, +2% for segmentation in figure 7.12 means 87.3% coincidence rate within ±20ms deviation.
For **recognition** the accuracies are plotted relative to 30% (table 7.30), in that +2% in figure 7.12 means 32% ASR-accuracy.

Figure 7.12 shows that the *recognition accuracy* increased with increasing pole value, with maximum at ρ=0.98-0.99, whereas best *segmentation accuracy* was obtained with ρ=0.88 for Norwegian.
In the other end of the scale, with ρ=0, the segmentation results were similar with that obtained with ordinary MFCCs, whereas ASR obtained only 15.2% accuracy.

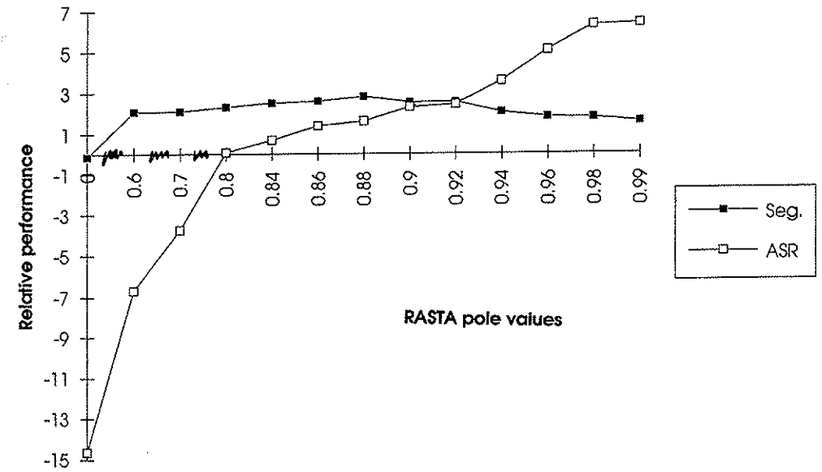


Figure 7.12 *Segmentation fine errors within ±20ms deviation for Norwegian SI,VC-test and recognition accuracy for the Norwegian SI,VC-test (with 150% o.s), as a function of the pole placement in the RASTA-filter.*
The results are given relative to the performance with the ordinary MFCC parameters. (See text).

7.5.4 ENLARGING TRAINING SET ACROSS LANGUAGES

The pure HMM phoneme recognition accuracy for the Norwegian SI-test dropped from 25.09% when trained on Norwegian only, to 16.53% when employing models trained on Danish, Italian, and English EUROM0 in addition to the Norwegian recordings (for all languages cepstral filtering, i.e. RASTA-filter with ρ=0.98 was applied). The SI-test for the constrained HMM phoneme recognition showed similar drops in accuracy rates.
Compared with the segmentation performances with different training sets, the only difference was that for ASR an increased number of mixtures in the phoneme models increased the ASR accuracy.

As for segmentation, the best results with enlarged training set were obtained when augmenting with English only, in that training on Norwegian and English only, gave 22.56% accuracy, training on Norwegian and Italian only, gave 16.42% accuracy, and training on Norwegian and Danish only, gave 14.54% accuracy for the pure HMM phoneme recognition.

Comparing confusion matrices shows that using models trained on English and Norwegian led to less confusions for /t/ and /m/, but more confusions especially for /b/, /v/, and /A/.

In automatic segmentation the task is to *discriminate* between the given labels, whereas in automatic recognition the *identity* of the segments has to be discovered. The increased variability in the phoneme models implied more overlap between models, which reduced the phoneme recognition accuracy even more than the segmentation accuracy.

7.6 SUMMARY

In this chapter several experiments with our automatic segmentation algorithm have been evaluated both quantitatively and qualitatively. The quantitative assessments of the automatic segmentation have been performed as comparisons with the corresponding manual segmentation and the results have been provided in terms of fine errors and gross errors. Since the endpoints of the sentences were given to the automatic segmentation algorithm, these boundaries were excluded in the counts of coincidence between automatic and manual segment boundaries.

The acoustic segmentation was evaluated by a DTW-comparison with the manually annotated reference, i.e. each phonemic boundary was compared with its closest acoustic segment boundary only. The acoustic segmentation algorithm was robust against slightly different recording conditions and yielded approximately the same segmentation accuracy for the English, Norwegian, Danish, and Italian EUROM0 recordings. With more than 75% oversegmentation the acoustic segmentation obtained high DTW-coincidence rates with manually placed boundaries, and most of the acoustic segments could be given a phonetic interpretation.
Based on these analyses, we proposed the acoustic segmentation used as a *pre-segmenter tool* for multi-lingual manual segmentation. Such a tool will reduce the degree of randomness in the manual segmentation.

In almost all experiments on phonemic segmentation, the pure HMM segmentation performances have been compared with constrained HMM segmentation performances. When the acoustic segment boundaries were used to constrain the Viterbi forward recursion, the number of gross errors decreased, the coincidence rates at small deviations increased for most speakers, and the largest fine error for each speaker was reduced compared with pure HMM segmentation. Constraining the Viterbi recursion increased the segmentation performance when poorly trained phoneme models were used and also increased the segmentation performance for the language with lowest coincidence rates for pure HMM segmentation. That is, constraining the Viterbi recursion made the phonemic segmentation more robust against different languages and recording conditions.
In addition the accuracy of the HMM based phoneme recogniser was improved when constrained by acoustic segment boundaries.

The EUROM0 recordings were not phonemically balanced and consisted of only four speakers for each language. Thus, the training sets were too small to make reliable estimates of the phoneme HMMs. We wanted to increase the training material by utilising the existing annotated EUROM0 speech database. In one experiment we grouped the phonemes in a language into phoneme classes and trained the HMMs on these. We also enlarged the training set for a language by using poly-phonemes from the other languages. However, these enlargements of training sets also increased the variability in the phoneme realisations, resulting in a deteriorated segmentation performance.

Cepstral filtering improved the segmentation performances both with respect to gross errors and fine errors for all languages. Thus *table 7.24 provides the best SI-test results with our algorithm on the selected speech corpora.*
Table 7.31 summarises the performance in terms of fine errors for some of the automatic segmentation methods described in this thesis. However, the comparison in table 7.31 is only *tentative* in that the experimental setups differ with respect to e.g. the speech material, (e.g.

recording equipment and environment, phoneme content, speakers, speaking style, and language), annotation strategy applied on the reference material, which subword units are used, and training and test conditions (e.g. SI-test, TI-test, or Full HMM test, and inclusion of the given pause boundaries in the coincidence scores). In addition, without confidence intervals for the coincidence rates it is difficult to state whether the differences in performance are statistically significant or not.
However, for the tests on the EUROM0 recordings, our algorithm obtained significantly better performances than those reported elsewhere, e.g. [Dalsgaard'90], [Dalsgaard'91], [Barry'91b], and [Svendsen'90].

Deviation threshold			Database	References
± 10 ms	± 15 ms	± 20 ms		
		57 % 78 %	Danish EUROM0 English EUROM0	Dalsgaard & Andersen, 1990
48.7 % 62.6 % 34.2 %	58.9 % 72.5 % 45.8 %	65.5 % 77.5 % 52.0 %	Danish EUROM0 English EUROM0 Italian EUROM0	Dalsgaard, Andersen & Barry, 1991
70 %		90 %	TIMIT	Glass & Zue, 1988
		86 %	?	Hemert, 1987
75 %	90 %		TIMIT	Cole, 1988
75 %		90 %	TIMIT	Leung & Zue, 1984
	80 %		TIMIT	Ljolje & Riley, 1991
68.9 % 55.2 %		82.3 % 71.7 %	English EUROM0 Italian EUROM0	Svendsen & Kvale, 1990
70.3 % 66.2 % 64.2 % 68.1 %	81.0 % 76.4 % 77.7 % 80.4 %	86.1 % 82.3 % 84.5 % 86.4 %	English EUROM0 Danish EUROM0 Italian EUROM0 Norwegian EUROM0	This thesis, table 7.24

Table 7.31 *Accuracy of some automatic phonemic segmentation methods. For our method in this thesis the results for the speaker-independent test with constrained HMM segmentation with 150% o.s. are given (see table 7.24).*
The results in [Svendsen'90] include the known pause transitions.

Ideally, the performance of the automatic segmentation algorithm should be similar to the coincidence rates between two manual segmentation and labellings. Unfortunately very few evaluations of manual annotation exist. For the speech material used in this thesis only the manual segmentation of speaker AFN was evaluated, see section 4.2.4. The fine errors and gross errors for the resemblance of two manual segmentations are in table 7.32 compared with the DTW-coincidences for the acoustic segmentation with 150% oversegmentation for speaker AFN (cf. table 7.2), and the best performance for speaker AFN with our phonemic segmentation (cf. table D.31). Since only 748 (out of 1144) boundaries were checked in the manual segmentation, the confidence intervals for the manual and the automatic segmentation performances have different widths. The missed pauses in the second manual segmentation are here regarded as gross errors.

Table 7.32 shows that the number of gross errors was low with our automatic segmentation algorithm (cf. also table 7.19). Other investigations, e.g. [Leung'84], [Cosi'91], [Ljolje'93],

indicate that for different human labellers the coincidence rates within the ±20ms deviation threshold may be lower than 95%, e.g. for Italian 93.5% coincidence was reported [Cosi'91] and for American English 90% coincidence was reported [Leung'84]. In this chapter qualitative assessment of the automatic segmentation showed that some of the "errors" made by the automatic segmentation were due to special conventions defined in the manual segmentation. In these cases the automatic boundary placements could be argued for phonetically to be as "correct" as the manual segmentation.
Thus, with respect to coincidence with a manual segmentation the automatic segmentation performances summarized in table 7.24 are probably close to an upper limit.

Speaker AFN	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Manual segmentation	0.3 - 1.3 %	59.5 - 66.3 %	85.4 - 90.1 %	94.8 - 97.5 %	96.4 - 98.6 %
Acoustic segmentation	-	55.7 - 61.4 %	79.0 - 83.5 %	94.0 - 96.4 %	96.1 - 98.0 %
Phonemic segmentation	0.3 - 1.3 %	48.1 - 54.2 %	70.5 - 75.9 %	87.2 - 91.0 %	89.4 - 92.8 %

Table 7.32 *95% confidence intervals for the coincidence rates and the number of gross errors for two manual segmentations and the Norwegian SI-test for speaker AFN.*

The constrained HMM segmentation obtained very similar results for the four languages. With cepstral filtering the difference in segmentation performance on different speakers within a language and between languages was even further reduced. When phoneme HMMs trained on some languages were used when segmenting another language, the segmentation performances were still comparable with those obtained when training on Norwegian only. Hence, poly-phonemes can be used to train, or at least initialise, phoneme HMMs for languages which have no annotated speech material available.
These results indicate that the phonemic segmentation algorithm combined with cepstral domain filtering is rather language and recording independent.

Although similar techniques are often applied in automatic speech segmentation and recognition, they have two different tasks. For automatic segmentation it is enough to discriminate between phonemes, whereas in automatic recognition the phonemes have to be identified. Thus, experiences with ASR cannot be transferred directly to ASS and visa versa.
However, the new idea of utilising the segmental information provided by the acoustic segmentation to constrain the possible state transitions in an *HMM based phoneme recogniser*, has been shown to improve the recognition accuracy significantly.

Chapter 8

CONCLUSION

The aim of this thesis has been to contribute towards the standardisation of segmentation and labelling of multi-lingual speech databases. The main contribution is a proposal for a language-independent algorithm for automatic segmentation and labelling of speech.

The performance of automatic segmentation algorithms should in general be evaluated in terms of how they may improve the intended use of the annotated speech material. For instance, letting the automatic segmentation system generate the subword library for a text-to-speech system and comparing the final synthetic speech with that generated by a hand-segmented subword library. However, our aim was to annotate a general purpose multi-lingual speech database. The assessment of the automatic segmentation algorithm was hence performed in terms of coincidence with a manually segmented and labelled reference speech material. In this case the automatic segmentation performance should ideally be at least as good as the agreement between expert labellers.
No single segmentation strategy can be claimed to be "correct". Hence, such a test of automatic segmentation performance is not an evaluation in the sense of what is correct, but rather a confrontation between the labeller's strategy and the automatic segmentation criteria. It is thus essential to investigate both the manual and automatic segmentation.

8.1 MANUAL SEGMENTATION

It is difficult to lay down detailed guidelines for the annotation of speech, due to the great variability and intrinsic complexity of the speech signal. A study of the actual annotations of speech recordings of some languages and their corresponding (rather sparse) documentations, showed that no agreed standards existed for segmentation and labelling of speech. However, in this thesis we made some manual annotation conventions explicit and compared them across languages, and we argued that such a comparison may serve as a starting point for the development of a standard procedure for manual multi-lingual speech annotation.

We proposed a set of manual segmentation conventions and applied them consistently for Norwegian. Our approach was to annotate what we heard and saw (waveform and broadband spectrogram) with phonemic labels and to mark the endpoints for each phoneme. If we heard a sound in context we segmented and labelled this sound even if we could not see any acoustic cues in the waveform or the spectrogram or could not hear the sound in isolation. The audible, but invisible phoneme was then squeezed in between the two surrounding phonemes by taking a couple of pitch periods from each of them. For some phoneme-to-phoneme transitions where no clear visual acoustic cues were available for marking boundaries, we introduced an iterative listening procedure.

Most of the detailed segmentation conventions suggested can probably be used for phonemic annotation of other languages as well.

Manual segmentation contains random human errors and inconsistencies. A comparison of two annotations by the same labeller with three months' interval showed that 96.5% of the boundaries coincided within ±20ms deviation, and 0.5% gross errors were detected. This coincidence was better than reported for similar comparisons elsewhere. The uncertainty in the manual segmentation varied for the different phoneme transitions for the same speaker and between speakers. Such an investigation should always be taken into account when assessing the accuracy of automatic segmentation algorithms in terms of coincidences with manual segmentation.

Manual annotation of speech is time consuming. When the segmentation conventions were known we annotated about 6 phonemes per minute. In comparison, an auditive transcription marked off about 10 phonemes per minute.

8.2 AUTOMATIC SEGMENTATION

We developed an automatic segmentation algorithm consisting of two parts; an acoustic and a phonemic segmentation module. In addition to the sampled speech signal a transcription of the speech and the endpoints of each utterance were required as inputs to the algorithm. In order to test performance of the automatic segmentation with respect to the manual annotation, we simply used the label string from the manual segmentation as input. If only an orthographic text is available, we argue that in a fully automatic annotation system a transcription generated by a Text-To-Phoneme (TTP) program can be used as input to the automatic segmentation algorithm without causing much deterioration.

The first pass of the automatic segmentation algorithm was language independent and segmented any speech signal into acoustic segments without any prior training. The number of acoustic segments was selected to be larger than the number of phonemes. We found that at least 75% oversegmentation was needed to achieve satisfactory coincidence with the manual segmentation. Some of the acoustic boundaries fall outside the ±20ms deviation because of some special conventions or inconsistencies in the manual segmentation. When the number of acoustic segments was twice the number of phonemes in a sentence, most of the acoustic segments could be given a phonetic interpretation.

In order to reduce the randomness in manual segmentation the acoustic segmentation may be used as a pre-segmenter tool. Such a tool should be combined with explicit rules for selecting acoustic boundaries for the different sound transitions and segmentation conventions such as the ones suggested for Norwegian in this thesis. Since the acoustic segmentation is language independent and does not require a training session, it may be used as basis for standardised multi-lingual segmentation.

The second pass of the automatic segmentation algorithm was language dependent in that the hidden Markov phoneme models had to be trained for each new language. By constraining the forward recursion step of the Viterbi algorithm with the acoustic segment boundaries this algorithm reduced the number of gross errors and became more robust against different language and recording conditions than the pure Viterbi decoding algorithm.

Cepstral domain filtering improved the segmentation performance for all languages and reduced both the interspeaker performance variation within a language and the variation in average performance between languages. In contrast to automatic speech recognition the automatic segmentation performance was rather insensitive to the pole value in the RASTA-filter.

Even when phoneme HMMs trained on different languages were used when segmenting Norwegian, the coincidence rates decreased less than 10%. These results indicate that the poly-phonemes defined in this thesis for Danish, English, Italian, and Norwegian really convey some phonetic similarities and that our phonemic segmentation algorithm is robust against different training conditions.

Since the exact placement of segment boundaries is subject for discussion, fine errors are not as critical as the gross errors. However, automatic segmentation algorithms are seldom assessed in terms of gross errors, although the number of gross errors indicates whether the manual and automatic segmentation correspond or not. For many purposes a low number of gross errors may be more valuable than a high coincidence rate for the fine errors within e.g. the ±20ms deviation threshold.
The best coincidences between the automatically placed segment boundaries and manual segmentation were close to that of comparing two manual segmentations. Especially the constrained HMM segmentation resulted in very few gross-errors, i.e. about 98-99% of the segments provided by this algorithm overlapped to some extent the manually marked reference segments. With respect to fine errors the constrained HMM segmentation obtained a coincidence rate of 64.2%-70.3% within ±10ms deviation and 82.3%-86.4% within ±20ms deviation for the different languages investigated in this thesis. Thus our algorithm obtained significantly better results than other algorithms on the same speech material and comparable results with methods trained and tested on larger speech corpora.

Some of the lack of correspondence between manual and automatic segmentation was due to special conventions or inconsistencies in the manual segmentation. Thus, some manual segmentation conventions could be slightly adjusted to better correspond with the automatic segmentation and still be argued for to be "correct". We believe that if the automatic segmentation had been evaluated manually, more boundaries would have been accepted as "correct" ones.

The experiments on speech recognition were performed with a very simple HMM phoneme recogniser trained on a sparse and phonemically unbalanced speech material. However, since the effect of constraining the Viterbi recursion was independent of imposing a grammar, we expect that constraining the Viterbi recursion will improve a more complex recognition system as well.

8.3 FURTHER WORK

The speaker independence test used for phonemic segmentation in this thesis was difficult because the phoneme models were trained on only three speakers and the segmentation was performed on the fourth speaker. That is, phoneme HMMs for each language were trained on a small, phonemically unbalanced speech corpus for each language. The limited training sets restricted the number of parameters which could be reliably estimated. With a larger training set available it would be possible to incorporate e.g. delta-delta cepstrum coefficients in the

parameter vectors and to use full covariance matrices and more mixtures in the HMMs. In addition, context dependent models and explicit duration modelling will probably improve the segmentation performances.

Although the acoustic segmentation coincided well with manual segmentation, and few errors made by the constrained HMM segmentation were due to bad acoustic segmentation, the acoustic segmentation could be improved by enhancing the parameter vectors to include e.g. delta-cepstrum and delta-energy parameters. The cepstral domain filtering may also improve the acoustic segmentation accuracy.

For most of the experiments with the constrained HMM segmentation the acoustic segment boundaries were calculated with 150% oversegmentation. It may be better to try all possible oversegmentation factors for each sentence and select the one providing the highest likelihood in the constrained Viterbi decoding.

Since the automatic segmentation algorithm makes mistakes, a practical segmentation and labelling system should consist of an automatic segmentation software with an interactive component incorporated. This semi-automatic approach makes it possible to check and refine segment boundaries. At the phoneme-transitions known to be difficult for the automatic segmentation algorithm the interactive module should display the speech waveform and broad band spectrogram and allow the human labeller to place the boundary. The interactive module should show the boundary suggestions provided by the acoustic segmentation with e.g. 100% oversegmentation. When the human labeller asks for help, a menu with agreed on conventions for manual segmentations should appear.

A natural extension of the automatic segmentation algorithm is to allow for auditory transcription or TTP-generated transcription as input. The main difference between a TTP-generated label-string and the manual labelling of the actual speech was the high number of substitutions. However, many of the substitutions were not regarded as serious "errors", e.g. [@] substituted /e/ and /O:/ substituted /O/. Thus alternative paths must be possible in the automatic segmentation. Allowing for alternative paths with our segmentation algorithm did not reduce the segmentation accuracy significantly. This indicates that the segmentation performance will not decrease much when TTP-generated label strings are used as input.

The TTP-program also created too many labels, i.e. this transcription may contain phonemes that have no clear acoustic cues. This may force the automatic segmentation algorithm to segment sounds which are not pronounced. However, since our phonemic segmentation algorithm was robust against ripple errors, such problems will probably only cause local "errors".
If a TTP-generated label string should be taken as input to the automatic segmentation algorithm, a speech/silence detector has to be incorporated.

Prosody has not been dealt with in this thesis, but prosodic annotation is important for e.g. improvement of text-to-speech systems, and should be done in the same consistent and standardised manner as for the phonemic labelling used in this thesis.
By use of multiple sensors in the speech recording other signals such as laryngograph signals and nasal air flows, can be measured. Segmentation and labelling of such signals will provide more detailed information for speech research than the speech annotation investigated in this thesis.

For further evaluation of the automatic segmentation algorithm, it should be tested on other

databases containing different speaking styles, e.g. spontaneous speech, and different dialects and phonation types.
For practical use, e.g. for shared use across laboratories, the automatic segmentation algorithm should be evaluated with respect to other criteria, such as *user friendliness* (e.g. easy installation and use, documentation, and flexibility for different test sets), *hardware requirements* (e.g. computer system and memory size), and the amount of manually annotated speech material required for training the ASS software.

In order to annotate speech databases more accurately and reliably, more effort must be spent on standardisation of segmentation conventions.

8.4 CONTRIBUTIONS OF THIS THESIS

The main contributions of this thesis can be summarized as:

* Segmentation and labelling of speech has been argued for and the terminology has been defined.

* Subword units have been compared with respect to the needs and requirements of automatic speech recognition, text-to-speech synthesis, and basic speech research.

* Manual segmentation and labelling strategies for different languages have been compared. We argued that a multi-lingual standard for manual segmentation can be based on these strategies.

* We suggested and applied one manual segmentation and labelling strategy for Norwegian. This strategy can be used as a guideline for other manual segmentation projects for Norwegian, and many of the conventions are general enough to be applied for other languages too.

* We developed a multi-lingual automatic segmentation algorithm which requires minimum training for each new language.

* The connection between manual segmentation and "errors" made by the automatic segmentation has been shown.

* The phonemic segmentation part of the automatic segmentation algorithm has been improved by speaker and environment normalisation by cepstral domain filtering.

* We defined a set of poly-phonemes for Norwegian, Danish, English, and Italian.

* The accuracy of an HMM-based phoneme recogniser was improved by constraining the Viterbi forward recursion with acoustic segment boundaries.

Appendix A

SAMPA - SAM Phonetic Alphabet

SAMPA is defined for phonemic representation, that is, the symbols are used according to the analysis of distictive sound oppositions within each language. Thus although their relation to sound category symbols of the IPA is given, they are symbols of intra-language convention, and do not have an exact language-independent phonetic (auditory or acoustic) equivalence, nor do they represent a single sound within a language [Wells'92, p.7].

IPA	SAMPA	IPA	SAMPA	IPA	SAMPA
a	a	b	b	s	s
ɑ	A	c	c	ʃ	S
æ	{	ç	C	t	t
ɐ	6	d	d	θ	T
ɒ	Q	ð	D	v	v
ɔ	O	f	f	w	w
e	e	g	g	x	x
ɛ	E	ɣ	G	ɥ	H
ə	@	h	h	z	z
ɜ	3	j	j	ʒ	Z
i	i	k	k		
ɪ	I	l	l		
o	o	ʎ	L		
ø	2	m	m		
œ	9	n	n		
Œ	&	ɲ	J		
u	u	ŋ	N		
ʊ	U	p	p		
ʉ	}	r	r		
ʌ	V	ʁ, ʀ	R		
y	y				
ʏ	Y				

Table A.1 *Some IPA and SAMPA symbols (After [Wells'92, pp.4-6]).*

Norwegian Phoneme Inventory

Symbol	Example-Word	Transcription
Vowels:		
A:	hat	hA:t
A	hatt	hAt
{:	vær	v{:r
{	vært	v{rt
O:	vât	vO:t
O	vått	vOt
e:	sen	se:n
e	send	sen
i:	vin	vi:n
i	vind	vin
2:	søt	s2:t
2	søtt	s2t
u:	bok	bu:k
u	bukk	buk
y:	lyn	ly:n
y	lynne	lyne
}:	lun	l}:n
}	lund	l}n
Diphthongs:		
{i	vei	v{i
2y	høy	h2y
A}	sau	sA}
Ai	kai	kAi
Oy	konvoy	ku:nvOy
}i	hui	h}i
ui	hoi	hui

Symbol	Example-Word	Transcription
Plosives:		
p	hopp	hOp
b	labb	lAb
t	lat	lA:t
d	ladd	lAd
k	takk	tAk
g	tagg	tAg
Fricatives:		
f	fin	fi:n
v	vin	vi:n
s	lass	lAs
S	skyt	Sy:t
C	kino	Ci:nu
j	gi	ji:
h	ha	hA
Nasals, lateral and trill:		
m	lam	lAm
n	vann	vAn
N	sang	sAN
l	fall	fAl
r	prøv	pr2:v

In addition there are important allophonic variants (which are phonemes in many dialects) for which transcription has been agreed:

rt	hardt	hArt	(retroflex t)
rd	verdi	v{rdi:	(retroflex d)
rn	gam	gA:rn	(retroflex n)
rl	ærlig	{:rlig	(retrofleks l)
rL	blå	brLO:	(retroflex flap)

In cases where the dental consonants do not change into retroflexes, they are transcribed using the separator sign (ASCII 45); e.g.:

r-d	verdig	v{:r-di

Table A.2 *The Norwegian Phoneme Inventory (tones excluded). After [Kvale'91].*

English Phoneme Inventory

Symbol	Example-Word	Transcription
Vowels:		
"Checked" vowels (short):		
I	pit	pIt
e	pet	pet
{	pat	p{t
Q	pot	pQt
V	cut	kVt
U	put	pUt
Short central vowel:		
@	another	@nVD@
"Free" vowels:		
i:	ease	i:z
ei	raise	reIz
ai	rise	raIz
Oi	noise	nOIz
u:	lose	lu:s
@U	nose	n@Uz
aU	rouse	raUz
3:	furs	f3:z
A:	stars	stA:z
O:	cause	kO:z
I@	fears	fI@z
e@	stairs	ste@z
U@	cures	kjU@z

Symbol	Example-Word	Transcription
Plosives:		
p	pin	pIn
b	bin	bIn
t	tin	tIn
d	din	dIn
k	kin	kIn
g	give	gIv
Affricates:		
tS	chin	tSIn
dZ	gin	dZIn
Fricatives:		
f	fin	fIn
v	vim	vIm
T	thin	TIn
D	this	DIs
s	sin	sIn
z	zing	zIN
S	shin	SIn
Z	measure	meZ@
h	hit	hIt
Sonorants:		
m	mock	mQk
n	knock	nQk
N	thing	TIN
r	wrong	rQN
l	long	lQN
w	wasp	wQsp
j	yacht	jQt

Table A.3 *The English Phoneme Inventory (After [Wells'89]).*

Danish Phoneme Inventory

Symbol	Example-Word	Transcription
Vowels:		
i:	mile	"mi:l@
i	ville	"vil@
e:	mele	"me:l@
e	visse	"ves@
E:	mæle	"mE:l@
E	tække	"tEk@
a:	male	"ma:l@
a	malle	mal@
A:	parken	"pA:g=n
A	pakken	pAg=n
y:	hyle	"hy:l@
y	hylde	"hyl@
2:	køle	"k2:l@
2	kølle	"k2l@
9:	høne	"h9:n@
9	hønse	"h9nse
u:	kule	"ku:l@
u	kulde	"kul@
o:	fole	"fo:l@
o	foto	"fodo
O:	måne	"mO:n@
O:	munde	"mOn@
Q:	kåre	"kQ:r@
Q	kors	kQs
Diphthongs:		
-"are most economically analysed as vowel plus j, v or r" , e.g.:		
aj	hegn	haj?n
Oj	hoj	hOj
uj	huj	huj
iv	ivrig	"ivri
ev	peber	"pev@r
Ev	evne	"Evn@
av	havn	hav?n
yv	syv	syv
2v	døvstum	"d2vsdom
9v	støvle	"st9vl@
ir	kirke	"kirk@
Er	bær	bEr
9r	smør	sm9r
Plosives:		
p	pande	"pan@
b	bande	"ban@
t	tand	tan?
d	dan	dan?
k	kal@	"kal@
g	galde	"gal@
Fricatives:		
f	finde	"fen@
s	stand	sdan?
Approximants:		
v	vinde	"ven@
D	bide	"bi:D@
j	Jul	ju:?l
h	hest	hEsd
Nasals:		
m	mile	"mi:l@
n	ny	ny?
N	lunge	"loN@
Liquids:		
l	land	lan?
R	ride	"Ri:D@

Stød is symbolised by ?

Table A.4 *The Danish Phoneme Inventory (After [Wells'89]).*

Swedish Phoneme Inventory

Symbol	Example-Word	Transcription
Vowels:		
i:	vit	v:t
e:	vet	vet
E:	säl	sE:l
y:	syl	sy:l
}:	hus	h}:s
2:	föl	f2:l
u:	sol	su:l
o:	hål	ho:l
A:	hal	hA:l
I	vitt	vIt
e	vett	vet
E	rätt	ret
Y	bytt	bYt
u0	buss	bu0s
2	föll	f2l
U	bott	bUt
a	hall	hal
Plosives:		
p	pil	pi:l
t	tal	tA:l
k	kal	kA:l
b	bil	bi:l
d	dal	dA:l
g	gås	go:s
Fricatives:		
f	fil	fi:l
v	vår	vo:r
s	sil	si:l
S	sjuk	S}:k
h	hal	hA:l
C	tjock	COk
Sonorants:		
m	mil	mi:l
n	nål	no:l
N	ring	rIN
r	ris	ri:s
l	lös	l2:s
j	jag	jA:g

Allophones and some detailed phonetic transcriptions that are used:

Pre-r allophones:		
{:	här	h{:r
9:	för	f9:r
{	herr	h{r
9	förr	f9r

Retroflexed consonants:		
rt	hjort	jUrt
rd	bord	bu:rd
rn	barn	bA:rn
rs	fors	fOrs
rl	karl	kA:rl

Schwa vowel allophone:

@	pojken	pOjk@n

In cases where the dental consonants do not change into retroflexes, they are transcribed using the separator sign (ASCII 45): r-t, r-d.

Table A.5 *The Swedish Phoneme Inventory (After [Wells'92]).*

Italian Phoneme Inventory

Symbol	Example-Word	Transcription	Symbol	Example-Word	Transcription
Vowels:			Plosives:		
i	mite	"mite	p	pane	"pane
e	rete	"rete	b	banco	"baNko
E	meta	"mEta	t	tana	"tana
a	rata	"rata	d	danno	"danno
O	moto	"mOto	k	cane	"kane
o	dove	"dove	g	gamba	"gamba
u	muto	"muto	pp	coppa	"kOppa
			bb	gobba	"gObba
Fricatives:			tt	zitto	"tsitto
			dd	cadde	"kadde
f	fame	"fame	kk	nocca	"nOkka
v	vano	"vano	gg	fugga	"fugga
s	sano	"sano	Affricates:		
z	sbaglio	"zbaLLo			
S	scendo	"Sendo	ts	zitto	"tsitto
ff	beffa	"bEffa	dz	zona	"dzOna
vv	bevvi	"bevvi	tS	cena	"tSena
ss	cassa	"kassa	dZ	gita	"dZita
SS	ascia	"aSSa	tts	bozza	"bOttsa
Nasals:			ddz	mezzo	"mEddzo
			ttS	braccio	"brattSo
m	molla	"mOlla	ddZ	oggi	"Oddzi
n	nocca	"nOkka			
J	gnocco	"JOkko	Liquids:		
mm	grammo	"grammo			
nn	panna	"panna	r	rete	"rete
JJ	bagno	"baJJo	l	lama	"lama
			L	gli	"Li
Glides:			rr	ferro	"fErro
			ll	colla	"kOlla
j	ieri	"jEri	LL	foglia	"fOLLa
w	uomo	"wOmo			

Table A.6 *The Italian Phoneme Inventory (After [Wells'92]).*

Appendix B

ANALYSES OF THE NORWEGIAN EUROM0 RECORDING

B.1 ANALYSES OF THE ANNOTATED SPEECH MATERIAL

Table B.1 shows on the basis of manual annotation that speaker AFN had 32 phonemes more than speaker TBN, and 72 pauses more than speaker TGN. Other recordings of the same text material suggest that with normal reading tempo the recording time will range between 120 and 140 seconds. According to this AFN read unusually slow.

Speaker	Recording time [sec.]	Speech time [sec.]	Pause time [sec.]	Number of Segments	Number of Pauses	Number of Phonemes
AFN	199	135	64	1270	126	1144
SHN	137	103	34	1214	87	1127
TBN	130	102	28	1183	71	1112
TGN	117	92	25	1171	54	1117
Total	583	432	151	4838	338	4500

Table B.1 *The reading time and the number of segments, pauses and phonemes for the four readers based on the manual segmentation. (SHN means speaker SH with language code N for Norwegian)*

We notice that the slowest reader, AFN, also has most pauses and longest pause time. The pauses constitute a bigger part of the total recording for AFN than for the others; i.e. when we calculate the pause time as the fraction of the recording time, we get 32.2% for AFN and 21.4% for TGN. Hence, AFN used 2.56 times the pause time of TGN, but only 1.46 times longer speech time.

From table B.1 we calculated the **articulation rate** as the average number of phonemes per second within the portions of the recordings which are marked as speech. **Speaking rate** is the average number of phonemes per second when the total recording time is taken into account. Both rates are listed in table B.2 below.

The tendency that differences in recording time manifest themselves most clearly in the number and duration of pauses, and not that much in articulation rates, is also reported in similar experiments, e.g. [Fant'91]. This can also be expressed as the articulation rate of speaker TGN is 1.43 times the articulation rate of speaker AFN, whereas TGN's speaking rate is 1.66 times faster than that of AFN. All speakers were easily understood. Hence, there is no direct correlation between faster speaking rate and an unprecise or slurred articulation.

Speakers	Articulation rate	Speaking rate	Average phoneme duration [ms]	Max. phoneme duration [ms]	Min. phoneme duration [ms]
AFN	8.5	5.8	118	390	22
SHN	10.9	8.2	91	288	15
TBN	10.9	8.6	92	322	15
TGN	12.1	9.6	82	316	16
All	10.4	7.7	96	390	15

Table B.2 *Articulation rate, speaking rate, average phoneme duration and maximum and minimum phoneme duration for the annotated Norwegian EUROM0 speech material.*

Dew, [Dew'77], claimed that a speaking rate of about 10 phonemes per second on average or 160-170 words per minute is normal when reading a text aloud. Speaking rates below 7 phonemes/sec., corresponding to about 130 words per minute, was regarded as "awfully slow" [Dew'77, p.224]. An articulation rate of 15 phonemes/sec. is here taken as an upper limit. When trying to talk faster than this problems with coordinating articulatory gestures occur.
As seen in table B.2, the 4 Norwegians had a speaking rate below 10 phonemes/sec. Our text contained 297 (orthographic) words. Dividing this number by the recording time, we get approximately 90 words/min for speaker AFN, 130 for SHN, 137 for TBN, and 152 for TGN - that is, rather slow reading according to Dew's measure.
In a Swedish text reading experiment, [Fant'91], a professional TV speaker read a short story at different speeds. Fant described 130 words per minute as "normal" reading mode. Hence, calculating "words per minute" may not be an appropriate measure of reading tempo, at least as long as the text material is not better documented.

Speaker TGN was the most difficult one to segment and label, mainly due to TGN's fast articulation rate compared to the others. Especially in long voiced sequences the acoustic cues tended to be smeared much more than was the case for the other speakers. However, we also noticed what Fant, [Fant'91], calls *lack of uniformity*, in that the fastest reader did not articulate fastest in all parts of the text.

In [Fant'91] the average phoneme duration for normal reading mode was 75 ms, for faster 70 ms, for slower 78 ms, and for distinct reading mode 89 ms. The average phoneme duration is calculated as number of phonemes divided by the speech time. Compared to this, the Norwegian informants read very slowly (see table B.2).

We also noticed that the phonemically defined long vowel segments really had longer duration than the short ones. The mean duration of a long vowel was 145 ms, and a short 85 ms. But, as can be seen in figure B.1 below, the distributions overlap to such an extent that duration is not a good clue for distinguishing the long and short vowels e.g. in an ASR system.

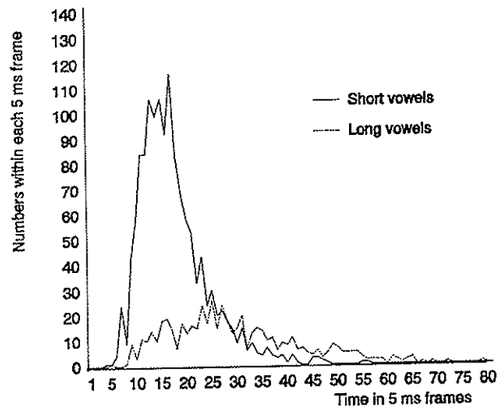


Figure B.1 *Duration of phonemically short vowels (whole line) and phonemically long vowels (dotted line). The calculation is based on all four Norwegian EUROM0 speakers.*

Figure B.2 shows that the unvoiced plosives have longer duration than the voiced ones, with average durations of 105 ms and 65 ms respectively.

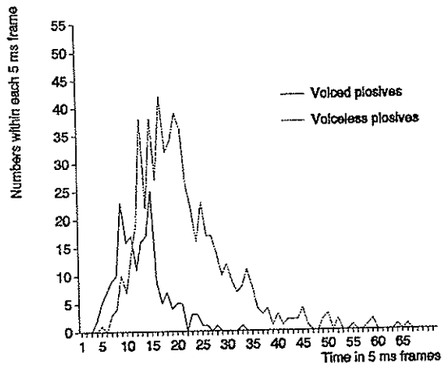


Figure B.2 *Duration of unvoiced plosives (whole line) and voiced plosives (dotted line). The calculation is based on all four Norwegian EUROM0 speakers.*

If the /r/ is realised as an alveolar tap it is per definition, impossible to extend. However, if /r/ appears at the end of an utterance the last part of it may change into an approximant with slowly decreasing voicing and amplitude. This part is included in the /r/-segment yielding a "long" /r/. In our material, the mean duration of /r/ was 11 frames, see figure B.3 below.

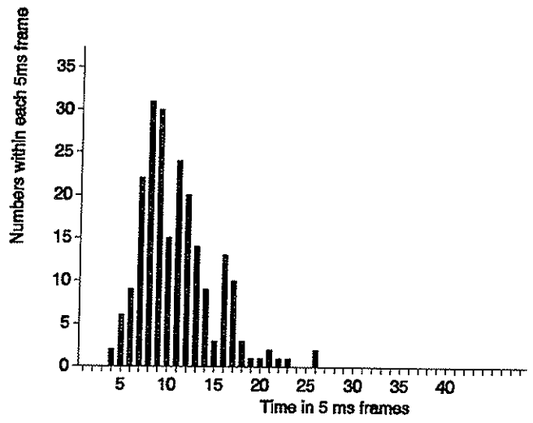


Figure B.3 *Duration of /r/ segments in the Norwegian EUROM0 recording.*

Most of the examples from the Norwegian EUROM0 recordings are taken from the male speaker AFN, who spoke most clearly. In figure B.4 below one second of the waveform for the three other speakers is shown. The second male speaker TGN, in the top row, spoke fastest of them all, and produced 14 phonemes /i A:ritmetiken kA/ within the one second time slot. The female speakers SHN and TBN uttered 9 and 11 phonemes respectively. But, e.g. the /m/-segment for TBN is shorter than the corresponding segment for TGN. That is, the fastest speaker is not the fastest one for all sounds.
We also notice the creaky voice realisation of the vowel to vowel transition across word boundary /i:/- /A:/, which is especially clear for SHN.

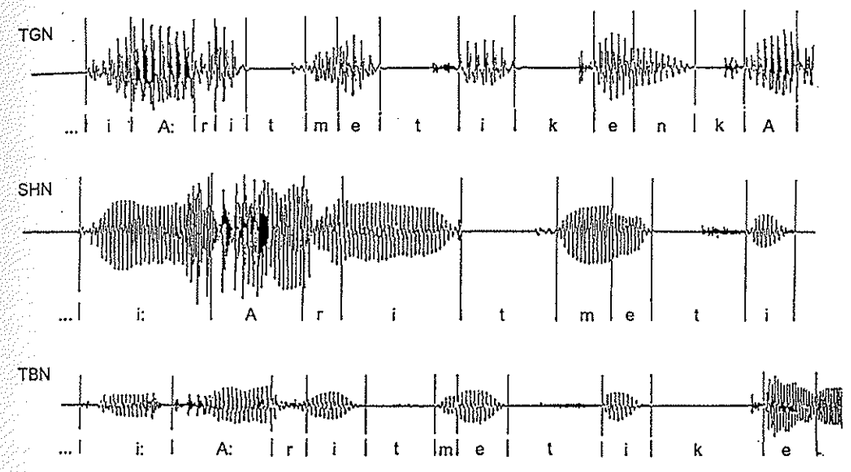


Figure B.4 *Examples of the three other speakers on The Norwegian EUROM0 recordings; in the top row TGN, in the middle row SHN, and in the bottom row TBN. One second was excerpted from the beginning of the sentence "i aritmetikken" (=in the arithmetics).*

B.2 SOME DETAILS OF THE TRANSCRIPTION LEVELS

Below some typical examples from the five transcriptions are shown to illustrate the difference between them. In the examples the phoneme(s) we want to highlight are typed in bold letters.

The **Text-to-Phoneme, TTP** transcription assumes an especially clear articulation, as shown below:

Orthography	Transcription used:	Expected phoneme sequence:
uendelig	/}:@n@li/	/}:@n**d**@li/
bokstaver	/b**O**kstAv@r/	/bu**k**stAv@r/
tjue	/**tj**}:e/	/**C**}:e/
språket	/sprOk@**t**/	/sprOk@/
ettersom	/et@**rs**Om/	/et@**S**Om/
forskjellig	/fOr**S**eli/	/fO**S**eli/
komma	/k**u**mA/	/k**O**mA/

In addition, the **TTP** transcription does not differ between "en" ('one') /e:n/ and "enn" ('than') /en/.

In the **citation** form of the words every "og" is transcribed /O:/ (not /O:**g**/). For the word "av" both /A:/ and /A:v/ can be used. We have chosen /A:v/. The word "avhengig" is transcribed /A:vheN**g**i/.

When comparing **phonotypical** transcription with the TTP or Citation form, the most typical difference is the retroflex assimilation over word boundaries, as when /b@nyt@**r ti**/ becomes /b@nyt@ **rt**i/, /gAN@**r ti**/-/gAN@ **rt**i/, /O:**r s**e:n@r@/-/O: **S**e:n@r@/, and /s}me:r@**r s**ifr@n@/-/s}me:r@ **S**ifr@n@/.

In the **phonotypical** transcription there are generally fewer lengthmarks than in the citation form. The /e/ is substituted by /@/, and some /@/s disappear as when /}:end@li/ becomes /}:endli/, and /tAl@n@/-/tAln@/. When the @ occurs in a plosive-@-nasal context, the @ is elided and a nasal release of the plosive is expected such as /syt@n/ becomes /syt=**n**/ and /At@n/ becomes /At=**n**/. (We do not mark syllabic nasal in our transcription, hence we transcribe /sytn/ and /Atn/).
In **phonotypical** transcription also some /r/s disappear, as when /nO**r**/ becomes /nO/ and /A:vj2:**r**@ls@/ becomes /A:vj2ls@/ (both /r/ and /@/ is elided). Other reductions and elisions are /Oph2yd/ becomes /Op2yd/, /**ve: e**n/-/ve n/, /**til**/-/te/, /A:**v**/-/A/, and /Se**l**dn@/-/Seln@/.

In **Broad Phonetic Auditory Transcription** even fewer lengthmarks than in phonotypical transcription are marked.
For speaker TBN the word "svært" in "svært sjeldne" is missed out. Speaker TBN had some pronunciations which differed from the other speakers, such as /nO vi/ instead of /nO**r** vi/, /si@**r d**}sin/ instead of /si@ **rd**}sin/, /eNgAN@r/ instead of /e**n** gAN@r/, and /{ **S**t2r@/ instead of /e**r s**t2r@/. Speaker AFN pronounced /ve jelp/ instead of /**m**e jelp/ and /j}ste:r@**r d**em/ instead of /j}ste:r@ **rd**em/.

In the **Manual segmentation** we found that speaker AFN said /niten/ instead of /nitn/ and that speaker TBN pronounced "sett av **b**okstaver" /setAvukstAver/ (/**b**/ disappeared), "posisjo**n**en" as /pusiSu**n**n/, "sammenfalne" as /sAmnfAlne/, and "Russland" is pronounced /r}S**l**An/ (instead of /r}s**l**An/). Speaker SHN articulated a clear /k/ in "fulgte" /f}l**k**te/ instead of /f}lte/ and "valgte" /vAl**k**te/ instead of /vAlte/. The word "snes" got an epenthetic [t] between /s/ and /n/, i.e. [stnes]. Speaker TGN pronounced "tallsystem**et**" as /tAlsystem/, "posisjo**n**en" as /pusiSun/, "systemene" as /systmne/, "valg**t**e å" as /vAltO/, "**s**om" as /sm/, "**m**ed hjelp" as /mjelp/ (instead of /mejelp/), "avhengig av" as /AveNjA/ (instead of /AvheNiA/), and "pluss ni" as /pl}seni/ (an extra /e/). Speaker TGN had no pause when reading the numbers.

In the **Manual segmentation** geminate phonemes with no acoustic cues to separate them, are marked off as one single segment. For instance the numbers eight one nine, "åtte en ni" will be labelled /Ote:ni:/ instead of /Ote e:n ni/, i.e. one number has disappeared.

The Norwegian EUROM0 recording was not phonemically balanced. Of a total of 4500 phonemes detected in manual segmentation and labelling, /e/ was found 446 times, /e:/ 94 times and /@/ 180 times, (together they constitute 16% of the database), while the phonemes /y: r r-l O }i u/ were not represented at all. The 10 most frequently occurring phonemes constitute over 60% of the database.
The phonemes were divided into classes and the number of occurrences for each phoneme group was calculated: Short vowels (and @) 1385, long vowels 544, diphthongs 32, fricatives 696, plosives 745, nasals 597, /r/ 219, /l/ 214 and retroflexed consonants 68, as shown in figure B.5.

Training statistically based automatic speech segmentation procedures on this speech material will give poor models for most of the phonemes.

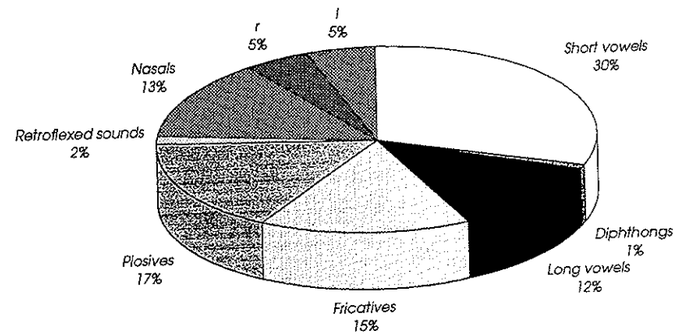


Figure B.5 *Distribution of the phoneme classes in the Norwegian EUROM0 recordings.*

Table B.3 below provides data for the number of occurrences of the phonemes for each speaker in the manual segmentation and auditory transcription of speaker AFN and TBN and in the phonotypical, citation, and TTP-generated transcriptions.

	Manual segmentation						Aud.Trans.		Orthographic based		
Symbol	AFN	SHN	TBN	TGN	Sum	Mean	TB	AF	Ph.typ	Cita.	TTP
b	10	10	9	10	39	10	10	10	10	10	10
d	20	18	19	19	76	19	19	20	19	24	22
g	17	17	17	16	67	17	17	17	17	17	16
p	27	27	27	27	108	27	27	27	27	27	27
t	77	79	77	79	312	78	78	78	78	87	94
k	36	37	35	35	143	36	36	36	36	36	36
l	53	55	52	54	214	54	53	54	53	55	55
m	52	53	53	53	211	53	53	52	53	53	53
n	85	82	83	83	333	83	86	87	87	87	87
N	13	13	14	13	53	13	14	13	13	13	13
s	98	94	93	94	379	95	95	98	96	98	100
v	23	22	21	22	88	22	20	21	21	26	26
f	29	29	29	29	116	29	29	29	29	29	29
S	19	20	20	20	79	20	22	20	22	19	18
C	1	1	1	1	4	1	1	1	1	1	0
j	5	5	4	6	20	5	8	6	8	5	6
h	4	3	3	0	10	3	4	10	4	10	10
r	59	54	51	55	219	55	55	57	50	73	75
rt	12	10	11	10	43	11	10	11	11	2	2
rd	4	6	6	5	21	5	5	4	5	1	1
rn	0	0	0	1	1	0	0	0	0	0	0
rl	1	0	1	1	3	1	0	0	0	0	0
rL	0	0	0	0	0	0	0	0	0	0	0
r-l	0	0	0	0	0	0	0	0	0	0	0
A	52	54	54	52	212	53	61	61	61	52	57
e	112	124	115	95	446	112	65	71	63	58	70
i	49	52	56	56	213	53	68	71	63	51	64
O	37	36	34	35	142	35	55	52	54	33	44
u	5	4	4	5	18	5	9	10	6	5	9
{	2	3	2	2	9	2	6	6	5	3	3
2	4	5	5	3	17	4	5	5	5	3	3
y	12	12	12	12	48	12	12	12	12	12	12
}	25	26	24	25	100	25	28	28	29	24	23
@	48	37	40	55	180	45	97	90	99	110	112
A:	19	17	17	19	72	18	10	10	10	19	14
e:	26	22	23	23	94	24	20	21	21	30	16
i:	47	43	40	41	171	43	25	25	30	46	33
O:	24	22	23	23	92	23	4	7	5	26	10
u:	12	11	12	11	46	12	8	7	11	12	13
{:	5	5	4	5	19	5	1	2	3	5	4
2:	2	1	1	3	7	2	1	1	1	3	3
y:	0	0	0	0	0	0	0	0	0	0	0
}:	10	10	12	11	43	11	8	8	7	12	12
{i	1	1	1	1	4	1	1	1	1	1	1
2y	6	6	6	6	24	6	6	6	6	6	6
A}	1	1	1	1	4	1	0	0	0	0	1
Oy	0	0	0	0	0	0	0	0	0	0	0
}i	0	0	0	0	0	0	0	0	0	0	0
ui	0	0	0	0	0	0	0	0	0	0	0

Table B.3 *Number of occurrences of the phonemes in the Norwegian EUROM0 recordings (cont. next page).*

	Manual segmentation						Aud.Trans.		Orthographic based		
	AFN	SHN	TBN	TGN	Sum	Mean	TB	AF	Ph.typ	Cita.	TTP
Sum phonemes:	1144	1127	1112	1117	4500	1125	1132	1145	1132	1184	1190
Long vowels:	145	131	132	136	544	138	77	81	88	153	105
Short vowels:	298	316	306	285	1205	301	309	316	298	241	285
Short+long+@:	491	484	478	476	1929	482	483	487	485	504	502
Diphthongs:	8	8	8	8	32	8	7	7	7	7	8
Sum voiced:	841	817	816	822	3296	824	830	835	828	875	874
Sum unvoiced:	303	300	296	295	1194	298	302	310	304	309	316
Plosives:	187	188	184	186	745	187	188	188	187	201	205
Fricatives:	179	174	171	172	696	175	179	185	181	188	189
Nasals:	150	148	150	149	597	149	153	152	153	153	153
Retroflexed:	17	16	18	17	68	17	15	15	16	3	3

Table B.3 *Number of occurrences of the phonemes in the Norwegian EUROM0 recordings. Sum* means the sum of the occurrences of the manual segmentation, and *Mean* is the sum divided by 4 and rounded to nearest integer. The abbreviations are explained in chapter 4.

B.3 THE NORWEGIAN EUROM0 TEXT

I språket kan vi skrive uendelig mange ord med et lite sett av bokstaver. I aritmetikken kan uendelig mange tall settes sammen av bare noen få siffer - med hjelp av symbolet null, posisjonsprinsippet og grunntallsprinsippet. Rene systemer med grunntall fem og seks sies å være svært sjeldne, men grunntallene tolv og tjue benyttes på norsk når vi sier "dusin" og "snes", som i "tre og et halvt snes".

Ettersom tiden gikk kunne ingen systemer holde tritt med det desimale - eller arabiske - tallsystemet som benytter 10 siffer, tallene null, en, to, tre, fire, fem, seks, sju, åtte og ni i tillegg til komma.

Sifrene får forskjellig verdi avhengig av posisjonen. Slik kan tallet åtte en ni komma seks fem skrives som åtte ganger ti opphøyd i to pluss en ganger ti opphøyd i en pluss ni ganger ti opphøyd i null pluss seks ganger ti opphøyd i minus en pluss fem ganger ti opphøyd i minus to.

Desimaltallene er en hjelp ved addisjon. Vi summerer sifrene i sammenfallende posisjon og justerer dem som er større enn ni ved en prosess som kalles mente. Som et godt eksempel; prøv å legge sammen tallene fire sju komma ni en fem, fem komma tre fem sju og seks ni komma åtte. Først legger du sammen fem og sju og får tolv ganger ti opphøyd i minus tre. Summen du får, skal ikke være sju tre sju komma null sju seks.

Pengesystemene har etter hvert gått over til å benytte ti som grunntall. Frankrike var det første "desimale" land i Europa i sytten nittini etterfulgt av Belgia, Italia og Sveits i atten sekstifem. Tysklands avgjørelse fulgte åtte år senere, og de skandinaviske land og Russland skiftet i atten syttifem. Storbritannia valgte å innføre desimale penger bare nittiseks år etter dette, i nitten syttien.

B.4 THE NORWEGIAN TI-TEST

The division of the speech corpus for the Norwegian Text Independency test was as follows:

Speaker	Part 1 (samples / utterances)	Part 2 (samples / utterances)
AFN	17280-1610591 / 0-62	1612767-3212191 / 63-126
SHN	7248-1091312 / 0-45	1100527-2200368 / 46-87
TBN	16256-1051504 / 0-38	1052399-2089968 / 39-71
TGN	21711-937664 / 0-24	956000-1892112 / 25-54

Table B.4 *Division of the Norwegian EUROM0 passage for the TI-test.*

This division assigned 2257 phonemes to the first part and 2243 phonemes to the second. The results will be presented as an average of these two parts since the results are similar for each of them.

B.5 RECORDING PROTOCOLS

The speech data was recorded in the anechoic chamber at the department of electrical and computer engineering of the Norwegian Institute of Technology with the following equipment:

Microphone: AKG CK22, no. 8760, with muffler.
Emitter-follower: AKG C 460 B. 50Hz highpass, 0 dB, no. 5083.
Two-channel microphone amplifier. Upper channel. 48V phantom.
Signal processing card: OROS AU21.
AD/DA-filter: 7kHz.
Amplifier: Norwegian Electronics, LN 420 (Lab. no. CB4044). Amplified +10dB.

Appendix C

ANALYSES OF THE OTHER EUROM0 ANNOTATIONS

C.1 ANNOTATION OF THE DANISH EUROM0 RECORDINGS

The Danish manual segmentation and labelling was performed at a level corresponding to phoneme units [Dyhr'92]. The phonemic SAMPA units for Danish are listed in Appendix A. As shown below it is not a purely broad phonetic segmentation and labelling approach, but rather an intermediate level between roughly narrow phonetic and a broad phonetic level of annotation. As pointed out in chapter 3.1, it may be difficult to stick consistently to such a level.
"The boundaries are primarily placed using wide-band spectrograms" [Dyhr'92]. That is, the Danish neglect the most accurate source of information, namely the waveform. It is not explained how boundaries were placed when relying on auditory characteristics only.

We investigated the actual annotation of a part of the Danish EUROM0 continuous passage, speaker BL, and discovered many discrepancies from the definitions given in Appendix A. For instance:

a) No length distinction for vowels is employed.
b) The plosives are divided into two segments, e.g. for the /d/, the closure part is denoted /d0/, and the burst part /d/.
c) Use of five unrounded front vowels.
d) The symbol 0 is used as a silent pause mark, instead of SAMPA symbol ...
e) The symbol /u/ is used for [}]-sounds.

Comments on these points:

a) Since there is phonemic contrast between long and short vowel phonemes in Danish this should be labelled.

b) This division is not phonemicly based but is rather an acoustic or signal processing annotation approach. The closure and the burst phase constitute a plosive together, and should be marked as one segment with one phonemic symbol.
It may be argued that the closure and burst segments can easily be joined afterwards, and for some applications, such as automatic phoneme recognition, a splitting of the plosives into a closure part and a burst part may be useful. However, when the task is broad phonetic annotation, one should constrain oneself to phonemic symbols only.

c) Phonemicly Danish has four unrounded front vowels as defined in Appendix A. In Danish an allophonic rule states that if an unrounded front vowel precedes an /R/ the vowel is pronounced more open; e.g. a succeeding /r/ will cause [e] to become an [E]. Including a fifth unrounded front vowel in a broad phonetic annotation is thus unnecessary. But at a more detailed level, as chosen here, more symbols will give more accurate description of the actual speech sounds. At such a level a vowel trapeze for each speaker or dialect is even more important for the documentation of the database.

d) and e) This seems unnecessary, but if it is done consistently it is easy to substitute the correct symbols.

In the following discussion the numbers refer to figure C.1. Number 1, 2, 3, 7, 8, 22, 23, 25, 27 and 28 are informative and illustrates the strategy of [Dyhr'92]. More doubtful annotation is shown at 4, 5, 6, 11, 12, 14, 15, 16, 19, 23, 24, 26 and 30. Segmentation errors are pointed out at 9, 10, 13, 17 and 20.

1: When a plosive begins a sentence the start is marked at the onset of the burst.
2: All the frication is included in the fricative, even if it is superimposed on the voicing of the surrounding vowels. This is the same convention as ours, but it is not applied consistently throughout the speech file.
3: The boundaries for /v/ in intervocalic position are placed on an auditory basis. For such a short segment this decision has to be rather arbitrary.
4, 21: Many closure parts of what is denoted voiced plosives, i.e. b0, d0 and g0, are completely voiceless, for instance this /d0/ in "stod" (see also /b0/ in "spis" and /d0/ in "frokostpausen"). Hence, these sounds must have been labelled entirely based on a subjective perception of the phonemes and not based on what is seen in the actual spectrogram or waveform. This is in contradiction to the strategy applied when dividing a plosive into its closure and burst parts, because this strategy has to be entirely based on visual information.
5: The segment labelled /e/ to us sounds like [i] both in isolation and in context.
6, 14, 18: When two plosives follow immediately after each other, the first one is not released, as /d0 p0/ in "midt på". It is not explained how to place the boundary in the silent /d0/-/p0/ area, and it seems arbitrarily placed.
A natural convention for placing the boundary in the middle of the silent segment would be easy to apply consistently. However, the Danish annotation has not divided the silent area into two segments of equal duration.
7: A very short vowel (13ms).
8: A strange plosive burst waveform.
9: The word "torvet" is labelled /tQD/ with /Q/ as the only vowel although the /Q/-segment sounds like [AOQ]. Hence the /Q/-segment should be further segmented. (Clear cues between the vowels are found in the spectrogram, as an intensity drop between /A/ and /O/ and a creaky voice portion between /O/ and /Q/).
10: A serious mistake in that the perceived /v/ is not segmented. The /i/ segment should be divided into a /i/ and a /v/ segment.

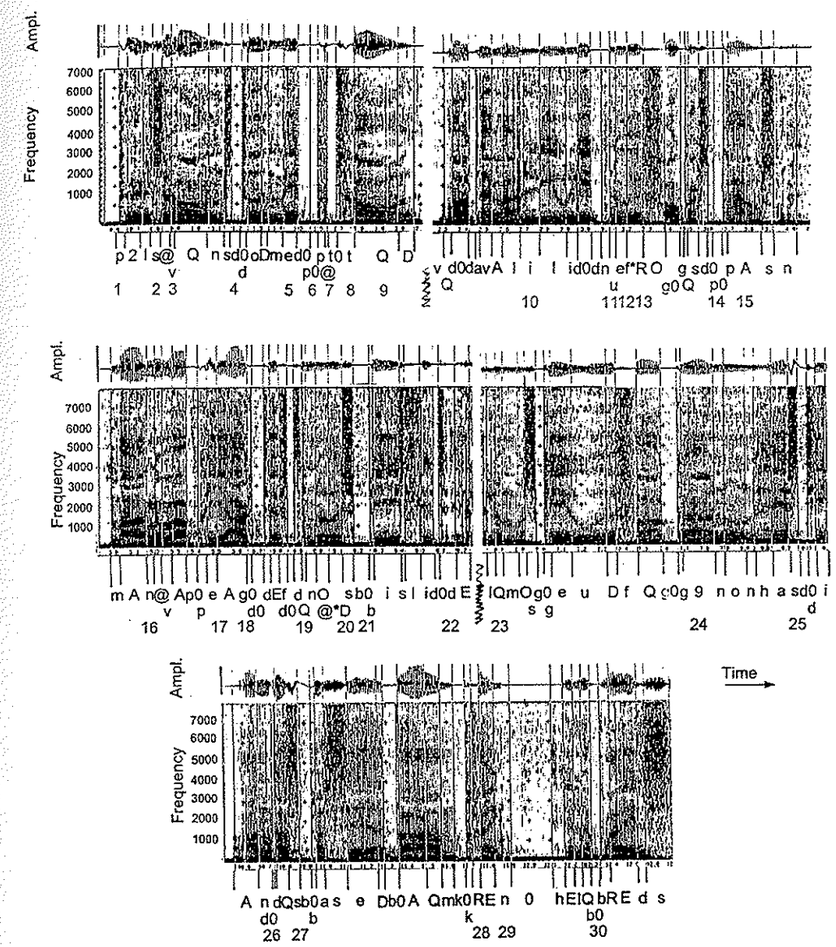


Figure C.1 *Waveform and spectrogram for the first part of the Danish EUROM0 recordings of speaker BLD, manually segmented and labelled according to [Dyhr'92].*
*At the top left side: "pølsevognen stod midt på torvet" /...p2ls@vQn sd0oD med0 p0p@ tQD/, after the break: "hvor der var livligt nu i frokostpausen" /vQ d0da vA lilid0d nu e f*ROg0gQsd0p0pAsn/.*
*In the middle left side: "mange var på jakt efter noget spiseligt e(ller)" /mAn@ vA p0pe Ag0d0d Efd0dQ nO@*D sb0bislid0d E/, which continues after the break: "ller måske ude for at gøre nogle hasti(ge)" /lQ mOsg0ge uD fQ g0g9 non hasd0di/.*
In the bottom: "andre spaserede bare omkring eller bredte s(ig)" /And0dQ sb0baseD b0A Qmk0kREn hElQ b0bREd s/.

11: The segment labelled /u/ sounds like [}] (see point e above).
12: The segment labelled /e/ sounds like [2].
13: In some instances the very same segment is labelled with two different symbols, where the first one is marked with an additional asterisk, as shown in the segment labelled /f*R/ in the word "frokost". This segment sounds like [f]. When listening to a larger portion of the speech signal we perceive an [R] succeeding the /f/, but no cues for [R] is found in the signal.
According to the definitions of linear segmentation and labelling, this labelling is not correct. In a similar problem as described above we would have squeezed an /R/-segment in between /f/ and the following /O/. Then the labels are marked in perceived order and the linear segmentation approach is maintained.
(The Danes should at least document and explain the strategy and indicate which symbol to use when e.g. the database is employed for training an ASS-method).
15: The word "pausen" labelled /p0pAsn/ sounds as [pA}sn], i.e. with a diphthong (or, alternatively, two vowels). A creaky voice area is seen in the spectrogram in the middle of the segment, hence two vowel segments is the best choice.
16, 29: Some /n/-segments should be marked as /N/, as in "mange" labelled as /m A n @/, where the nasal is perceived as a /N/, and a "velar pinch" is observed in the spectrogram.
See also "omkring" labelled as /Q m k0 k R E n/.
17: The words "på jakt" are labelled /p0 p e A g0 d0 d/, but since the phoneme sequence /p @ j A g d/ is perceived (and expected), these labels should be aligned with the signal. Thus the /e/-segment should be divided into a short /@/-segment and a /i/ or /j/ segment.
Another example of missing /j/ is "vej" labelled /vA/; which should has been labelled /vAi/ or perhaps /vAj/.
19, 23: This is an example of segments labelled with /Q/ but that sounds [@] or [{] to us.
20: Same type of error as in 13. We would have placed a /d/ at the end of the /@/ and let the burst of /d/ include e.g. 10 ms of the succeeding /s/.
22: After the burst of /d/ is finished a period of silence indicating a word boundary appears in the waveform. This silence is here included in the /d/. We would have marked a segment of silence.
24: The word "gøre" ('do') is labelled /g0g9/, but the /9/ could be divided into /2/ and /@/, which is perceived when listening, and the boundary between them is seen as intensity difference in the fourth formant in the spectrogram.
25, 27: Notice that the /s/ is realised as frication superimposed on a single wave (the whole /s/ is one wave).
26, 30: The subdivision of plosives often implies unnecessarily short segment durations, as in the word "andre" /A n d0 d Q/ where the duration is one pitch period for each of /d0/ and /d/. In the word "det" /d0de/ the /d0/ is a segment of no duration (starts and ends at sample 280754).
28: /R/ is often realised as voiceless, see also "indryk" ('impression') /ent0tR9g0/, "tre ristede" ('three shook') /t0tRE REsd0d@*D/. The /R/ is also voiced as in "sceneri" ('scenery') /senQRi/, where the acoustic realisation is similar to the South Western Norwegian /R/.

Below the tables C.1 and C.2 correspond to the Norwegian analyses in tables B.1 and B.2 respectively:

Speaker	Recording time [sec.]	Speech time [sec.]	Pause time [sec.]	Number of Segments	Number of Pauses	Number of Phonemes
BLD	120.8	86.5	34.3	1204	46	1158
CHD	94.8	73.4	21.4	1142	38	1104
JDD	121.2	96.8	24.4	1230	49	1181
LFD	105.1	82.3	22.8	1184	45	1139
Total	441.9	340.0	102.9	4760	178	4582

Table C.1 *The recording, speech and pause time and the number of segments, pauses and phonemes for the four readers based on the manual annotation of the Danish EUROM0 continuous passage.*

Speakers	Articulation rate	Speaking rate	Average phoneme duration [ms]	Max. phoneme duration [ms]	Min. phoneme duration [ms]
BLD	13.4	9.6	74.7	300	4
CHD	15.0	11.6	66.5	209	6
JDD	12.2	9.7	82.0	306	10
LFD	13.8	10.8	72.2	335	12
All	13.5	10.4	74.0		

Table C.2 *Average articulation and speaking rate, average phoneme duration and maximum and minimum duration for the Danish EUROM0 continuous passage.*

C.2 ANNOTATION OF THE SWEDISH EUROM0 RECORDINGS

The Swedish strategy was to "stick to phoneme strings as much as possible and regard the different realisations as a research objective" [Nord'90, p.3]. This principle manifests itself in that e.g. different voicing realisations do not alter the phoneme label. For instance, even when a /v/ is devoiced and sounds as [f], it is still labelled /v/. Extra sounds, irregular onsets and noisy disturbances are ignored and epenthetic sounds are mostly included in the adjacent phoneme segments. These examples are similar to our approach, e.g. [p@rAte] labelled /prate/.
However, in some instances the strategy may have been applied too restrictively, as in "har ti" ('has ten') labelled /hA:r ti:/, not /hA:rti/, i.e. retroflexing over word boundaries, and /elEr arA:bIska/ where the segments /Era/ sounds [Ea] only. These discrepancies from my labelling may be due to the Swedish approach of manually aligning the phoneme string from a text-to-speech program, which does not take reductions across word boundaries into account.

On the other hand phonemic assimilation where not only the voicing feature is altered, is labelled as it is perceived. For instance, "har du sett" usually pronounced as [hAr} set], (the /d/ is not pronounced), is labelled /hAr} set/, and "är" [{:r] pronounced [e:] is labelled /e:/.
Geminate sounds are labelled as one phoneme, e.g. "med en" /me:n/ and "ett decimal" /etesImA:l/.

The Swedish did "not particularly try to adjust their thinking to acoustic events, although this would result in better correspondence with automatically generated segment boundaries" [Nord'90, p.3]. For instance in the word /sifrUrna/ no acoustic cues are seen for the /r/.

We investigated some sentences of speaker GW. Figure C.2 below shows the waveform and spectrogram for his first sentence /UEndlIt mONa u:rd kan i: spro:k skrivas me:n liten mEN bu:kstEv{r/, and exemplifies the Swedish annotation approach. However, at number 1 a mistake is detected and at 4 and 7 the boundary placement deviate from our approach.

1. They do not follow their own principle of placing the boundary at "the zero-crossing preceding a strong positive going peak" [Nord'90].
2. Both plosives are released as in Norwegian.
3. A brief period of silence between /s/ and /m/ indicates a word boundary. This pause is included in the /m/-segment. I would probably marked a pause segment.
4. Geminate sounds are not separated; "med en" is labelled /me:n/.
5. No visible burst for the /b/ is noticed.
6. The start of /E/ is placed according to the guidelines, but we would have placed it one pitch period (11 ms) earlier.
7. Difficult to make a decision, but perhaps the end of /r/ should be marked two pitch periods later?

Nord's conventions for the start of voicing and for the transition between voiceless and voiced segments are contradictory, and one of them has to be selected. In the word /samamsEtas/ the onset of voicing convention is applied.

Due to practical computer problems only the qualitative analysis is provided for the Swedish EUROM0.

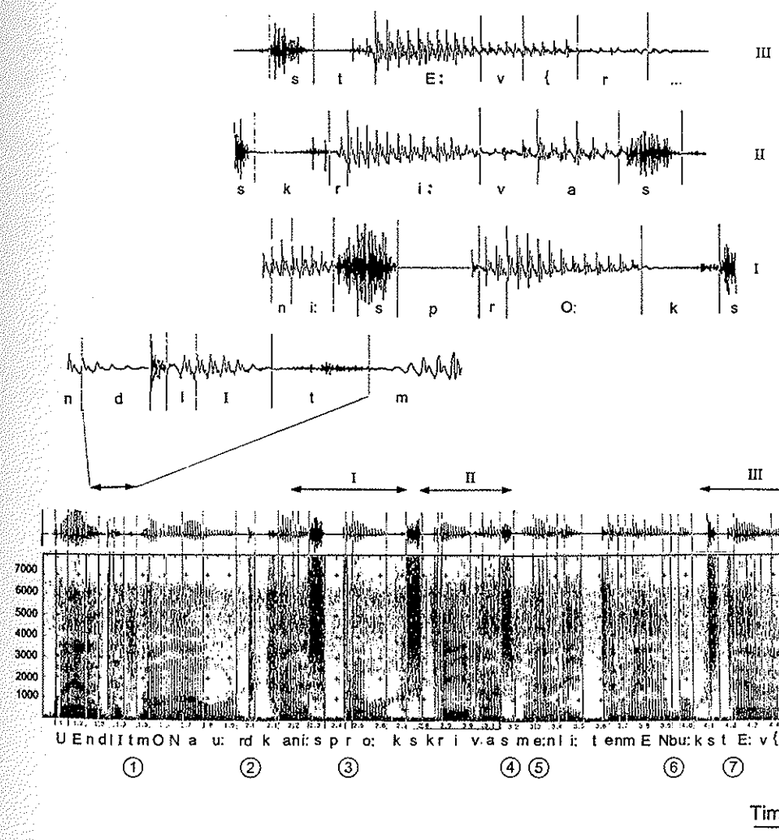


Figure C.2 *Waveform and spectrogram of the utterance by speaker GW /UEndlIt mONa u:rd kan i: spro:k skri:vas me:n li:ten mEN bu:kstE:v{r/ manually segmented and labelled according to [Nord'90]. The + signs in the spectrogram are separated by 0.1 sec. horizontally and 1000 Hz vertically.*

C.3 ANNOTATION OF THE ENGLISH TEST PASSAGE

Since the English Test Passage was used for testing the competing ASS algorithms in the SAM project [Barry'91b], the manual annotation was more carefully investigated.
The English segmentation and labelling strategy, as described in [Barry'90 b], seems similar to ours with respect to e.g. nasals placed between vowels. Also the level of accuracy in the annotation strategy is appreciated, exemplified by the convention of marking the boundary between a plosive and a succeeding vowel at the first positive zero-crossing of the vocalic voice period seen in the waveform. Another convention states that all friction of a fricative should be included in the fricative, even when the voicing begins before the frication is finished[1]. However, these rules may be contradictory when the aspiration of a plosive continues into the voicing of the following vowel.

The strategy of the segmentation is marking segment boundaries at acoustic discontinuities in the waveform or spectrogram when these correspond to transitions between phonemes. The labels are phonemic as listed in Appendix A, and, as for Norwegian, the actual voicing and friction content of a phoneme realisation does not alter the phoneme label; for instance a fully devoiced /z/, looks like a [s] and sounds as [s] in isolation, but the segment is labelled /z/ and not /s/, as in point 36 (the last /z/ in figure C.4).

The broad phonetic manual labelling of the Test Passage contained nine sentences. Each sentence of the Test Passage is analyzed and commented on and some labels and boundaries are marked with numbers, e.g. 23 means the third point in the second sentence. The orthographic transcription and the phonemic label strings for these sentences are (spaces are introduced here to improve the readability):

s1:"When a sailor in a small craft"
/[11]wen @ seIl@r n @ s[12]mO:l kr[13]A:ft/
s2:"faces the might of the vast Atlantic Ocean today,"
/feIsIz[21] D@ maIt[22] @v D@ [23] A:s[24]t @[25]tl{[26]nS[27]tIk @USn tdeI/
s3:"he takes the same risks that generations took before him."
/hi: teIks D@ seIm[31] r[32]Isk[33]s D@t[34] dZe[35]n@reISnz[36] tUk[37] bI[38]fO: hI[39]m/
s4:"But,"
/bV[41]t/
s5:"in contrast to them,"
/In kQnt[51]rA:st@ Dem/
s6:"he can meet any emergency that comes his way"
/hi: k@n[61] mi:t[62] eni: @m3:dZnsi: D@[63]t[64] kVmz[65] Iz w[66]eI/
s7:"with a confidence that stems from a profound trust in the advances of science."
/wID @ kQn[71]f@d@ns D[72]@t stemz[73] rm @ p[74]f@f@Und[75] t[76]rVst @n i: @d[77]v[78]A:nsz @v saI@ns/
s8:"Boats are stronger and more stable,"
/b@Ut[80]s @ str[81]QNg3:r[82] @n mO: steIbl/
s9:"protecting against undue exposure;"
/pr[91]@tek[92]tIN @geInst Vn[93]dZu:@k[94]sp@UZ@/

[1] This is in contrast to our convention stating that all voicing should be included in the vowel.

We immediately notice the typical **reductions** as "a" becomes /@/, "to"-/t@/, "confidence"-/kQnf@d@ns/, and "protecting"-/pr@tektIN/ and usual **elisions** where "and" becomes /n/, "comes his"-/kVmz Iz/, "are"-/@/, "more"-/mO:/, and "in" becomes /n/. Examples of **geminates** are "contrast to"-/kQntrast@/ (geminate /t/) and "with the"-/wID@/ (geminate /D/). More special elisions are "from" labelled /rm/ and "today" labelled /tdeI/.
That is, *for elisions, reductions, and geminates an acoustic-phonetic approach is followed, whereas for the voicing feature a phonemic approach is selected.*

Below the points are described and suggested changes of some of them are given (note that the alternatives at 12, 62 and 71 will lead to 3 new sentences). Afterwards some less serious points such as voicing and frication problems are listed.

11 Is the half-way convention used?
Suggestion: put the end boundary of /r/ where the F_3 appears in the spectrogram.
12 An epenthetic silence of 30ms is included in the /s/-segment.
Suggestion: mark the epenthetic silence with /.../

22 The /t/ has no closure phase and sounds as a [s] both in isolation and in context, and it appears as a [s] in the waveform and spectrogram.
Suggestion: label this segment with a /s/.
23 The segment labelled with a /@/ is perceived as [@] followed by [v] when listening to this portion in isolation and in context.
Suggestion: the /@/-segment should be divided into a /@/ and /v/ segment at the intensity drop where F_2 and F_3 disappear (which is seen in the spectrogram at time instance 4.400 sec.).
24 The end boundary of /s/ is placed too late, that is, after the closure phase of /t/ has begun and no amplitude is seen in the waveform (it seems as if the boundary decision is based on the spectrogram only, which is unnecessary for the fricative-unvoiced plosive transition).
26 An intensity drop is clearly visible in the spectrogram at 5.000 sec., but the boundary between /{/ and /n/ is marked at 4.983 sec. where no acoustic cues are seen, i.e. a mismatch of 17 ms.
Suggestion: Move the end boundary of /{/ to the right.
27 Where does the /S/ come from? It is not heard, it is not expected and no acoustic cues for it are seen. Despite all this, the /S/ is marked as a segment of 9 ms duration taken as the start of the closure part of the succeeding /t/.
Suggestion: remove the /S/ and include the segment in /t/.

32 The /r/-segment should contain the formant-transition in the beginning of the following /l/ according to the convention for segmenting semivowels.
33 The /k/-segment does not contain a burst, which can be seen in the spectrogram before the /s/-frication.
34 The /t/-segment contains a voiced waveform with decreasing amplitude and the /dZ/-segment contains the burst.
Suggestion: Remove the /t/-segment.

41 A period of creaky voice after /V/ is included in the /t/.
Note: This is an example of a double articulation which is typical in English. The /t/ is preceded and accompanied by a glottal stop, where the glottal and alveolar closure are probably released simultaneously. The area of creaky voice after the vowel is assigned to the closure part of the plosive by the human labeller.

Suggestion: The voice excitation pulses are voiced and have a vowel quality. It is thus more natural to include a period of creak in the vowel instead of in the plosive.

51 The /r/ segment is marked off with start point in the /t/-aspiration and end point when the voicing of /A:/ begins.
(Our automatic segmentation program included all the voiceless aspiration in /t/, and /r/ was the transition phase of the /A:/ (the end boundary was placed too late compared to the half-way convention)).
Suggestion: Mark start /r/ where the voicing starts and mark end according to the half-way convention.

61 Although no cues for splitting the nasal segment are seen in the waveform or the spectrogram, the /n/ is segmented long (102ms) and /m/ short (17ms).
Suggestion: divide the nasal segment containing /n/ and /m/ into two parts of equal duration.
62 Epenthetic silence after t-burst is included in /t/. (It is probably a glottal stop).
Suggestion: Mark the silence with a pause segment /.../.
63,64 The /@/ is the voiced part, the /t/-segment contains the period of creaky voice, and the /k/ contains the silent closure part and the burst.
Suggestion: Let /@/ include the voiced part and the creaky voice area, /t/ the silent closure part, and /k/ the burst (i.e. move 2 boundaries).
66 The /w/-/eI/ boundary is placed when the F_2 begins its glide or approximately when F_3 is visible in the spectrogram. At this transition it must be difficult to apply the half-way convention because the F_3 has almost a constant value.
Suggestion: Move the boundary to the right to when the F_2 glide has finished.

71 Before the frication of [f] begins, a 40ms pause/silence is included in the /f/-segment.
Suggestion: mark the epenthetic silence with the pause symboi, i.e. /.../.
73 The /z/-segment is the first part of a voiceless area. The /r/ begins in the unvoiced /z/ and ends after the beginning of the voiced [m].
Suggestion: /z/=the voiceless area, /r/= only the beginning of the voiced [m], (i.e. move only the end boundary of /z/).
74 A short /p/ (16ms) contains the closure only, i.e. all the aspiration/frication is included in the /f/. Since /p/ and /f/ are almost homorganic phonemes this segmentation strategy is reasonable.
Suggestion: Include a short burst, e.g. 5ms, in the /p/-segment.
76 The /r/-segment starts in the /t/-aspiration (unvoiced) and ends when voicing of /V/ begins.
Suggestion: Move the end boundary of /r/ to the right to include some voicing.
77 No burst in the /d/-segment.
Suggestion: Include a small burst segment, e.g. 5ms, in /d/.

80 The /t/ is marked off as the segment containing a small random amplitude after /@U/, with the end of /t/ where this amplitude is finished. A small pause (5ms) and the /t/-burst is included in the /s/-segment.
Suggestion: Move end /t/ from 20.476 sec. to 20.489 sec.
82 The /r/-segment is very short and no acoustic cues for placing segment boundaries is seen in the waveform or spectrogram.
Suggestion: Remove the /r/ if it is not perceived by a native speaker of English.

94 No burst is included in the /k/-segment.
Suggestion: Include a burst in the /k/-segment.

Less serious problems:

Some doubtful boundary placements:

-At **22** almost one pitch period of voicing is included in the (otherwise) voiceless segment.
-At **25** the end boundary of /@/ is perhaps marked too early because amplitude is found in the closure part of the following /t/.
-At **37** the non-released /k/ is marked off as the first part of the closure phase where aperiodic amplitudes are seen in the waveform, and /b/ is the silent closure part and a short burst. This is however, done throughout, see e.g. **92**.
-At **81** and **91** the /r/-segment starts in the aspiration of the voiceless plosive and ends after the voicing of the vowel has begun.
-At **93** the /dZ/-segment contains only the burst frication and the short closure seen in the waveform is included in the /n/-segment.

Some doubtful labelling decisions:

-Some completely voiceless segments are given a voiced phoneme label, e.g. with /z/ as in **21, 36, 65,** and **73**, or with /D/ in **72**, /d/ in **75**. The /st/-segment looks like two /s/ with a brief pause between them, and /v/ in **78**.
-At **31** the segments labelled /m/ and /r/ are perceived as [m] when listened to in isolation and in context, and there are no visible cues indicating a boundary between /m/ and /r/ in the spectrogram or waveform. Hence, the /m/ and /r/ should be joined into one segment and labelled /m/ only.
-At **35** the /e/ should (perhaps) be divided into /j/ and /e/ because both sounds are perceived. The boundary can be marked where an abrupt change in the formant transition is seen in the spectrogram.
-At **38,39** a reduced /I/ could be labelled /@/.
-At **74** an r-sound is perceived when listened to in context.

Summary of observations:

-The **epenthetic silences** in **12**, **62** (possibly a glottal stop), and **71** all occur between a voiced and a voiceless sound, and the epenthetic silence is then included in the voiceless sound.

-The **glottalisation** or **creaky voice** between a vowel and a plosive at **41** and **63**, is included in the following plosive, not in the vowel.

-When no clear **plosive burst** can be isolated in the waveform or spectrogram, the end of the plosive is marked at the beginning of the next sound (i.e. where an amplitude is seen in waveform). That is, no burst is included in the plosive segment. When there are **two succeeding plosives with no audible release of the first one**, the general rule that a plosive or affricate starts at the beginning of the closure and ends where the aspiration (friction) ends has to be modified. At all occurrences some random amplitude is assigned to the first plosive, and the rest

of closure and burst is assigned to the last plosive; **34**,**36**,**64**, and **75**.

-A special problem in English is segmenting the semi-vowels /w/ and /j/ and some realisations of /r/. In the guidelines, [Barry'90 b, p.4], a **half-way convention for semi-vowels** of marking the half-way point in the F_3 glide between its minimum and maximum values is described. It is claimed that this half-way point will correspond "roughly to the half-way point in the accompanying change in the speech pressure waveform" [Barry'90 b]. This seems reasonable, but it often implies that the boundary is placed before F_3 and F_4 appear in the spectrogram. At point **11** for instance, a [w] in the /e/ segment is perceived with the given boundary placing. However, at points **11**, /...w/, **32** /mrI/, **51** /trA:/, and **66** /zweI/, the end boundary for /w/ or /r/ is placed too early with respect to the half-way convention.
To alleviate this, we suggest marking the boundary when the third formant appears in the spectrogram (as they also have done in /krA:ft/ in sentence 1). This approach is also consistent with the other conventions which use intensity changes in the spectrogram as boundary clues.

-Since this is a phonemic labelling, "stops and fricatives occur, either wholly or in part, with phonetic voicing properties that are opposite of the phonemically defined category. Thus, /t/ may have a (partially) voiced closure and /d/ incomplete or no closure voicing" [Barry'90b, p.5].

-The problem of segmenting a (devoiced) /r/ after an voiceless plosive is in the Test Passage solved in three different ways:
i) /r/ is *voiced*, and defined as the transition part in the beginning of the following vowel.
ii) /r/ is *voiceless*, and defined as the last part of the plosive aspiration.
iii) /r/ is *defined as half devoiced and half voiced.*
Hence, the /r/ is voiced, e.g **13**, and voiceless, e.g. **51** and **76**, and is sometimes segmented to include both a voiceless and voiced portion, e.g. **81** and **91**.
Since the manual segmentation has not applied any convention consistently, it will be impossible for an automatic procedure to place the boundaries similarly.
(In order to make the manual segmentation consistent, we would prefer alternative i)).

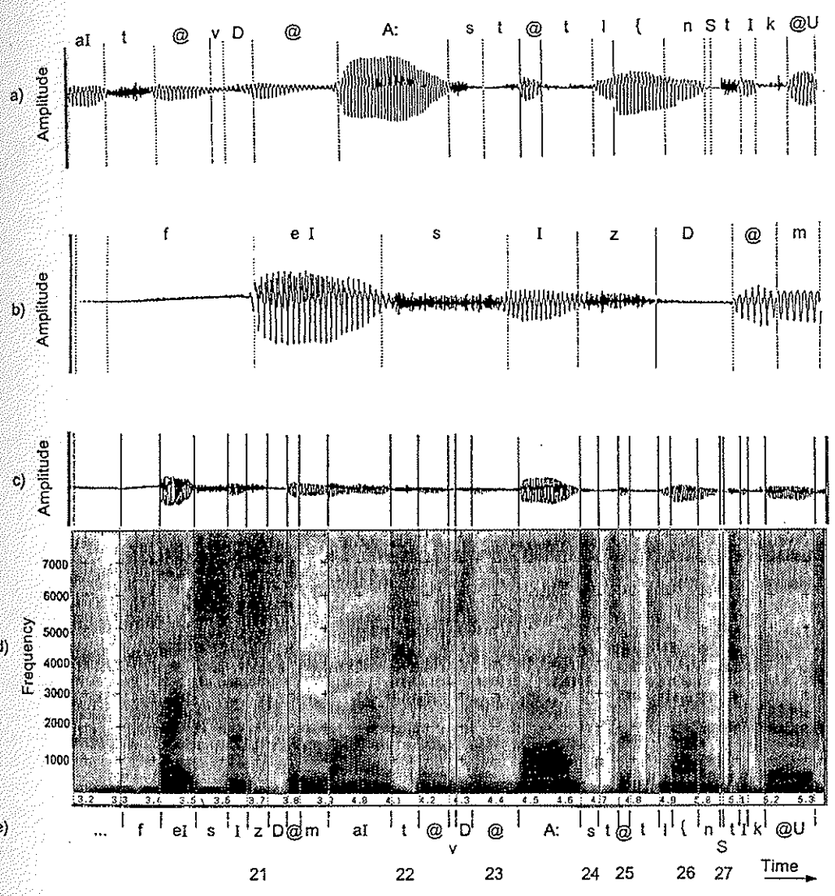


Figure C.3 *The second sentence from the Test Passage (female speaker: EAE).*
a) Waveform zoomed in for the last part of the sentence.
b) Waveform zoomed in for the first part of the sentence.
c) Waveform synchronised with the spectrogram.
d) Wide band spectrogram with segment boundaries.
e) The phonemic label string. The numbers refer to the discussion in the text.

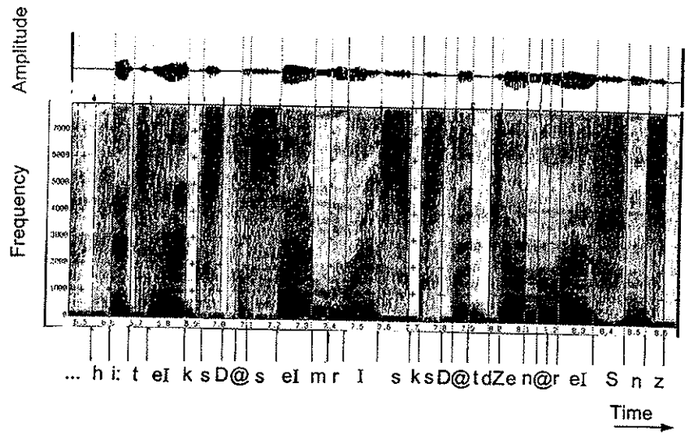


Figure C.4 *A part of the third sentence from the Test Passage (i.e. speaker EAE); "he takes the same risks that generations" /hi: teIks D@ seIm rIsks D@t dZen@reISnz/.*

Changes discussed in section 7.3.3

Table C.3 below shows the number of concidences within the deviation thresholds ±5ms, ±10ms, ±15ms, ±20ms, ±25ms, and ±40ms, and the number of gross errors and the mean deviation for the tests with the different alternative annotations.

Firstly, *1.corr.*, the three epenthetic silences at 12, 62, and 71 were segmented and labelled with /.../ (i.e. three sentences were divided).
When comparing manual and automatic segmentation with respect to the epenthetic silences, we found at 12 that the manual segmentation included the silence in the /s/-segment. The ASS assigned /s/ to the frication only and segmented the following epenthetic silence as a segment on its own. This segment was forced to be labelled /m/. However, ASS got only two segmentation errors at this point. At 62 the boundaries of manual and automatic segmentation coincided. At 71 the 40ms silence was included in the beginning of the /f/-segment, whereas ASS includes the silence in /n/, i.e. one segmentation error.

The effect of segmenting the epenthetic silences separately was hence minimal: at 12 the first new sentence was segmented as before but in the second part a smaller deviation at /O:/-/l/, and bigger at /l/-k/ was observed. In sentence 6 the same error-pattern was seen, but small changes in the boundary placements implied that 2 small gross-errors disappeared. At 71 the ASS segmented as before, except at the /p/-/f/ transition where the deviation had increased from about -30ms to about -90ms, i.e. the results were worse.

Secondly, *2. corr.*, we made changes at the points: 11,23,26,27,32,33,34,41,61,63,64,73,74,80, and 82 according to the suggestions in this Appendix, i.e. 11 boundaries were moved, 4 boundaries deleted, and 1 boundary inserted.

Thirdly, *3. corr.*, we substituted labels that were more phonetically correct in addition to the second correction, i.e. additional changes at points: 21,22,36,38,39,65,72,73[2],75, and 78, i.e. 10 new labels.

Finally, *4. corr.*, the epenthetic silences were segmented explicitly (i.e. as in the 1. correction), and some plosive to /r/ transitions and some half-way transitions were changed in addition to the second and third correction, i.e. additional changes at points: 12,24,51,62,66,71,76,77, and 94. That is, 3 boundaries were inserted and 8 extra boundaries were moved.
This last correction was also tested with full HMM's, both with 150% and 100% o.s. called *4b* and *4c* respectively in table C.3.

With the original annotation of the Test Passage, 5 gross errors were only 1ms displaced compared to the corresponding manual segmentation and 2 gross errors were 2ms displaced. Thus, only small changes in the boundary placements may remove some gross errors.

	Original	1.corr	2.corr	3.corr	4.corr	4b	4c
< 5 ms	106	108	111	113	121	130	125
< 10 ms	133	135	140	143	151	160	156
< 15 ms	158	161	164	169	174	181	182
< 20 ms	172	175	178	184	187	193	191
< 25 ms	181	183	186	191	194	200	200
< 40 ms	200	202	204	207	213	216	216
> 40 ms	44	45	38	35	32	29	29
Nos. gross err	14	12	12	11	10	5	6
mean dev. (ms)	19.6	19.4	17.9	16.6	15.5	14.5	14.7
total no. of transitions	244	247	242	242	245	245	245

Table C.3 *Segmentation performances for different manual annotation alternatives for the Test Passage. See text.*

The first alternative that simply divided three sentences did not increase the accuracy of the automatic segmentation. Accordingly the ASS procedure was not sensitive to ripple effects.

The second correction improved the performance, especially at small deviations. The only difference between second and third correction was that in the third one 10 labels were substituted with *phonetically* more correct labels (all boundaries were fixed). The acoustic segmentation was thus the same, but the segmentation results with the third correction label file were better than that of the second correction, e.g. with 6 additional boundaries within the ±20ms deviation threshold and with 1 gross-error less.

[2] Note that point 73 covers two operations, both the end boundary of /z/ (changed in the second correction) and the label /z/ changed to /s/ (in the third correction).

The last correction gave segmentation accuracy similar to the ones obtained with the original Test Passage label file with full HMMs. By using the label file of the last correction instead of the original label file for the Test Passage, 15 more boundaries were detected within ±20ms deviation from the label file (i.e. an improvement of 6%).

Simulation 4b used the label file from the last correction, but now the HMMs trained on the whole EUROM0 were applied (full HMM). This increased the coincidence rate further by 2-3%.

At small deviations 4b gave comparable accuracy to the EUROM0 TI-test results, but when bigger deviations were allowed for, the TI-test was clearly best. Table C.3 shows a reduction in the number of gross-errors. Five (or 2.1%) gross-errors are comparable to the TI-test for EUROM0 which had 1.6% gross-errors.

Reducing the oversegmentation to 100%, i.e. simulation 4c, reduced the accuracy at small deviation with less than 2% and gave the same results as in 4b at ±25 ms deviation (although the degradation was at about 7% for small deviation and 2% for bigger deviation in the acoustic segmentation).

C.4 ANALYSIS OF THE ENGLISH EUROM0 RECORDINGS

The English EUROM0 recording was "transcribed by two trained and experienced phoneticians, hand-labelled by SAM research assistants, checked by two further project workers and counter-checked by the authors" (B.Barry and A.Fourcin) [Barry'91a,p.13].
The annotation of the English EUROM0 is believed to be fairly similar to that of the Test Passage described in C.3. Hence, only a summary of the speech data is given. Speaker JWE and MBE are males, and all are in the mid forties. Below the tables C.4 and C.5 correspond to the Norwegian analyses in tables B.1 and B.2 respectively:

Speaker	Recording time [sec.]	Speech time [sec.]	Pause time [sec.]	Number of Segments	Number of Pauses	Number of Phonemes
EAE	117.7	94.4	23.3	1169	48	1121
JHE	112.2	83.6	28.6	1160	51	1109
JWE	109.9	92.3	17.6	1149	33	1116
MBE	113.4	94.8	18.6	1177	58	1119
Total	453.2	365.1	88.1	4655	190	4465

Table C.4 *The recording, speech and pause time and the number of segments, pauses and phonemes for the four readers based on the manual annotation of the English EUROM0 continuous passage.*

Speakers	Articulation rate	Speaking rate	Average phoneme duration [ms]	Max. phoneme duration [ms]	Min. phoneme duration [ms]
EAE	11.9	9.52	84	376	3
JHE	13.3	9.88	75	339	3
JWE	12.1	10.15	83	1059	6
MBE	11.8	9.87	85	385	7
All	12.2	9.85	82		

Table C.5 *Average articulation and speaking rate, average phoneme duration and maximum and minimum duration for the English EUROM0 continuous passage.*

Although the number of phonemes was nearly equal in the Norwegian and English recordings, the English recording time was over 2 minutes shorter. This is because the English speakers had less pauses, longer sentences and higher speaking and articulation rates.
The total Norwegian pause time is 1.7 times the English, but the speech time is only 1.2 times longer.

We found extremely short and extremely long duration of some of the English phonemes (the plosives were only 3-4 ms for some speakers, and one /v/ is marked as 1.059 seconds for speaker JWE). Due to our annotation philosophy, no phone should be marked shorter than at least one pitch period. This would exclude the very short segments in the English material.

One may ask which criteria are applied to mark such short phoneme segments and how is an automatic segmentation algorithm supposed to locate such small segments? Such manual segmentation is a source for gross errors when evaluating automatically placed segment boundaries.

The English TI-test

The division of the speech corpus for the English EUROM0 Text Independency (TI) test was as follows:

Speaker	Part 1 (samples / utterances)	Part 2 (samples / utterances)
EAE	102411-1035254 / 0-28	1054277-1987654 / 29-48
JHE	105861-1039941 / 0-29	1069113-1902805 / 30-51
JWE	93885-982117 / 0-16	1006612-1854683 / 17-33
MBE	84340-971268 / 0-32	983671-1900995 / 33-58

Table C.6 *Division of the English EUROM0 passage for the TI-test yielding 2362 segments to part 1 and 2281 segments to part 2.*

Since one half of the text is used for training and the other half for test, some phonemes /e@ A: Z u/ occurred only in one of the sets. These phonemes had to be trained on the test set. Since these phonemes constituted only 29 out of 4465 phonemes (i.e. 0.6%), this "trick" was believed to not influence on the segmentation performance.

C.5 ANALYSIS OF THE ITALIAN EUROM0 RECORDINGS

The Italian EUROM0 recording was annotated by two "phonetically trained human labellers by listening to the speech signal and by visually examining displays of signal related time and frequency parameters" [Cosi'90]. Some examples of the Italian annotation are discussed in section 4.1.2.
Below the tables C.7 and C.8 correspond to the Norwegian analyses in tables B.1 and B.2 respectively:

Speaker	Recording time [sec.]	Speech time [sec.]	Pause time [sec.]	Number of Segments	Number of Pauses	Number of Phonemes
GCI	122.4	101.5	20.9	1338	45	1293
LCI	117.9	104.9	13.0	1341	49	1292
LVI	109.6	96.5	13.1	1300	26	1274
PUI	102.8	83.7	19.1	1294	28	1266
Total	452.7	386.6	66.1	5273	148	5125

Table C.7 *The recording, speech and pause time and the number of segments, pauses and phonemes for the four readers based on the manual annotation of the Italian EUROM0 continuous passage.*

Speakers	Articulation rate	Speaking rate	Average phoneme duration [ms]	Max. phoneme duration [ms]	Min. phoneme duration [ms]
GCI	12.7	10.6	78.5	230	8
LCI	12.3	11.0	81.2	272	7
LVI	13.2	11.6	75.7	277	6
PUI	15.1	12.3	66.1	291	9
All	13.1	11.3	75.4		

Table C.8 *Average articulation and speaking rate, average phoneme duration and maximum and minimum duration for the Italian EUROM0 continuous passage.*

The analyses of the EUROM0 recordings in Appendix B and C, show that the Norwegian speakers had the lowest articulation rates, and even the fastest Norwegian speaker, TGN, had lower articulation and speaking rate than all the Danish and Italian speakers.

Appendix D

COMPLEMENTARY RESULTS

In this appendix detailed results for the experiments in chapter 7 are provided. Firstly, in section D.1 to D.3, results of the acoustic segmentation experiments discussed in section 7.2 are provided. Then, in section D.4 to D.7, the performances of the phonemic segmentation with the optimal parameter set defined in section 7.4.1 are given for the English, Norwegian, Danish, and Italian EUROM0 recordings respectively. Section D.8 provides the segmentation results when using cepstral domain filtering, and in section D.9 the segmentation performances on Norwegian using enlarged training set is given.
In section D.10 some phoneme recognition results are given.

For all the phonemic segmentation results, V means pure HMM segmentation and VC means constrained HMM segmentation. VC_150 means that 150% oversegmentation is used in the acoustic segmentation which constrains the Viterbi recursion. In the tables providing results as confidence intervals, the gross error results are given as 95% confidence intervals for the number of gross errors, whereas the fine errors are given as 95% confidence intervals for the coincidence rates within each deviation threshold.

D.1 ACOUSTIC SEGMENTATION ON NORWEGIAN EUROM0

In this section the acoustic segmentation performance with different oversegmentation factors on the Norwegian EUROM0 recording are given, see section 7.2.4.

	Deviation from manual broad phonetic segmentation				
Speakers	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	62.81	86.06	94.21	97.64	98.43
SHN	68.62	92.01	97.69	99.23	99.81
TBN	67.34	91.35	97.60	99.04	99.42
TGN	71.40	92.47	97.74	98.87	99.81
Average	67.54	90.47	96.82	98.70	99.37

Table D.1 *DTW-coincidence in percent for the acoustic segmentation with 200% o.s. for each speaker in the Norwegian EUROM0.*

Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	60.26	83.81	93.03	96.76	98.14
SHN	66.03	90.38	96.73	98.65	99.52
TBN	66.19	90.20	96.83	98.94	99.23
TGN	68.30	90.78	96.99	98.31	99.44
Average	65.20	88.80	95.90	98.17	99.08

Table D.2 *DTW-coincidence in percent for the acoustic segmentation with 175% o.s. for each speaker in the Norwegian EUROM0.*

Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	55.35	79.10	89.60	94.11	95.78
SHN	60.54	84.31	92.01	95.77	97.88
TBN	62.54	86.46	94.91	97.69	98.75
TGN	64.16	87.58	94.26	96.99	98.49
Average	60.65	84.36	92.70	96.14	97.73

Table D.3 *DTW-coincidence in percent for the acoustic segmentation with 125% o.s. for each speaker in the Norwegian EUROM0.*

Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	50.34	73.31	84.00	88.62	90.87
SHN	52.94	76.52	86.43	91.34	94.23
TBN	54.37	79.35	89.15	92.70	95.00
TGN	55.79	79.68	88.24	92.66	94.83
Average	53.38	77.22	86.96	91.33	93.73

Table D.4 *DTW-coincidence in percent for the acoustic segmentation with 75% o.s. for each speaker in the Norwegian EUROM0.*

Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	47.01	68.20	78.70	83.51	85.77
SHN	48.70	72.28	82.10	86.81	91.15
TBN	49.18	74.54	85.11	88.57	92.12
TGN	50.81	73.22	81.96	86.32	89.65
Average	48.93	72.06	81.97	86.30	89.67

Table D.5 *DTW-coincidence in percent for the acoustic segmentation with 50% o.s. for each speaker in the Norwegian EUROM0.*

Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	39.74	57.61	65.95	71.05	74.48
SHN	40.71	59.58	69.39	75.17	78.83
TBN	40.06	61.48	71.95	76.37	80.31
TGN	43.27	61.24	69.99	74.98	78.08
Average	40.95	59.98	69.32	74.39	77.92

Table D.6 *DTW-coincidence in percent for the acoustic segmentation with 25% o.s. for each speaker in the Norwegian EUROM0.*

Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	21.30	30.13	35.03	38.08	40.82
SHN	19.92	29.16	34.36	37.63	39.36
TBN	19.69	28.53	33.24	35.93	38.14
TGN	17.38	24.88	29.06	31.62	34.47
Average	19.57	28.18	32.92	35.82	38.20

Table D.7 *DTW-coincidence in percent for the acoustic segmentation with 0% o.s. for each speaker in the Norwegian EUROM0.*

D.2 UNIFORM SEGMENTATION ON NORWEGIAN EUROM0

In this section the uniform segmentation performance with 100% oversegmentation and no oversegmentation on the Norwegian EUROM0 recording are given, see section 7.2.4.1.

Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	17.66	36.02	53.78	69.19	84.69
SHN	23.10	46.01	70.55	89.32	95.86
TBN	24.40	48.70	69.36	89.63	97.02
TGN	25.12	49.58	75.07	93.41	98.21
Average	22.57	45.08	67.19	85.39	93.95

Table D.8 *DTW-coincidence in percent for the uniform segmentation with 100% o.s. for each speaker in the Norwegian EUROM0.*

Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
AFN	4.91	9.03	13.35	16.98	22.28
SHN	6.35	11.55	17.81	22.04	27.33
TBN	4.51	10.37	14.70	20.65	23.92
TGN	4.61	8.47	12.98	15.80	18.63
Average	5.10	9.86	14.71	18.87	23.04

Table D.9 *DTW-coincidence in percent for the uniform segmentation with 0% o.s. for each speaker in the Norwegian EUROM0.*

D.3 COMPARING ACOUSTIC SEGMENTATION AT 100% OVERSEGMENTATION ACROSS LANGUAGES

In this section the acoustic segmentation performance with 100% oversegmentation on the English, Italian, and Danish EUROM0 recording are given, see section 7.2.5.

English Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
EAE	62.69	84.70	90.86	94.59	95.71
JHE	63.20	85.62	91.96	94.42	96.59
JWE	61.46	83.73	91.31	93.90	95.47
MBE	63.02	83.87	90.57	94.06	96.04
Average	62.59	84.48	91.18	94.24	95.95

Table D.10 *DTW-coincidence in percent for the acoustic segmentation with 100% o.s. for each speaker in the English EUROM0.*

Italian Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
GCI	49.88	75.26	88.60	94.18	96.77
LCI	46.80	73.07	86.05	92.62	94.89
LVI	56.27	82.46	91.59	95.80	97.49
PUI	53.88	80.24	90.86	95.35	97.39
Average	51.70	77.76	89.28	94.49	96.64

Table D.11 *DTW-coincidence in percent for the acoustic segmentation with 100% o.s. for each speaker in the Italian EUROM0.*

Danish Speakers	Deviation from manual broad phonetic segmentation				
	< 5 ms	< 10 ms	< 15 ms	< 20 ms	< 25 ms
BLD	58.03	83.55	90.78	94.36	96.18
CHD	62.59	85.62	92.60	95.34	96.76
JDD	63.61	84.97	90.66	94.30	96.04
LFD	55.58	81.42	88.83	93.32	96.09
Average	59.95	83.89	90.72	94.33	96.28

Table D.12 *DTW-coincidence in percent for the acoustic segmentation with 100% o.s. for each speaker in the Danish EUROM0.*

D.4 PHONEMIC SEGMENTATION ON ENGLISH EUROM0

This section provides detailed results for the phonemic segmentation of the English EUROM0 recording using the optimal parameter set defined in section 7.4.1.

English	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	81	2038	2891	3672	3801
SI, VC_150	64	2226	2955	3628	3773
SI, VC_100	76	2212	2952	3586	3740
TI, V	93	2011	2900	3661	3796
TI, VC_150	61	2209	2945	3623	3770
Full HMM, V	48	2166	3070	3830	3939
Full HMM, VC_150	27	2329	3080	3771	3909

Table D.13 *Number of gross errors and number of coincidences within different deviation thresholds for the SI, TI, and Full HMM-tests on English EUROM0.*

The corresponding 95% confidence intervals are:

English	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	1.5 - 2.3 %	45.2 - 49.2 %	66.3 - 69.1 %	84.9 - 87.0 %	88.0 - 89.9 %
SI, VC_150	1.1 - 1.8 %	50.6 - 53.6 %	67.8 - 70.6 %	83.8 - 86.0 %	87.3 - 89.3 %
SI, VC_100	1.4 - 2.1 %	50.3 - 53.3 %	67.7 - 70.5 %	82.8 - 85.0 %	86.5 - 88.5 %
TI, V	1.7 - 2.5 %	45.6 - 48.6 %	66.5 - 69.3 %	84.6 - 86.7 %	87.9 - 89.8 %
TI, VC_150	1.1 - 1.8 %	50.2 - 53.2 %	67.5 - 70.3 %	83.7 - 85.9 %	87.3 - 89.2 %
Full HMM, V	0.8 - 1.4 %	49.2 - 52.2 %	70.5 - 73.2 %	88.7 - 90.6 %	91.4 - 93.0 %
Full HMM, VC_150	0.4 - 0.9 %	53.0 - 56.0 %	70.8 - 73.4 %	87.3 - 89.2 %	90.7 - 92.3 %

Table D.14 *95% confidence intervals for gross errors and fine errors for the SI, TI, and Full HMM-tests on English EUROM0.*

The English SI-test results for each speaker:

English	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
EAE, V	1.2 - 2.7 %	44.8 - 50.8 %	66.1 - 71.7 %	83.5 - 87.7 %	86.0 - 89.9 %
EAE, VC_150	0.8 - 2.1 %	49.9 - 55.9 %	68.7 - 74.1 %	82.6 - 86.9 %	86.1 - 90.0 %
JHE, V	1.1 - 2.6 %	40.4 - 46.4 %	60.1 - 65.9 %	83.9 - 88.1 %	87.9 - 91.6 %
JHE, VC_150	1.0 - 2.5 %	47.2 - 53.2 %	63.4 - 69.0 %	81.2 - 85.7 %	86.0 - 89.9 %
JWE, V	1.8 - 3.7 %	43.9 - 49.8 %	64.9 - 70.5 %	82.6 - 88.9 %	85.3 - 89.3 %
JWE, VC_150	1.6 - 3.4 %	47.0 - 52.9 %	64.8 - 70.4 %	82.7 - 87.0 %	85.7 - 89.6 %
MBE, V	0.7 - 2.0 %	50.6 - 56.6 %	68.8 - 74.3 %	85.7 - 89.7 %	89.4 - 92.8 %
MBE, VC_150	0.3 - 1.1 %	52.6 - 58.6 %	68.7 - 74.1 %	84.2 - 88.4 %	87.4 - 91.1 %

Table D.15 *95% confidence interval for the gross errors and fine errors for the individual speakers in the SI-test on English EUROM0.*

D.5 PHONEMIC SEGMENTATION ON NORWEGIAN EUROM0

Phonemic segmentation on Norwegian EUROM0 using the optimal parameter set defined in section 7.4.1. See section 7.4.4 for details.

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	72	2018	2831	3550	3670
SI, VC_150	63	1964	2731	3512	3663
SI, s+l, V	66	2009	2817	3527	3656
SI, s+l, VC_150	51	1994	2774	3525	3665
TI, V	45	2105	2943	3663	3782
TI, VC_150	27	2076	2865	3652	3777
Full HMM, V	20	2155	3030	3747	3867
Full HMM, VC_150	20	2095	2902	3685	3807

Table D.16 *Number of gross errors and number of coincidences within different deviation thresholds for the SI, TI, and Full HMM-tests on Norwegian EUROM0. For the SI-test the results with short and long vowels grouped (SI, s+l) are given.*

The corresponding 95% confidence intervals are:

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	1.3 - 2.0 %	47.0 - 50.1 %	66.7 - 69.5 %	84.3 - 86.4 %	87.3 - 89.2 %
SI, VC_150	1.1 - 1.8 %	45.7 - 48.8 %	64.2 - 67.1 %	83.3 - 85.5 %	87.1 - 89.0 %
SI, s+l, V	1.2 - 1.9 %	46.8 - 49.8 %	66.3 - 69.2 %	83.7 - 85.9 %	86.9 - 88.9 %
SI, s+l, VC_150	0.9 - 1.5 %	46.4 - 49.5 %	65.3 - 68.1 %	83.7 - 85.8 %	87.1 - 89.1 %
TI, V	0.8 - 1.3 %	49.1 - 52.1 %	69.4 - 72.1 %	87.1 - 89.0 %	90.0 - 91.8 %
TI, VC_150	0.4 - 0.9 %	48.4 - 51.4 %	67.5 - 70.3 %	86.8 - 88.8 %	89.9 - 91.7 %
Full, V	0.3 - 0.7 %	50.3 - 53.3 %	71.5 - 74.2 %	89.2 - 91.0 %	92.2 - 93.7 %
Full, VC_150	0.3 - 0.7 %	48.9 - 51.9 %	68.4 - 71.2 %	87.6 - 89.6 %	90.7 - 92.4 %

Table D.17 *95% confidence intervals for gross errors and fine errors for the SI, TI, and Full HMM-tests on Norwegian EUROM0. For the SI-test the results with short and long vowels grouped (SI, s+l) are given.*

The Norwegian SI-test results for each speaker:

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
AFN, V	6	486	694	893	917
AFN, VC_150	4	461	668	866	901
SHN, V	16	505	689	886	926
SHN, VC_150	9	486	653	861	913
TBN, V	42	481	676	824	854
TBN, VC_150	36	479	661	850	883
TGN, V	8	546	772	947	973
TGN, VC_150	14	538	749	935	966

Table D.18 *Number of gross errors and number of coincidences within different deviation thresholds for the individual speakers in the SI-test on Norwegian EUROM0.*

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
AFN, V	0.3 - 1.1 %	44.7 - 50.9 %	65.3 - 71.0 %	85.7 - 89.7 %	88.2 - 91.8 %
AFN, VC_150	0.3 - 1.1 %	42.3 - 48.4 %	62.7 - 68.5 %	82.8 - 87.2 %	86.5 - 90.4 %
SHN, V	0.9 - 2.3 %	45.6 - 51.6 %	63.4 - 69.1 %	83.0 - 87.3 %	87.1 - 90.9 %
SHN, VC_150	0.5 - 1.5 %	43.8 - 49.8 %	59.9 - 65.7 %	80.5 - 85.0 %	85.8 - 89.7 %
TBN, V	2.8 - 5.1 %	43.2 - 49.2 %	62.0 - 67.8 %	76.6 - 81.5 %	79.6 - 84.2 %
TBN, VC_150	2.4 - 4.4 %	43.0 - 49.0 %	60.5 - 66.4 %	79.2 - 83.9 %	82.5 - 86.9 %
TGN, V	0.4 - 1.4 %	48.4 - 54.4 %	69.9 - 75.2 %	87.1 - 90.8 %	89.7 - 93.1 %
TGN, VC_150	0.8 - 2.1 %	47.6 - 53.6 %	67.7 - 73.1 %	85.9 - 89.8 %	89.0 - 92.5 %

Table D.19 *95% confidence intervals for the gross errors and fine errors for the individual speakers in the SI-test on Norwegian EUROM0.*

Table D.20 and D.21 show how acoustic segmentation affects the constrained HMM segmentation (correspond to table 7.22).

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	72	2018	2831	3550	3670
SI, VC_225	52	1964	2742	3535	3666
SI, VC_200	54	1969	2744	3538	3663
SI, VC_175	45	1976	2747	3545	3670
SI, VC_150	63	1964	2731	3512	3663
SI, VC_125	48	1980	2756	3514	3660
SI, VC_100	65	1946	2720	3507	3653
SI, VC_75	70	1893	2679	3455	3604
SI, VC_50	74	1800	2585	3345	3498
SI, VC_25	250	1524	2202	2912	3074
SI, VC_0	1262	815	1173	1489	1588

Table D.20 *Number gross errors and number of coincidences within different deviation thresholds for the SI-test on the Norwegian EUROM0 for the pure HMM segmentation (V) and constrained HMM segmentation (VC), using different oversegmentation factors in the acoustic segmentation.*

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	1.3 - 2.0%	47.0 - 50.1 %	66.7 - 69.5 %	84.3 - 86.4 %	87.3 - 89.2 %
SI, VC_225	0.9 - 1.6 %	45.7 - 48.8 %	64.5 - 67.4 %	83.9 - 86.1 %	87.2 - 89.1 %
SI, VC_200	0.9 - 1.6 %	45.8 - 48.9 %	64.5 - 67.4 %	84.0 - 86.1 %	87.1 - 89.0 %
SI, VC_175	0.8 - 1.3 %	46.0 - 49.0 %	64.6 - 67.5 %	84.1 - 86.3 %	87.3 - 89.2 %
SI, VC_150	1.1 - 1.8 %	45.7 - 48.8 %	64.2 - 67.1 %	83.3 - 85.2 %	87.1 - 89.0 %
SI, VC_125	0.8 - 1.4 %	46.1 - 49.1 %	64.8 - 67.7 %	83.4 - 85.6 %	87.0 - 89.0 %
SI, VC_100	1.1 - 1.8 %	45.3 - 48.3 %	64.0 - 66.8 %	83.2 - 85.4 %	86.8 - 88.8 %
SI, VC_75	1.2 - 2.0 %	44.0 - 47.0 %	63.0 - 65.9 %	81.9 - 84.2 %	85.6 - 87.7 %
SI, VC_50	1.3 - 2.1 %	41.8 - 44.8 %	60.7 - 63.6 %	79.2 - 81.6 %	83.0 - 85.2 %
SI, VC_25	4.9 - 6.3 %	35.2 - 38.1 %	51.4 - 54.5 %	68.6 - 71.4 %	72.6 - 75.2 %
SI, VC_0	26.8 - 29.4 %	18.4 - 20.8 %	26.9 - 29.6 %	34.4 - 37.3 %	36.7 - 39.7 %

Table D.21 *95 % confidence intervals for the gross errors and fine errors for the SI-tests on Norwegian EUROM0 using pure HMM segmentation (V) and constrained HMM segmentation (VC) with different oversegmentation factors in the acoustic segmentation.*

Table D.22 and D.23 show the effect different types of acoustic segmentation had to the performance of the constrained HMM segmentation (correspond to table 7.23).

SI, VC_100	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Ordinary	65	1946	2720	3507	3653
Corrected	60	3352	3479	3718	3842
Uniform	21	794	1495	2893	3326

Table D.22 *Number gross errors and number of coincidences within different deviation thresholds for the constrained HMM segmentation (VC) of the SI-test on the Norwegian EUROM0, using different types of acoustic segmentation at 100% o.s..*

SI, VC_100	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Ordinary	1.1 - 1.8 %	45.3 - 48.3 %	64.0 - 66.8 %	83.2 - 85.4 %	86.8 - 88.8 %
Corrected	1.0 - 1.7 %	79.4 - 81.8 %	82.5 - 84.8 %	88.4 - 90.3 %	91.6 - 93.2 %
Uniform	0.3 - 0.7 %	17.9 - 20.3 %	34.5 - 37.4 %	68.2 - 71.0 %	78.7 - 81.2 %

Table D.23 *95% confidence intervals for the gross errors and fine errors for the constrained HMM segmentation (VC) of the SI-test on the Norwegian EUROM0, using different types of acoustic segmentation at 100% o.s.. (See section 7.4.4.3).*

D.6 PHONEMIC SEGMENTATION ON DANISH EUROM0

Phonemic segmentation on Danish EUROM0 using the optimal parameter set defined in section 7.4.1. See also section 7.4.3.

Danish	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	155	1945	2769	3496	3642
SI, VC_150	133	2057	2847	3519	3673
SI, VC_100	94	2065	2859	3552	3712

Table D.24 *Number of gross errors and number of coincidences within different deviation thresholds for the SI-tests on Danish EUROM0.*

Danish	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	2.9 - 3.9 %	42.7 - 45.7 %	61.5 - 64.3 %	78.2 - 80.6 %	81.6 - 83.9 %
SI, VC_150	2.5 - 3.4 %	45.3 - 48.2 %	63.3 - 66.1 %	78.8 - 81.1 %	82.4 - 84.5 %
SI, VC_100	1.7 - 2.5 %	45.5 - 48.4 %	63.6 - 66.4 %	79.5 - 81.9 %	83.3 - 85.4 %

Table D.25 *95% confidence intervals for the gross errors and fine errors for the SI-tests on Danish EUROM0.*

Danish	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
BLD, V	2.3 - 4.4 %	37.1 - 42.8 %	53.6 - 59.5 %	72.0 - 77.1 %	75.6 - 80.4 %
BLD, VC_150	1.3 - 3.0 %	39.2 - 45.0 %	57.6 - 63.4 %	72.4 - 77.5 %	76.2 - 81.0 %
CHD, V	1.8 - 3.6 %	43.9 - 49.9 %	62.5 - 68.2 %	78.6 - 83.3 %	82.0 - 86.4 %
CHD, VC_150	1.1 - 2.7 %	43.5 - 49.5 %	61.6 - 67.4 %	80.1 - 84.7 %	83.3 - 87.5 %
JDD, V	2.7 - 4.9 %	41.6 - 47.4 %	63.4 - 68.9 %	80.6 - 85.0 %	83.1 - 87.2 %
JDD, VC_150	1.6 - 3.3 %	48.3 - 54.1 %	67.1 - 72.4 %	80.4 - 84.8 %	83.3 - 87.4 %
LFD, V	3.1 - 5.4 %	42.5 - 48.4 %	60.6 - 66.3 %	76.6 - 81.4 %	81.0 - 85.4 %
LFD, VC_150	2.9 - 5.1 %	43.9 - 49.8 %	60.8 - 66.4 %	77.1 - 81.9 %	81.9 - 86.2 %

Table D.26 *95% confidence intervals for the gross errors and fine errors for the individual speakers in the SI-test on Danish EUROM0.*

D.7 PHONEMIC SEGMENTATION ON ITALIAN EUROM0

Phonemic segmentation on Italian EUROM0 using the optimal parameter set defined in section 7.4.1. See also section 7.4.3.

Italian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	165	2164	3081	4010	4224
SI, VC_150	104	2225	3095	4074	4278
SI, VC_100	109	2153	3035	4054	4254

Table D.27 *Number of gross errors and number of coincidences within different deviation thresholds for the SI-tests on Italian EUROM0.*

Italian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	2.8 - 3.7 %	42.1 - 44.9 %	60.6 - 63.3 %	79.5 - 81.7 %	83.9 - 85.9 %
SI, VC_150	1.7 - 2.5 %	43.4 - 46.1 %	60.9 - 63.6 %	80.8 - 83.0 %	85.0 - 87.0 %
SI, VC_100	1.8 - 2.6 %	41.9 - 44.7 %	59.7 - 62.4 %	80.4 - 82.6 %	84.5 - 86.5 %

Table D.28 *95% confidence intervals for the gross errors and fine errors for the SI-tests on Italian EUROM0.*

Italian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
GCI, V	4.5 - 7.0 %	38.6 - 44.1 %	54.3 - 59.8 %	71.8 - 76.7 %	76.7 - 81.2 %
GCI, VC_150	2.7 - 4.7 %	40.5 - 46.0 %	56.0 - 61.4 %	73.6 - 78.4 %	77.3 - 81.8 %
LCI, V	2.0 - 3.8 %	35.5 - 40.9 %	56.2 - 61.7 %	78.2 - 82.6 %	82.9 - 86.9 %
LCI, VC_150	0.8 - 2.0 %	35.9 - 41.3 %	55.7 - 61.1 %	79.7 - 84.0 %	85.2 - 88.9 %
LVI, V	1.3 - 2.8 %	44.0 - 49.6 %	64.4 - 69.6 %	82.1 - 86.1 %	86.1 - 89.7 %
LVI, VC_150	0.7 - 1.9 %	47.2 - 52.8 %	64.7 - 69.9 %	83.5 - 87.4 %	87.4 - 90.8 %
PUI, V	1.8 - 3.5 %	45.1 - 50.7 %	62.3 - 67.6 %	81.7 - 85.8 %	86.0 - 89.6 %
PUI, VC_150	1.5 - 3.1 %	44.5 - 50.0 %	61.9 - 67.3 %	82.3 - 86.3 %	86.4 - 90.0 %

Table D.29 *95% confidence intervals for the gross errors and fine errors for the individual speakers in the SI-test on Italian EUROM0.*

D.8 CEPSTRAL DOMAIN FILTERING

With RASTA_88 (i.e. optimal pole value) on Norwegian EUROM0, the following segmentations were obtained, see section 7.4.5:

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	41	2072	2955	3686	3802
SI, VC_150	34	2032	2833	3594	3740
AFN, V	7	520	745	908	928
AFN, VC_150	6	456	665	850	897
SHN, V	5	514	743	933	969
SHN, VC_150	3	527	726	915	948
TBN, V	16	517	729	904	937
TBN, VC_150	13	512	712	899	932
TGN, V	13	521	738	941	968
TGN, VC_150	12	537	730	930	963

Table D.30 *Number of gross errors and number of coincidences within different deviation thresholds for the individual speakers in the SI-test on Norwegian EUROM0.*

Norwegian	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
SI, V	0.7 - 1.2 %	48.3 - 51.4 %	69.7 - 72.4 %	87.6 - 89.6 %	90.5 - 92.3 %
SI, VC_150	0.5 - 1.1 %	47.4 - 50.4 %	66.7 - 69.5 %	85.4 - 87.4 %	89.0 - 90.8 %
AFN, V	0.3 - 1.2 %	48.1 - 54.2 %	70.5 - 75.9 %	87.2 - 91.0 %	89.4 - 92.8 %
AFN, VC_150	0.3 - 1.2 %	41.8 - 47.9 %	62.4 - 68.2 %	81.2 - 85.7 %	86.1 - 90.0 %
SHN, V	0.2 - 1.0 %	46.4 - 52.5 %	68.7 - 74.2 %	87.8 - 91.5 %	91.6 - 94.6 %
SHN, VC_150	0.1 - 0.7 %	47.7 - 53.7 %	67.0 - 72.8 %	86.0 - 89.9 %	89.4 - 92.8 %
TBN, V	0.9 - 2.3 %	46.6 - 52.7 %	67.2 - 72.7 %	84.7 - 88.8 %	88.0 - 91.7 %
TBN, VC_150	0.7 - 2.0 %	46.2 - 52.2 %	65.5 - 71.1 %	84.1 - 88.3 %	87.5 - 91.2 %
TGN, V	0.7 - 2.0 %	46.0 - 52.0 %	66.6 - 72.1 %	86.5 - 90.3 %	89.2 - 92.6 %
TGN, VC_150	0.6 - 1.8 %	47.5 - 53.5 %	65.8 - 71.4 %	85.4 - 89.3 %	88.7 - 92.2 %

Table D.31 *95% confidence intervals for the gross errors and fine errors for the individual speakers and the average in the SI-test on Norwegian EUROM0.*

COMPARING LANGUAGES WHEN APPLYING CEPSTRAL DOMAIN FILTERING

In tables D.32 and D.33 below the segmentation results when using optimal pole value in the RASTA-filter for each language are shown, see section 7.4.5.

	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
English, SI, V	45	2059	2934	3728	3857
English, SI, VC_150	49	2254	3002	3678	3832
Italian, SI, V	85	2250	3309	4274	4467
Italian, SI, VC_150	76	2297	3195	4201	4406
Danish, SI, V	108	2035	2909	3644	3800
Danish, SI, VC_150	102	2131	2913	3622	3773

Table D.32 *Gross errors and fine errors for the SI-tests on English, Italian, and Danish EUROM0 recordings where the cepstral coefficients were RASTA-filtered with optimal pole value for the given language.*

	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
English, SI, V	0.8 - 1-3 %	46.7 - 49.7 %	67.3 - 70.1 %	86.3 - 88.3 %	89.4 - 91.2 %
English, SI, VC_150	0.8 - 1.4 %	51.3 - 54.3 %	68.9 - 71.6 %	85.0 - 87.1 %	88.8 - 90.6 %
Italian, SI, V	1.3 - 2.0 %	43.9 - 46.6 %	65.2 - 67.8 %	85.0 - 86.9 %	89.0 - 90.6 %
Italian, SI, VC_150	1.2 - 1.9 %	44.8 - 47.6 %	62.9 - 65.6 %	83.4 - 85.5 %	87.7 - 89.5 %
Danish, SI, V	2.0 - 2.8 %	44.8 - 47.7 %	64.7 - 67.5 %	81.7 - 83.9 %	85.3 - 87.3 %
Danish, SI, VC_150	1.8 - 2.7 %	47.0 - 49.9 %	64.8 - 67.6 %	81.2 - 83.4 %	84.7 - 86.8 %

Table D.33 *95% confidence intervals for the gross errors and fine errors for the SI-tests on English, Italian, and Danish EUROM0 recordings where the cepstral coefficients were RASTA-filtered with optimal pole value for the given language.*

For a more detailed assessment of the constrained HMM segmentation (SI,VC) when using optimal pole in the RASTA-filter, table D.34 to D.41 below show how many boundaries that were placed within ±20ms deviation from the manually segmented boundaries for each phoneme class transition for the English, Norwegian, Italian, and Danish EUROM0 recordings.

English:

Eng.	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids	All
Plos	18/33	0/0	64/86	506/525	15/24	13/28	31/61	647/757
Affr	1/2	0/0	2/3	41/41	0/0	0/0	0/0	44/46
Fric	133/165	1/1	30/51	449/465	71/81	21/30	27/38	732/831
Vow	274/313	19/20	502/534	12/30	521/566	11/17	60/133	1399/1613
Nas	169/201	11/12	144/149	200/221	1/1	26/28	20/25	571/637
Glid	0/0	0/0	0/1	93/131	0/1	0/0	0/0	93/133
Liq	12/16	4/4	26/28	133/185	12/15	4/4	1/2	192/254
All	607/730	35/37	768/852	1434/1598	620/688	75/107	139/259	3678/4271

Table D.34 *Number of automatic phoneme boundary placements within ± 20 ms deviation from manual segmentation in relation to number of occurrences in the material for each phoneme class transition for the English SI,VC -test (150% o.s.).*

The corresponding 95% confidence intervals for the coincidence rates within ±20ms from the manual segmentation:

Eng.	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids
All/Class	80-86 %	84-97 %	88-92 %	88-91 %	88-92 %	61-78 %	48-60 %
Class/All	83-88 %	87-98 %	86-90 %	85-88 %	87-92 %	62-77 %	70-80 %

Table D.35 *Coincidence within ±20ms deviation given as 95% confidence intervals for phoneme class transitions for the English SI,VC -test (150% o.s.).*

Norwegian:

Nor.	Plosives	Fricatives	Vowels	Diphthongs	Nasals	Liquids	Retroflex	All
Plos.	8/16	52/55	489/503	23/24	33/33	96/108	0/0	701/739
Fric.	97/108	15/20	461/486	3/4	13/17	17/19	2/2	108/656
Vow.	321/374	393/418	37/106	3/4	440/497	211/294	62/66	1466/1759
Dipht.	27/28	0/0	0/0	0/0	0/0	4/4	0/0	31/32
Nas.	72/94	63/69	296/353	0/0	2/8	0/1	0/0	433/525
Liq.	36/48	34/38	203/268	0/0	18/26	0/0	0/0	291/380
Retr.	3/4	6/6	54/56	0/0	0/0	1/1	0/0	64/167
All	564/672	563/606	1540/1772	29/32	506/581	329/427	64/66	3594/4158

Table D.36 *Number of automatic phoneme boundary placements within ± 20 ms deviation from manual segmentation in relation to number of occurrences in the material for each phoneme class transition for the Norwegian SI,VC -test (150% o.s.).*

The corresponding 95% confidence intervals for the coincidence rates within ±20ms from the manual segmentation:

Nor.	Plosives	Fricatives	Vowels	Diphth.	Nasals	Liquids	Retroflex
All/Class	81-87 %	91-95 %	85-88 %	76-96 %	84-90 %	73-81 %	86-97 %
Class/All	93-96 %	90-94 %	82-85 %	86-97 %	79-85 %	72-81 %	88-98 %

Table D.37 *Coincidence within ±20ms deviation given as 95% confidence intervals for phoneme class transitions for the Norwegian SI,VC -test (150% o.s.).*

Italian:

Ita.	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids	All
Plos	0/0	0/0	0/0	592/611	3/3	128/136	64/70	787/820
Affr.	0/0	0/0	0/0	177/183	0/0	19/20	0/0	196/203
Fric.	52/59	0/0	0/0	349/371	0/0	13/13	26/26	440/469
Vow.	452/517	169/195	378/415	74/157	411/489	5/12	303/428	1791/2213
Nas.	100/126	3/4	14/15	316/375	0/0	4/4	0/3	437/527
Glid.	0/0	0/0	0/0	131/192	0/0	0/0	0/0	131/192
Liq.	55/76	0/0	10/13	329/417	11/24	6/7	4/8	415/545
All	659/778	172/199	402/443	1968/2306	425/516	175/192	397/535	4201/4973

Table D.38 *Number of automatic phoneme boundary placements within ± 20 ms deviation from manual segmentation in relation to number of occurrences in the material for each phoneme class transition for the Italian SI,VC -test (150% o.s.).*

The corresponding 95% confidence intervals for the coincidence rates within ±20ms from the manual segmentation:

Ita.	Plosives	Affricates	Fricatives	Vowels	Nasals	Glides	Liquids
All/Class	82-87 %	87-90 %	88-93 %	84-87 %	79-85 %	86-94 %	70-78 %
Class/All	94-97 %	93-98 %	91-96 %	79-83 %	79-86 %	61-64 %	72-80 %

Table D.39 *Coincidence within ±20ms deviation given as 95% confidence intervals for phoneme class transitions for the Italian SI,VC -test (150% o.s.).*

Danish:

Dan.	Plosives	Approximants	Fricatives	Vowels	Nasals	Liquids	All
Plos	23/57	3/12	24/30	571/603	39/44	62/82	722/828
Appr.	27/29	7/12	35/41	131/200	27/31	0/0	227/313
Fric.	145/150	2/4	10/12	328/367	15/25	19/32	519/590
Vow.	427/458	132/268	336/357	58/141	362/416	147/203	1462/1843
Nas.	75/88	18/20	88/90	248/280	1/3	11/16	441/497
Liq.	16/20	1/2	18/21	209/273	5/11	2/2	251/329
All	713/801	163/318	511/551	1545/1864	449/530	241/335	3622/4400

Table D.40 *Number of automatic phoneme boundary placements within ± 20 ms deviation from manual segmentation in relation to number of occurrences in the material for each phoneme class transition for the Danish SI,VC -test (150% o.s.).*

The corresponding 95% confidence intervals for the coincidence rates within ±20ms from the manual segmentation:

Dan.	Plosives	Approximants	Fricatives	Vowels	Nasals	Liquids
All/Class	87-91 %	46-57 %	90-95 %	81-85 %	81-88 %	67-76 %
Class/All	85-89 %	67-77 %	85-90 %	77-81 %	86-91 %	71-81 %

Table D.41 *Coincidence within ±20ms deviation given as 95% confidence intervals for phoneme class transitions for the Danish SI,VC -test (150% o.s.).*

RESULTS WHEN USING NO POLE IN THE RASTA FILTER

In tables D.42 and D.43 below the segmentation results when using the RASTA-filter without pole for each language are shown, see section 7.4.5.

	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
English, SI, V	92	2127	2934	3652	3774
English, SI, VC_150	74	2292	3035	3659	3797
Italian, SI, V	141	2290	3284	4161	4340
Italian, SI, VC_150	135	2262	3153	4115	4302
Norwegian, SI, V	93	2104	2894	3545	3658
Norwegian, SI, VC_150	68	2028	2800	3536	3722
Danish, SI, V	204	1980	2786	3477	3636
Danish, SI, VC_150	163	2115	2896	3550	3707

Table D.42 *Gross errors and fine errors for the SI-tests on English, Italian, and Danish EUROM0 recordings where the cepstral coefficients were filtered with a RASTA filter without pole.*

	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
English, SI, V	1.7 - 2.5 %	48.3 - 51.3 %	67.3 - 70.1 %	84.4 - 86.5 %	87.4 - 89.3 %
English, SI, VC_150	1.3 - 2.1 %	52.2 - 55.2 %	69.9 - 72.4 %	84.6 - 86.7 %	87.9 - 89.8 %
Italian, SI, V	2.3 - 3.2 %	44.7 - 47.4 %	64.7 - 67.3 %	82.6 - 84.7 %	86.3 - 88.2 %
Italian, SI, VC_150	2.2 - 3.1 %	44.1 - 46.9 %	62.1 - 64.7 %	81.7 - 83.8 %	85.5 - 87.4 %
Norwegian, SI, V	1.7 - 2.5 %	49.1 - 52.1 %	68.2 - 71.0 %	84.1 - 86.3 %	87.0 - 89.0 %
Norwegian, SI, VC_150	1.2 - 1.9 %	47.3 - 50.3 %	65.9 - 68.7 %	83.9 - 86.1 %	88.5 - 90.4 %
Danish, SI, V	3.9 - 5.1 %	43.5 - 46.5 %	61.9 - 64.7 %	77.8 - 80.2 %	81.5 - 83.7 %
Danish, SI, VC_150	3.1 - 4.1 %	46.6 - 49.5 %	64.4 - 67.2 %	79.5 - 81.8 %	83.1 - 85.3 %

Table D.43 *95% confidence intervals for the gross errors and fine errors for the SI-tests on English, Italian, and Danish EUROM0 recordings where the cepstral coefficients were filtered with a RASTA filter without pole.*

D.9 ENLARGING TRAINING SET ACROSS LANGUAGES

Table D.44 and D.45 show the results for the Norwegian SI-test when enlarging the training set for Norwegian by phonemes from other languages, see section 7.4.6.

Training conditions	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Combined, V	92	1871	2691	3487	3626
Combined, VC	56	1940	2718	3514	3650
Comb. + RASTA, V	54	1831	2802	3588	3724
Comb. + RASTA, VC	25	1938	2754	3582	3731
Comb. + RASTA + 2 mix, V	37	1761	2734	3623	3762
Comb. + RASTA + 2 mix, VC	28	1917	2727	3578	3736
Comb. + sex dep. + RASTA, V	76	1685	2645	3486	3620
Comb. + sex dep. + RASTA, VC	48	1855	2651	3476	3631
Eng. + Nor. + RASTA, V	48	2065	2933	3607	3739
Eng. + Nor. + RASTA, VC	34	2042	2840	3591	3734
Ita. + Nor. + RASTA, V	47	1771	2750	3602	3749
Ita. + Nor. RASTA, VC	29	1926	2742	3551	3711
Dan. + Nor. + RASTA, V	44	1778	2767	3591	3742
Dan. + Nor. + RASTA, VC	18	1931	2745	3547	3703

Table D.44 *Number of gross errors and number of coincidences within different deviation thresholds for the individual speakers in the SI-test on Norwegian EUROM0 using pure HMM segmentation (V) and constrained HMM segmentation (VC) with acoustic segmentation at 150% o.s..*

Training conditions	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Combined, V	1.7 - 2.5 %	43.5 - 46.5 %	63.3 - 66.2 %	82.7 - 84.9 %	86.2 - 88.2 %
Combined, VC	1.0 - 1.6 %	45.1 - 48.2 %	63.9 - 66.8 %	83.4 - 85.6 %	86.8 - 88.7 %
Comb. + RASTA, V	0.9 - 1.6 %	42.5 - 45.5 %	65.9 - 68.8 %	85.2 - 87.3 %	88.6 - 90.5 %
Comb. + RASTA, VC	0.4 - 0.8 %	45.1 - 48.1 %	63.5 - 66.4 %	85.1 - 87.2 %	88.8 - 90.6 %
Comb. + RASTA + 2 mix, V	0.6 - 1.1 %	40.9 - 43.9 %	64.3 - 67.2 %	86.1 - 88.1 %	89.5 - 91.3 %
Comb. + RASTA + 2 mix, VC	0.4 - 0.9 %	44.6 - 47.6 %	64.1 - 67.0 %	85.0 - 87.1 %	88.9 - 90.7 %
Comb. + sex dep. + RASTA, V	1.4 - 2.1 %	39.0 - 42.0 %	62.1 - 65.1 %	82.7 - 84.9 %	86.0 - 88.0 %
Comb. + sex dep. + RASTA, VC	0.8 - 1.4 %	43.1 - 46.1 %	62.3 - 65.2 %	82.4 - 84.7 %	86.3 - 88.3 %
Eng. + Nor. + RASTA, V	0.8 - 1.4 %	48.1 - 51.2 %	69.1 - 71.9 %	85.7 - 87.7 %	89.0 - 90.8 %
Eng. + Nor. + RASTA, VC	0.5 - 1.1 %	47.6 - 50.6 %	66.9 - 68.7 %	85.3 - 87.4 %	88.8 - 90.7 %
Ita. + Nor. + RASTA, V	0.8 - 1.4 %	41.1 - 44.1 %	64.7 - 67.6 %	85.6 - 87.6 %	89.2 - 91.0 %
Ita. + Nor. + RASTA, VC	0.5 - 0.9 %	44.9 - 47.8 %	64.5 - 67.4 %	84.3 - 86.4 %	88.3 - 90.2 %
Dan. + Nor. + RASTA, V	0.7 - 1.3 %	41.3 - 44.3 %	65.1 - 68.0 %	85.3 - 87.4 %	89.0 - 90.9 %
Dan. + Nor. + RASTA, VC	0.3 - 0.6 %	44.9 - 48.0 %	64.6 - 67.4 %	84.2 - 86.3 %	88.1 - 90.0 %

Table D.45 *95% confidence intervals for gross errors and coincidence rates within different deviation thresholds for the SI-test on Norwegian EUROM0 using pure HMM segmentation (V) and constrained HMM segmentation (VC) with acoustic segmentation at 150% o.s..*

Training conditions	Gross errors	Deviation from manual broad phonetic segmentation			
		< 5 ms	< 10 ms	< 20 ms	< 25 ms
Others, V	3.6 - 4.7 %	33.7 - 36.7 %	52.7 - 55.7 %	74.5 - 77.1%	79.0 - 81.5 %
Others, VC	1.8 - 2.6 %	40.0 - 43.0 %	57.3 - 60.3 %	77.4 - 79.9 %	81.4 - 83.7 %
Others + RASTA, V	2.6 - 3.6 %	35.1 - 38.0 %	57.4 - 60.4 %	78.7 - 81.2 %	82.9 - 85.1 %
Others + RASTA, VC	1.4 - 2.2 %	41.1 - 44.1 %	59.7 - 62.6 %	79.8 - 82.2 %	84.3 - 86.4 %
Others, init., V	1.5 - 2.3 %	46.8 - 49.9 %	66.3 - 69.1 %	83.7 - 85.9 %	86.6 - 88.6 %
Others, init., VC	0.9 - 1.5 %	46.9 - 49.9 %	65.3 - 68.1 %	84.1 - 86.2 %	87.3 - 89.2 %
Others + RASTA, init., V	0.7 - 1.3 %	47.3 - 50.4 %	68.9 - 71.7 %	87.4 - 89.4 %	90.4 - 92.2 %
Others + RASTA, init., VC	0.3 - 0.7 %	47.6 - 50.3 %	66.8 - 69.6 %	85.7 - 87.7 %	89.5 - 91.3 %

Table D.46 *95% confidence intervals for the coincidence rates and gross errors for the Norwegian SI-test with phoneme HMMs trained on polyphonemes from the English, Italian, and Danish EUROM0. Results with pure HMM segmentation (V) and constrained HMM segmentation (VC) with acoustic segmentation at 150% o.s. are provided.*

D.10 PHONEME RECOGNITION

Table D.47 and D.48 provide some phoneme recognition results as described in section 7.5.

Norwegian	Del.	Sub.	Ins.	H=N-D-S	%Correct	%Accuracy
Full HMM, V	128	1423	1370	2949	65.53	35.09
Full HMM +grammar, V	167	1068	822	3265	72.56	54.29
Full HMM, VC	198	1464	734	2838	63.07	46.76
Full HMM +grammar, VC	243	1049	410	3208	71.29	62.18
SI, V	173	1839	1928	2488	55.29	12.44
SI + grammar, V	211	1665	1240	2624	58.31	30.76
SI, VC	232	1869	1047	2399	53.31	30.04
SI + grammar, VC	337	1575	607	2588	57.51	44.02
SI, l+s, V	147	1650	1859	2703	60.07	18.76
SI, + grammar, l+s, VC	355	1363	588	2782	61.82	48.76

Table D.47 *Recognition results for the Norwegian SI and Full HMM tests, with pure HMM recognition (V) and constrained HMM recognition (VC) with acoustic segmentation at 150% o.s., and with and without the grammar. For the SI-test, also the results with long and short vowels joined (l+s), are given.*

	Del.	Sub.	Ins.	H=N-D-S	%Correct	%Accuracy
AFN , SI, V	23	412	623	709	61.98	7.52
SHN, SI, V	41	442	410	644	57.14	20.76
TBN, SI, V	61	491	395	560	50.36	14.84
TGN, SI, V	48	494	500	575	51.48	6.71
AFN, SI +grammar, V	31	389	413	724	63.29	27.19
SHN, SI +grammar, V	41	419	288	667	59.18	33.63
TBN, SI +grammar, V	71	437	272	604	54.32	29.86
TGN, SI+ grammar, V	68	420	267	629	56.31	32.41
AFN, SI, VC	35	413	333	696	60.84	31.73
SHN, SI, VC	63	453	237	611	54.21	33.19
TBN, SI, VC	78	499	220	535	48.11	28.33
TGN, SI, VC	56	504	257	557	49.87	26.86
AFN, SI,+ grammar, VC	51	362	178	731	63.90	48.34
SHN, SI, +grammar, VC	95	390	156	642	56.97	43.12
TBN, SI, +grammar, VC	109	412	135	591	53.15	41.01
TGN, SI, +grammar, VC	82	411	138	624	55.86	43.51

Table D.48 *Recognition results for the Norwegian SI test for each speaker, with pure HMM recognition (V) and constrained HMM recognition (VC) with acoustic segmentation at 150% o.s., and with and without the grammar.*

Constrained HMM phoneme recognition as a function of acoustic oversegmentation factor:

	Deletions	Subst.	Insertions	H=N-D-S	%Corr	%Acc.
SI, V	173	1839	1928	2488	55.29	12.44
SI + grammar, V	211	1665	1240	2624	58.31	30.76
SI, VC_150	232	1869	1047	2399	53.31	30.04
SI + grammar, VC_150	337	1575	607	2588	57.51	44.02
SI, VC_100	265	1870	893	2365	52.56	32.71
SI + grammar, VC_100	361	1572	502	2567	57.04	45.89
SI, VC_75	330	1836	760	2334	51.87	34.98
SI + grammar, VC_75	425	1565	398	2510	55.78	46.93
SI, VC_50	405	1795	621	2300	51.11	37.31
SI + grammar, VC_50	497	1500	341	2503	55.62	48.04
SI, VC_25	535	1736	428	2229	49.53	40.02
SI + grammar, VC_25	632	1433	247	2435	54.11	48.62
SI, VC_0	770	1648	258	2082	46.27	40.53
SI + grammar, VC_0	896	1365	172	2239	49.76	45.93

Table D.49 *Recognition performance for the Norwegian SI-test with pure HMM recognition (V) and constrained HMM recognition (VC) for different oversegmentation factors, with and without the use of grammar.*

GLOSSARY

Accent A cover term related to the syllable, and includes length, stress and tone.

ACCOR ESPRIT II Basic Research Action-project which aims at investigating the articulatory-acoustic correlations in coarticulatory processes in seven European languages [Marchal'92].

Affricate A speech sound where a plosive precedes a homorganic fricative, as [tS] in the beginning of the English word "church".

Airstream mechanism describes by the initiating air source, i.e. pulmonic (from the lungs), glottalic (caused by vertical movements of the larynx with closed vocal cords) or velaric (caused with a closure between the back of the tongue against the soft palate and some oral closure further forward in the mouth), and the direction of the airstream, i.e. egressive (outgoing) or ingressive (ingoing).

Allophone "A variant of a phoneme. The allophones of a phoneme form a set of sounds that (1) do not change the meaning of a word (2) are all very similar to one another, and (3) occur in phonetic contexts different from one another - for example, syllable initial as opposed to syllable final. The differences among allophones can be stated in terms of phonological rules" [Ladefoged'82, p.280].

Allophonic rule A description of the allophones belonging to a phoneme and in which context they appear.

Alveolar An articulation involving the tongue and the alveolar ridge (the area behind the roots of the teeth, see e.g. figure 2.1). Typical alveolar sounds are [d] and [s].

Analysis-synthesis systems A speech coding scheme where the speech wave is represented by a set of parameters based on the speech production model (as opposed to the waveform coding).

Analysis-by-synthesis The process of determining parameters which characterise a system by comparing the output of the model with the original signal and change the parameters according to the deviation. This is done iteratively until the deviation is less than a given threshold.

ANN Artificial Neural Net (see e.g. [Lippmann'87]).

ANOVA ANalysis Of VAriance (see e.g. [Johnson'88]).

Annotation Used here as a cover term for segmentation and labelling of speech.

Approximant An articulation in which one articulator is close to another but without the vocal tract being narrowed to such extend that a turbulent airstream is produced. In many forms of English /j l w/ are approximants [Ladefoged'82], and in Norwegian the phonemes /v j l/ are approximants.

AR-model Auto Regressive model (see e.g. [Orfanidis'88]).

Articulation rate The average number of phonemes per time unit within the portions of the recordings which are marked as speech.

Aspiration A period of voicelessness and friction noise caused by the outgoing airstream after the release of the (complete) closure of a plosive, as in English "pie" [p^{h}Ai] and in Norwegian "på" [p^{h}O:]. (Cf. burst below).

ASR Automatic speech recognition.

ASS Automatic speech segmentation.

Assimilation A sound becomes more like the following sound (regressive assimilation) or the preceding (progressive assimilation) or both. The sound may also be influenced by more distant sounds.

In **allophonic** assimilation the product, i.e. the new sound, is an allophone of original phoneme as in the Norwegian word "salt" /salt/ ('salt') where the lateral may become devoiced. In **phonemic** assimilation the product is a sound which is an allophone of another phoneme as e.g. in the change of underlying [n] to [m] in the English word "input" and the Norwegian word "kronprins".

Biallophone An allophone that occurs only in special contexts in contrast to the *main allophone* of a phoneme which has more general distribution [Endresen'88, p.54].

Bilabial A sound produced with the lips pressed together as in [m] and [b].

Breathy voice See Murmur.

Broad phonetic transcription Labelling of continuous speech using only symbols with phonemic values.

Burst A plosive burst consists of the impulse (or impulses) exceeding the amplitude of the following friction [Barry'90b]. The burst may be difficult to distinguish from the aspiration.

Cardinal vowels A set of reference vowels first defined by Daniel Jones. The vowels of any language can be described by stating their relations to the cardinal vowels [Ladefoged'82].

CCS Constrained Clustering Vector Quantization Segmentation [Svendsen'87].

Centroid The generalised centroid of a vector sequence is the vector which represents the sequence with minimum distortion.

Citation form The form a word has when it is cited or pronounced in isolation [Ladefoged'82].

Coarticulation Articulatory movements for successive phones overlapping in time.

In **anticipatory** coarticulation (or right-to-left coarticulation), articulators not needed in producing the actual sound move toward their next required state. For example in the Norwegian word "lå" /lO:/ ('lay'), the lip rounding for /O:/ starts before the tongue moves from the /l/ position to the /O:/ position. This imposes a lowering of the formants in /l/ which is usually not perceived. (Anticipatory express that this coarticulation is due to the planning of speech production).

In **carry over** coarticulation (or left-to-right coarticulation) features of one phoneme affect the realisation of the succeeding phoneme(s), as noticed in the formant transitions of /O:/ in the Norwegian word "må" /mO:/ ('must').

Constrained HMM segmentation Segmentation where the Viterbi forward recursion is constrained by acoustic segment-boundaries. Is abbreviated *VC*.

Content words Words carrying most of the meaning in a sentence.

Contoid Phonetically defined term for sounds which are not vocoids (see vocoids).

Creaky voice A type of phonation in which the artyenoid cartilages hold the posterior end of the vocal cords together, so that they can vibrate only at the other end [Ladefoged'82].

CVC-word Words with the structure: Consonant-Vowel-Consonant.

Demi-syllable The part of the syllable extending from syllable boundary to syllable nucleus (initial demi-syllable), and the part extending from nucleus to next syllable boundary (final demi-syllable). The nucleus is split "in the middle" (stationary region) [Ruske'85].

Deviation threshold By a deviation of 20ms we understand a deviation threshold of ±20ms for the boundary displacement between automatic and manual segmentation.
The terms *deviation* or *deviation threshold* will here denote the *absolute deviation values*.

Diacritics Additional marks placed near a symbol that can be used in transcription to distinguish different values of a symbol. For example the : denotes length; see length (duration).

Diphone A segment of speech that contains the last part of one phoneme, the transition, and the first part of the next phoneme.

Diphthong A vowel in which there is a change in quality during a single syllable, as in English [Ai] in "high" or Norwegian "mai" ('May').

Discourse A sequence of sentences which cohere, as in a conversation, a story, or book [Taylor'90].

Distinctive features The minimal number of phonetic features needed to separate one phoneme from all the other phonemes in the language.

DP Dynamic programming. An optimising algorithm based on the principle of optimality [Bellman'62], [Silverman'90].

DTW Dynamic Time Warping. A technique used e.g. in template based ASR where the templates are stretched or compressed relative to each other to obtain the best match between them.

Dyad A dyad is a transition between two successive phonemes and every utterance is assumed to be a sequence of unique transitions and "steady states" [Pols'83].

EEG Electroencephalogram, which is electrical signal recorded from a human's skull.

ELSA Esprit Labelling Software Assessment [Bourjot'91].

Epenthetic sound An extra sound between two others due to synchronism problems of the articulators, e.g. the Norwegian name "Henrik" is often realised [h { n d r i k].

EPG Electropalatography, a technique where an artificial palate registers contact between different parts of the palate and the tongue (see e.g. [Marchal'92]).

ESPRIT "**E**uropean **S**trategic **P**rogramme for **R**esearch and Development in **I**nformation **T**echnology".

ETR The ESPRIT-SAM project was divided into three parts: Speech recognition assessment, speech synthesis assessment, and ETR = **E**nabling **T**echnologies and **R**esearch.

EUROM0 The first speech recording at CD-ROM in the ESPRIT-SAM project.
The EUROM0 CD-ROM contains recordings of Danish, Dutch, English, French, and Italian. For each language single digits, digit triples, and a continuous passage were recorded [Grice'89]. The label-files are not included on the CD-ROM, but they are available in ASCII-format on floppy disks.
The corresponding Norwegian and Swedish recordings are not included in the CD-ROM, but they are available on the same format as the other languages from National Physical Laboratory, United Kingdom.
In this thesis the term *EUROM0 recordings* means the continuous passage recording for all the languages mentioned above.

EUROM1 The second speech recordings in the ESPRIT-SAM project. These recording are not on CD-ROM yet, but there is planned to make 3 CD-ROMs for each language participating in the SAM-project. Details can be found in [Sherwood'92].
This speech corpus is not used in this thesis.

Extralinguistic sounds Sounds that are not speech sounds, such as stomach rumbling, breath noise, lip smacking etc.

FFT Fast Fourier Transform (see e.g. [Oppenheim'75]).

Fine error The deviation between the automatic and the corresponding manual segment boundary.

Flap An articulation in which one articulator, usually the tongue tip, is drawn back and then allowed to strike against another articulator before returning to its rest position. The /t/ in "ditty" is often a flap in American English [Ladefoged'82].

Formant A group of overtones corresponding to a resonating frequency of the air in the vocal tract [Ladefoged'82]. A vowel can be characterized by its three lowest formants, i.e. F_1, F_2, and F_3.

Fricatives Speech sounds as [f] and [s] produced by forcing air through a constriction in the vocal tract so that turbulence or hissing noise is generated.

Function words Words as articles, conjunctions and prepositions. Is used here for unstressed words that are not important for the meaning, such as in English "a, and, are, for, in, is ..." (see also content words).

Fundamental frequency (F_0) The rate at which the vocal cords vibrate in a voiced sound.

Geminate Adjacent segments that are the same (but belonging to two different syllables), such as the two consonants in the middle of Italian "folla" [folla] ('crowd') [Ladefoged'82].

Glide see semivowel.

Glottal sound A phone, where the glottis is the primary place of articulation, e.g. [h] in English "horse" and Norwegian "hest". A **glottal stop** is produced with the vocal cords held tightly together, as [?] in many forms of English "button" ['bV?n] or "kitten" [k^{h}I?n].

Glottis The space between the vocal cords.

Grammar A description of the syntax, often expressed as a set of formal rules.
In statistical language models a **N-gram grammar** estimates the probability of a word in a sentence given the N previous words. In practice, only N=2 and N=3 are used implying bigram and trigram grammars respectively.
Used in ASR to eliminate many candidates from consideration, or to assign higher probabilities to some word candidates than others.

Gross error When the automatically placed boundary is so displayed relative to the manual reference that its position passes beyond the region of the adjacent manually labelled SAMPA segments.

Homographs Words that have the same spellings but differ in meaning and pronunciation.

Homophones Words which have identical pronunciation but which may differ in meaning and/or spelling, such as "to, too, two".

Homorganic Sounds produced at the same place of articulation and with the same articulators, as e.g. [d] and [n] in English "hand".

HMM Hidden Markov Model, (see e.g. [Rabiner'89a]).

IIR Infinite impulse response, see e.g. [Oppenheim'75].

Intensity The amount of acoustic energy in a sound.

Intonation The pattern of pitch changes that occur during an utterance, which may be a complete sentence (it is superimposed on word tones) [Ladefoged'82].

IPA The International Phonetic Alphabet. A system for transcribing speech sounds adopted by International Phonetic Association, see e.g. [IPA'89].

Juncture A syllable boundary which is important for the meaning in a language, such as the difference between "that stuff" and "that's though", or in Norwegian "lettet anker" (=lifted the anchor) and "lette tanker" (=hight tanks).

Labelling Identification of the segments and giving names according to a predefined set of labels.

Labiodental An articulation involving the lower lip and the upper front teeth.

Language The set of conventions and signs that a group people agree to use for communicating meaning [Dew'77].

Lateral An articulation in which the airstream flows over the sides of the tongue and there is a median closure, as in English [l] in "leaf" or Norwegian [l] in "loff" (=white bread).

Lateral plosion The release of a plosive by lowering one or both sides of the tongue, as at the end of the word "bottle" or as in Norwegian "Atle" [Ladefoged'82].

LBG A VQ-algorithm named after Linde, Buzo and Gray [Linde'80].

Length (duration) Long and short vowels have in Norwegian phonemic value, for example the minimal pair "vin-vinn", /vi:n/ - /vin/ (=wine-win).

Lexicon A catalogue containing one or more descriptions of each word in the vocabulary. In each such description a sequence of subwords corresponding to a standard pronunciation of a word as an isolated utterance is included.

Linguistics The study of language as a formal and abstract system.

Liquid A cover term for laterals and various forms of r-sounds [Ladefoged'82].

Logatome Meaningless words, usually in the form of a consonant-vowel-consonant (CVC) sequence.

Lombard effect Speech produced when 90 Db pink noise is injected into the speakers ears [Shanton'89].

Loudness The auditory property of a sound that enables a listener to place it on a scale from soft to loud without considering the acoustic properties, such as the intensity of the sound [Ladefoged'82].

LPC Linear Predictive Coding (see e.g. [Rabiner'78]).

Masking The effect that strong frequency components mask the ear's response to weaker components, see e.g. [Zue'89].

MFCC Mel-scale cepstrum coefficients.

MLR Maximum Likelihood Ratio (distortion measure).

MLS Maximum Likelihood Segmentation [Bridle'77].

Monophthong A vowel in which there is no appreciable change in quality during a syllable, as in English [A:] in "father". Compare Diphthong [Ladefoged'82].

Monophonemes Mono-lingually defined phonemes, i.e. phonemes which are not phonetically equal to any phoneme in any other language, see polyphonemes [Dalsgaard'91].

Morph The smallest subword giving meaning or having grammatical function in a language. For example the English words "cats" and "dogs" have the morphs /kæt/ /s/, /dog/ /z/. (As an analogy to phone-allophone-phoneme, we get morph-allomorph-morpheme).

MRI Magnetic Resonance Imaging, see e.g. [Foldvik'88].

Murmur Another name for breathy voice, a type of phonation in which the vocal cords are only slightly apart so that they vibrate while allowing a high rate of airflow through the glottis [Ladefoged'82].

Narrow phonetic transcription A description of speech showing phonetic details by using a big inventory of symbols and diacritics.

Nasal A sound produced by lowering the soft palate so that all the air is expelled through the nose without any velic closure.

Nasalization Lowering of the soft palate during a sound resulting in the air passing out through the mouth as well as the nose, as in the vowel [A] between nasals in Norwegian "mann".

Nasal plosion The release of a plosive by lowering the soft palate so that air escapes through the nose, as at the end of the word "hidden" [Ladefoged'82].

Oversegmentation The number of acoustic segments are larger than the number of phonemes. By 150% oversegmentation we mean that the number of acoustic segments are 2.5 times the number of phonemes for each sentence.
Acoustic oversegmentation is abbreviated *o.s.*.

Perception An interpretation of a sensation in light of experience [Dew'77].

Phone The smallest sound unit abstracted from the continuum of speech by segmenting it on the basis of auditory impression and articulatory movement. The phone is unique and cannot be repeated identically. The phone system is language-independent.

Phoneme The smallest phonological unit that has distinctive function in a language.

Phoneme boundary effect In listening to speech sounds, discrimination of slightly differing sounds is good across the boundaries of phonemes but poor within a phoneme [Taylor'90].

Phoneme restoration The reconstruction of the phonemes from the speech signal, also called phoneme correction.

Phonemic cue Cue which gives difference in meaning of a word in the language (as opposed to a phonetic or allophonic cue).

Phonemically balanced A speech corpus containing equal number of all phonemes in the given language, i.e. "phonemic compact".

Phonemic transcription The transcription used in lexicon/dictionary.

Phonetics Describes the actual realisation of speech in terms of language independent features, which may be either articulatory, acoustically or auditorily defined.

Phonetic features Language independent features for describing a phone, such as [p] as voiceless, unaspirated, bilabial plosive.

Phonology An abstract, language dependent description of the system and function of the speech sounds.

Phonotactics Describes how the phonemes in a given language can be combined, i.e. the permissable sound sequences of that language.

Phonotypical A transcription predicting what would be written down in an auditory transcription of fluent continuous speech, including assimilations and elisions.

Pitch The auditory property of a sound that enables a listener to place it on a scale going from low to high, without considering the acoustic properties, such as the frequency of the sound [Ladefoged'82].

Plosive A phone characterized by complete closure at some point in the articulatory channel as [b],[d],[g],[p],[t] and [k] in English and Norwegian.

PLP Perceptual Linear Predictive technique [Hermansky'90].

Poly-phonemes Sets of phonemes in different languages with properties similar enough to be equated for certain purposes.

Prosody Literally "To the song" (pro=to, sode=song). Is used as a cover term for accent and intonation.

Psycholinguistic The study of how people learn and use language to communicate ideas.

Pure HMM segmentation Segmentation by Viterbi decoding (see constrained HMM segmentation).

RASTA Relative spectral processing [Hermansky'91].

Reduced vowel A vowel that is pronounced with a noncontrasting centralized quality, although in the underlying form of a word it is part of a full set of contrasts. The second vowel in "emphasis" is a reduced form of the vowel /{/, as in "emphatic" [Ladefoged'82].

Retroflex An articulation involving the tip of the tongue and the back part of the alveolar ridge [Ladefoged'82], as in Norwegian [rn] in "garn" (=net).

Ripple errors One miss in the automatic segmentation that may cause many subsequent errors.

SAM ESPRIT project no. 2589 is called "Multi-Lingual Speech Input/Output Assessment, Methodology and Standardisation", which for short is called SAM = "Speech Assessment Methods" [Fourcin'91].

SAMPA SAM Phonetic Alphabet [Wells'89]. Corresponds to IPA, but the symbols can be written with the ASCII characters, i.e. computer compatible. See Appendix A.

SAMPROSA SAM PROsodic Alphabet, [Gibbon'90]. A standard for prosodic transcription of the European languages in EEC.

Schwa An central and unstressed "vowel", symbolised with @ in SAMPA, as in the English word "agree" [@gri] and the Norwegian word "flytte" [flyt@] ('remove').

Segmental unit One phone.

Segmentation The process of dividing something continuous into discrete, non-overlapping entities; i.e. the process of deciding boundaries.

Semivowel A sound articulated in the same way as a vowel, but not forming a syllable on its own, as in [w] in the English word "we" [Ladefoged'82].

Sentence "A grammatical unit of one or more words, bearing the minimal syntactic relation to the words that precede or follow it, of preceded and followed in speech by pauses ..." [Webster].
In speech segmentation tasks as in this thesis, a sentence denotes the vocal activity between two pauses, i.e. an utterance.

Sonority The loudness of a sound relative to that of other sounds with same length, stress and pitch [Ladefoged'82, p.221].

Speaking rate The average number of phonemes per time unit when the total recording time is taken into account.

Spectrogram A machine-made graphic representation of sounds in terms of their component frequencies, in which time is shown on the horizontal axis, frequency on the vertical axis, and the intensity of each frequency at each moment in time by the darkness of the mark [Ladefoged'82].

Speech "The expression of ideas and thought by means of articulate vocal sounds, or the faculty of thus expressing ideas and thoughts" [Webster].

Speech chain The chain of events from the conception of a message in the speaker's brain to the arrival of the message in the listener brain.

Speech synthesis Speech produced by computers (see TTS).

SPIN ESPRIT project on **Sp**eaker **In**dependent continuous speech recognition.

Stress The use of extra respiratory energy during a syllable [Ladefoged'82].
Stress is phonemic in English e.g. "to export" vs. "an export".

Subwords Speech units (phonetic) or segments (acoustic), corresponding to simple or more complex sounds or speech intervals [Colla'89a].

Suprasegmental Phonetic features such as stress, length, tone and intonation, which are not properties of single consonants or vowels [Ladefoged'82].

Syllable A phonological, structural unit, describing the characteristic combination of vowel+consonant (VC) in a language giving combinations as e.g. VC, CV, CVC or V only or C only. No good phonetic definition exits.

Syntax Systematic arrangement of parts; the study of sentence structure [Taylor'90].

Systematic phonetic transcription A transcription that shows all the phonetic details that are part of the language and can be stated in terms of phonological rules [Ladefoged'82].

Target An idealized articulatory position that can be used as a reference point in describing how a speaker produces utterances [Ladefoged'82].

TOBI Tones and Break Indices [Silverman'92]. A standard for prosodic transcription of American English.

Tones Pitch variations that affect the meaning of a word. In Norwegian, the tones are related to a stressed syllable followed at least by one unstressed, as in rota (n) ('the root') vs. rota (v) ('messed').

Transcription A segmental description in terms of a set of previously defined symbols. The symbols are notated in the correct time order (i.e. corresponding to the time sequence in which the acoustic events have been perceived) and at a specified level of detail [Wells'88].

Triphone Two possible interpretations: (1) the speech signal corresponding to two subsequent diphones, i.e. from the midpoint of one phoneme, the whole following phoneme, and the half of the third subsequent phoneme.
(2) A phoneme in a given fixed phoneme context.

Trill An articulation in which one articulator is held loosely near to another so that the flow of air between them sets one or both of them in motion, alternatively sucking them together and blowing them apart [Ladefoged'82].

TTP Text-to-Phoneme. A module in the TTS-system which converts the orthographic text into a phonotypical transcription.

TTS Text-to-Speech. Synthetic speech generated by a computer from a given text.

Uncertain area The transition segment where two sounds are perceived (see chapter 4).

Velar pinch When a velar sound, e.g. [k], [g] or [N], appear in an intervocalically context, the formants F_2 and F_3 of the vowel move towards each other.

Vocal tract Consists of the pharynx, the oral tract and the nasal tract, i.e. it is the air passage above the larynx.

Voice Onset Time (VOT) The time between the onset of the plosive's burst and the onset of voicing of the succeeding sound. In utterance initial position in English and Norwegian [b] has very short VOT whereas [p] has long VOT.

Vocoid By vocoid we mean a speech sound articulated with the air escaping along the median line of the tongue with a minimum of local friction [Sivertsen'88].

VQ Vector Quantization (see [Buzo'80],[Linde'80],[Gray'84],[Pan'85]).

z-transform The z-transform X(z) of a discrete time sequence x(n) is defined as

$$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}$$

[Oppenheim'75, p.45]

BIBLIOGRAPHY

Abbreviations:

ASSP	Acoustic, Speech and Signal Processing
ESCA	European Speech Communication Association
EUROSPEECH	European Conference on Speech Communication and Technology
ICASSP	International Conference on Acoustics, Speech and Signal Processing
ICSLP	International Conference on Spoken Language Processing
IEEE	The Institute of Electrical and Electronics Engineers, Inc.
JASA	Journal of Acoustic Society of America
SAP	Speech and Audio Processing
SAM	An ESPRIT project, see Glossary
ETR	A subgroup of the SAM-project

Book titles are written in *italics*.

Algazi, R.V. and Brown, K.L., "Automatic Speech Recognition Using Acoustic Sub-Words and no Time Alignment", Proc. ICASSP-88, pp. 465-468, 1988.

Andre-Obrecht, R., "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals", IEEE Trans. on ASSP, Vol. 36. No.1, Jan. 1988.

Atal, S.B. and Schrøder, M.R., "Predictive coding of speech signals", Proc. 6th Int. Cong. Acoust., C-5-4, 1968.

Atal, S.B. and Hauner, S.L., "Speech analysis and synthesis by linear prediction of the speech wave", Proc. JASA, Vol. 50, pp. 637-655, 1971.

Atal, S.B., "Effectiveness of linear prediction characteristics of a speech wave for automatic speaker identification and verification" Proc. JASA, Vol. 55, pp. 1304-1312, June, 1974.

Atal, S.B., "Automatic recognition of speakers from their voices", Proc. IEEE, Vol. 64, pp. 460-475, April, 1976.

Atal, S.B., "Efficient coding of LPC parameters by temporal decomposition", Proc. ICASSP-83, pp. 81-84, 1983.

Auteserre, D., Perennou, G., and Rossi, M., "Methodology for transcription and labelling of speech corpus", Proc. in Journal of IPA, Vol.19, no.1., pp. 2-15, 1989.

Auteserre, D. and Meunier, C., "Segmentation criteria and labelling conventions", SAM-document, Aix, Feb. 1991.

Averbuch, A., et.al, "An IBM-PC based large-vocabulary isolated utterance speech recognizer", Proc. ICASSP-86, pp. 53-56, 1986.

Bahl, L.R., Bakis, R., Cohen, P.S., Cole, A.G., Jelinek, F., Lewis, B.L. and Mercer, R.L., "Continuous Parameter Acoustic Processing for Recognition of a Natural Speech Corpus", Proc. ICASSP-81, pp. 1149-1151, 1981.

Bahl, L.R., Jelinek, F. and Mercer, R.L., "A Maximum Likelihood Approach to Continuous Speech Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. Pami-5, No. 2, pp. 179-190, 1983.

Bahl, L.R., Das, S.K., de Sourza, P.V., Jelinek, F., Katz, S., Mercer, R.L. and Picheny, M.A., "Some Experiments with Large-Vocabulary Isolated-Word Sentence Recognition", Proc. ICASSP-84, pp. 26.5.1-4, 1984.

Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L. and Picheny, M.A., "Acoustic Markov Models used in the Tangora Speech Recognition System", Proc. ICASSP-88, pp. 497-500, 1988.

Bahl, L.R., Bakis, R., Bellegarda, J., Brown, P.F., Burshtein, D., Das, S.K., de Souza, P.V., Gopalakrishnan, P.S., Jelinek, F., Kanevsky, D., Mercer, R.L., Nadas, A.J., Nahamoo, D. and Picheny, M.A., "Large Vocabulary Natural Language Continuous Speech Recognition", Proc. ICASSP-89, pp. 465-467, 1989.

Baker, J.M., "Large vocabulary speaker-adaptive speech recognition overview at Dragon System", Proc. EUROSPEECH'89, pp. 29-32, Paris, 1989.

Baker, J.M., "Dragondictate™-30K: Natural language speech recognition with 30,000 words", Proc. EUROSPEECH'91, Vol.2, pp. 161-165, Genova, 1991.

Barry, W.J. and Fourcin, A.J., "Levels of labelling" in *Speech, hearing and language, work in progress 1990*, vol 4, University College London, Department of Phonetics and Linguistics, pp. 31-43, 1990.

Barry, W.J., "Labelling criteria: phonemic and acoustic-segment labelling", Document no. SAM-UCL-026, for workgroup ETR, oct. 1990.

Barry, W.J. and Grice, M., "Auditory and visual factors in speech database analysis", in *Speech, hearing and language, work in progress 1991*, vol. 5, University College London, Department of Phonetics and Linguistics, pp. 11-31, 1991.

Barry, W.J., "SALA labelling tests", Summary report, SAM-document: SAM UCL-020, 1991.

Barry, W. and Dalsgaard, P., "Speech database annotation. The importance of a multi-lingual approach", Proc. EUROSPEECH'93, pp. 13-20, Berlin, 1993.

Baudary, M. and Dupeyrat, B., "Speech Segmentation and Recognition Using Syntactic Methods on the Direct Signal", Proc. ICASSP-79, pp. 101-104, 1979.

Beinum, F.J.K., "Spectro-temporal reduction and expansion in spontaneous speech and read text: focus words versus non-focus words", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 36.1-5, Barcelona, 1991.

Bellman, R.E. and Dreyfus, S.E., *Applied Dynamic Programming*, Princeton University Press, Princeton, NJ, 1962.

Bengio, Y., Cardin, R., De Mori, R. and Normandin, Y., "A Hybrid Coder for Hidden Markov Models Using Recurrent Neural Network", Proc. ICASSP-90, pp. 537-540, 1990.

Bertsekas, D.P., *Dynamic Programming*, Prentice-Hall, Inc. 1987.

Bimbot, F., Chollet, G., Deleglise, P. and Montacie, C., "Temporal Decomposition and Acoustic-Phonetic Decoding of Speech", Proc. ICASSP-88, pp. 445-448, 1988.

Blaauw, E., "Phonetic characteristics of spontaneous speech and read-aloud speech", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 12.1-5, Barcelona, 1991.

Blomberg, M. and Nord, L., "Comparison between manually and automatically labelled speech files", ETR-document, oct. 1990.

Blomberg, M. and Carlson, R., "Labelling of speech given its text representation", Proc. EUROSPEECH'93, pp. 1775-1778, 1993.

Bocchieri, E.L. and Wilpon, J.G., "Discriminative analysis for feature reduction in automatic speech recognition", Proc. ICASSP'92, pp. I.501-504

Boeffard, O., Miclet, L. and White, S., "Automatic generation of optimized unit dictionaries for text to speech synthesis", Proc. ICSLP'92, pp. 1211-1214, Banff, Canada, 1992.

Boeffard, O., Cherbonnel, B., and White, S., "Automatic segmentation and quality evaluation of speech unit inventories for concatenation-based, multilingual PSOLA Text-To-Speech systems", Proc. EUROSPEECH'93, pp. 1449-1452, 1993.

Bourjot, C., Boyer, A., and Fohr, D., "Phonetic Decoder Assessment", Proc. EUROSPEECH'89, Vol II, (S.43), pp. 457-460, Paris, 1989.

Bourjot, C., Boyer, A., and Fohr, D., "ELSA - ESPRIT Labelling system assessment software. User's guide for v2.3", SAM document SAM-UCL-027, R9 in SAM interim report, year two, 1991.

Bourlard, H., Kamp, Y. and Wellekens, C.J., "Speaker Dependent Connected Speech Recognition via Phonemic Markov Models", Proc. ICASSP-85, pp. 1213-1216, 1985.

Bourlard, H. and Wellekens, C.J., "Links between Markov models and multilayer perceptrons", IEEE Trans. on Pattern analysis and machine intelligence, Vol. 12, No. 12, Dec. 1990.

Brassard, J.-P., "Integration of Segmenting and Nonsegmenting Approaches in Continuous Speech Recognition", Proc. ICASSP-85, pp. 1217-1220, 1985.

Bridle, J.S. and Sedgwick N.C., "A method for segmenting acoustic patterns, with application to automatic speech recognition", Proc. ICASSP-77, pp. 656-659, 1977.

Browman, C.P., "Rules for Demisyllable Synthesis using Lingua, a Language Interpreter", Proc. ICASSP-80, pp. 561-564, 1980.

Brown, K.L. and Algazi, R.V., "Characterization of Spectral Transitions with Applications to Acoustic Sub-Word Segmentation and Automatic Speech Recognition", Proc. ICASSP-89, pp. 104-107, 1989.

Brown, M.K., Maureen, A., Rabiner, L.R. and Wilpon, J.G., "Training Set Design for Connected Speech Recognition", Trans. on Signal Processing, Vol. 39, No. 6, pp. 1268-1281, June 1991.

Bush, M.A. and Kopec, G.E., "Evaluation of a Network-Based Isolated Digit Recognizer Using the TI Multi-Dialect Database", Proc. ICASSP-85, pp. 846-849, 1985.

Bush, M.A. and Kopec, G.E., "Network-Based Connected Digit Recognizer Using Vector Quantization", Proc. ICASSP-85, pp. 1197-1200, 1985.

Buzo, A., Gray, A.H. Jr., Gray, R.M. and Markel, J.D., "Speech Coding Based upon Vector Quantization", Proc. ICASSP-80, pp. 15-18, 1980.

Caerou, J.C. et.al., "PTS software v: 4.30, User Manual", GRECO I.C.P., 1991.

Castelaz, P.F. and Niederjohn, R.J., "A Comparison of Linear Prediction, FFT, and Zero-crossing Analysis Techniques for Vowel Recognition", Proc. ICASSP-78, pp. 541-545, 1978.

Chafcouloff, M., Chollet, G., Durand, P., Guizol, J. and Rodet, X., "Observation and Modelling of "Formant" Transitions using ISASS (An Interactive Speech Analysis-Synthesis System", Proc. ICASSP-80, pp. 146-149, 1980.

Chapanis, A., "Interactive human communication", Scientific American, Vol. 232, No. 3, pp. 36-49, 1975.

Chen, F.R. and Zue, V.W., "Application of Allophonic and Lexical Constraints in Continuous Digit Recognition", Proc. ICASSP-84, pp. 35.3.1-4, 1984.

Chigier, B. and Brennan, R.A., "Broad Class Network Generation Using A Combination Of Rules And Statistics For Speaker Independent Continuous Speech", Proc. ICASSP-88, pp. 449-452, 1988.

Chollet, G.F., Astier, A.B.P. and Rossi, M., "Evaluating the Performance of Speech Recognisers at the Acoustic-Phonetic Level", Proc. ICASSP-81, pp. 758-761, 1981.

Chomsky, N. and Halle, N., *The Sound Patterns of English*, New York, Harper & Row, 1968.

Chow, Y.L, Schwartz, R., Roucos, S., Kimball, O., Price, P., Kubala, F., Dunham, M., Krasner, M. and Makhoul, J., "The role of word-dependent coarticulatory effects in a phoneme-based speech recognition system", Proc. ICASSP-86, Apr. 1986.

Church, K.W., *Phonological Parsing in Speech Recognition*, Kluwer Academic Publishers, 1987.

Clark, J. and Colin, Y., *An Introduction to Phonetics and Phonology*, Basil Blackwell, 1990.

Cole, R.A. and Hou, L., "Segmentation and Broad Classification of Continuous Speech", Proc. ICASSP-88, pp. 453-456, 1988.

Cole, R.A. and Muthusamy, Y.K., "Perceptual studies on vowels excised from continuous speech", Proc. ICSLP'92, pp. 1091-1094, Banff, Canada, 1992.

Colla, A.M. and Sciarra, D., "Automatic Diphone Bootstrapping for Speaker-Adaptive Continuous Speech Recognition", Proc. ICASSP-84 (35.2), San Diego, 1984.

Colla, A.M., Scagliola, C. and Sciarra, D., "A Connected Speech Recognition System using a Diphone-Based Language Model", Proc. ICASSP-85 (31.9), Tampa, 1985.

Colla, A.M. and Sciarra, D., "Automatic Generation of Linguistic, Phonetic and Acoustic Knowledge for a Diphone-Based Continuous Speech Recognition System", in R. DeMori and C.Y. Suen (eds.), *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, NATO ASI Series, Vol. F16, pp. 361-387, Springer-Verlag, 1985.

Colla, A.M., "Some Considerations on the Definition of Sub-Word Units for a Template-Matching Speech Recognition System", Proc. Int. Symp. on Speech Recognition, pp. 55-56, Montreal 1986.

Colla, A.M., "Automatic Extraction of Acoustic Prototypes for Large Vocabulary Speech Recognition by Using Speaker-Independent Features", Proc. ICASSP-89, S3.4, Glasgow, 1989.

Colla, A.M., *On using sub-word units in ASR*, Notes from lecture series given at the Norwegian Institute of Technology, Trondheim 1989.

Cooper, L. and Cooper, M.W., *Introduction to Dynamic Programming* Pergamon Press Ltd. 1981.

Cosi, P., "Segmentation and Labelling of EUROM0 Italian Continuous Passage", Internal Report for ETR group of SAM, Document No. SAM-CSRF(CNR)-04, 1990.

Cosi, P., Falavigna, D. and Omologo, M., "A preliminary statistical evaluation of manual and automatic segmentation discrepancies", Proc. EUROSPEECH'91, pp. 693-697, Genova, Italy, 1991.

Cox, R.C. and Robinson, D.M., "Some Notes on Phase in Speech Signals", Proc. ICASSP-80, pp. 150-153, 1980.

Cravero, M., Pieraccini, R. and Raineri, F., "Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models", Proc. ICASSP-86, pp. 42.3.1-4, Tokyo, 1986.

Cravero, M., Pieraccini, R. and Raineri, F., "Definition of Recognition Units Through Two Levels of Phonemic Description", Proc. Int. Symp. on Speech Recognition, pp. 53-54, Montreal 1986.

Daaboul, F. and Adoul, J.P., "Parametric Segmentation of Speech into Voiced-Unvoiced-Silence Intervals", Proc. ICASSP-77, pp. 327-331, 1977.

Dalsgaard, P., "Semi-automatic phonemic labelling of speech data using a self-organising neural network", Proc. EUROSPEECH'89, pp. 541-544, 1989.

Dalsgaard, P. and Barry, W., "Acoustic-Phonetic Features in the framework of Neural-Network Multi-Lingual Label Alignment", Proc. of ICSLP'90, pp. 945-948, Kobe, Japan, 1990.

Dalsgaard, P., Andersen, O. and Barry, W., "Multi-Lingual Label Alignment using Acoustic-Phonetic Features derived by Neural-Network technique", Proc. ICASSP'91, pp. 197-200, Toronto, Canada, 1991.

Dalsgaard, P. and Andersen, O., "Identification of mono- and poly-phonemes using acoustic-phonetic features derived by self-organising neural network", Proc. ICSLP'92, pp. 547-550, Banff, Canada, 1992.

Das, S.K., "Some Experiments in Discrete Utterance Recognition", Proc. ICASSP-80, pp. 178-181, 1980.

Deleglise, P., Bimbot, F., Montacie, C., and Chollet, G., "Temporal decomposition and acoustic-phonetic decoding for automatic recognition of continuous speech", Proc. ICASSP'88, pp. 839-842, 1988.

Delegado-Martins, M.R. and Freitas, M.J., "Temporal structures of speech: "Reading news at TV"", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 19.1-5, Barcelona, 1991.

De Mori, R., Laface, P. and Piccolo, E., "Automatic detection and description of syllabic features in continuous speech", IEEE Trans. on ASSP 24:5, 365-79, 1976

De Mori, R. and Giordano, G., "A Parser for Segmenting Continuous Speech into Pseudo-Syllabic Nuclei", Proc. ICASSP-80, pp. 876-879, 1979.

Depuydt, L. and Martens, J.P., "Broad phonetic classification and segmentation of continuous speech by means of neural networks and dynamic programming", Proc. Speech Communication, Vol.10, No.1, pp.81-90, Feb. 1991.

Dew, D. and Jensen, P.J., *PHONETIC PROCESSING -The dynamics of speech*, Charles E. Merrill Publishing Company, 1977.

Diehl, R.L., "On segments and segment boundaries", Journal of Phonetics, pp. 289-290, 1987.

Dijk-Kappers, A.M.L.Van, *Temporal decomposition of speech and its relation to phonetic information*. Ph.D. Thesis, Technische Universiteit Eindhoven, 1989.

Dixon, N.R., "An Application Hierarchy for Heuristic Rules in Automatic Phonemic Segmentation of Continuous Speech", Proc. ICASSP-77, pp. 671-674, 1977.

Dixon, N.R. and Silverman, H.F., "What are the Significant Variables in Dynamic Programming for Discrete Utterance Recognition?", Proc. ICASSP-81, pp. 728-731, 1981.

Dours, C., Calmès, M., Kabré, H., Pécatte, J.M., Pérennou, G. and Vigouroux, N., "A Multi-Level Automatic Segmentation System: SAPHO and VERIPHONE", Proc. of EUROSPEECH'89, pp 83-86, Vol II, (S.29), Paris, 1989.

Duda, R.O. and Hart, P.E., *Pattern Classification and scene analysis*, John Wiley and Sons, 1973.

Dyhr, N., Andersen, O. and Dalsgaard, P. "Labelling criteria for the Danish EUROM.0 database", SAM-document no.: SAM-IES-055, 1992.

Ehara, T., Ogura, K. Morimoto, T., "ATR Dialogue Database", Proc. ICSLP'90, pp. 24.5.1-4, Kobe, Japan, 1990.

Elenius, K., "Comparing a Connectionist and a Rule Based Model for Assigning Parts-of-Speech", Proc. ICASSP-90, pp. 597-600, 1990.

Elman, J.L. and Zipser, D., "Learning the hidden structure of speech", JASA 83, pp. 1615-1626, April 1988.

Endresen, R.T., "Fonologi", pp. 38-75 in *Språkvitenskap, En elementær innføring*, Universitetsforlaget 1988, (in Norwegian).

Erp, A. Van and Boves, L., "Manual Segmentation and Labelling of Speech", Proc. Speech'88, pp. 1131-1138, Edinburgh, Scotland, 1988.

Erp, A. Van, Grice, M. and Barry, W., "Manual Labelling of Danish, Dutch, English and French Speech Material on EUROM-0", pp. 304-315, in *SAM, EXTENSION PHASE, FINAL REPORT, 1 April 1988 - 28 February 1989*.

Erp, A. Van, Houben, C., Barry, W., Grice, M., Boë, L.J., Braun, G., Cosi, P., Dyhr, N., Pérennou, G., Vigouroux, N. and Auteserre, D., "A Unified Approach to the Labelling of Speech : First Multilingual Results", Proc. of EUROSPEECH'89, pp. 88-91, Vol. II, (S.29), Paris, 1989.

Fant, G., "Analysis and synthesis of speech processes", in *Manual of Phonetics*, B. Malmberg (ed.) p. 223, North-Holland publishing Company, 1968.

Fant, G., *Speech Sounds and Features*, The MIT Press, 1973.

Fant, G., Kruckenberg, A. and Nord, L., "Some observations on tempo and speaking style in Swedish text reading", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 23.1-5, Barcelona, 1991.

Feng, G., et. al. "On-Line Speech Segmentation Using Adaptive Models: Application To Variable Rate Speech Coding", Proc. EUROSPEECH'91, pp. 705-708, 1991.

Fesseler, P. et. al : "Automatic Vocabulary Extension for a Speaker-Adaptive Speech Recognition System based on CVC Units", Proc. EUROSPEECH'89, pp. 75-79, Paris, 1989.

Fissore, L., Laface, P., and Micca, G., "Comparison of discrete and continuous HMMs in a CSR task over the telephone", Proc. ICASSP'91, pp. 253-256, 1991.

Flanagan, J.L., *Speech Analysis Synthesis and Perception*, Second, Expanded Edition, Springer-Verlag, 1972.

Flanagan, J.L., "Speech Technology and Computing: A Unique Partnership", Proc. EUROSPEECH'91, opening session, pp. 7-22, 1991.

Foldvik, A.K., "The change from apical to dorsal r in Norwegian". Proc. 11th International Congress of Phonetic Sciences, Tallinn, Vol.1, pp. 177-178, 1987.

Foldvik, A.K., Husby, O. and Kværness, J., "Magnetic Resonance Imaging", Proc. Speech'88, pp. 423-428, Edinburgh, 1988.

Foldvik, A.K., Private communication, 1992.

Fournol, D., et. al., "The SPIN Continuous-Speech Decoding System", Proc. EUROSPEECH'89, pp. 84-88, Paris, 1989.

Fourcin, A. "Assessment, Methodology and Standardisation in Multilingual Speech Technology", Proc. of the Annual ESPRIT Conference, pp. 3-14, Brussels, Nov. 1990.

Fourcin, A., "Linguistic engineering and speech assessment methods" in *Speech, hearing and language, work in progress 1991*, vol 5, University College London, Department of Phonetics and Linguistics, pp. 65-74, 1991.

Fu, K.S., *Syntactic Methods in Pattern Recognition*, Academic Press, New York, 1974.

Fujimura, O., Macchi, M.J. and Lovins, J.B., "Demisyllables and Affixes for Speech Synthesis", Proc. 9th ICA, Madrid #I-107, pp. 513, 1977.

Furui, S., *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, Inc., 1985.

Furui, S., "Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Transactions on ASSP, Vol. 34, pp. 52-59, Feb. 1986.

Furui, S., "On the Role of Spectral Transition for Speech Perception", JASA, Vol. 80, no. 4, pp. 1016-1025, Oct. 1986.

Furui, S., "Unsupervised Speaker Adaption Based on Hierarchical Spectral Clustering", Trans. on ASSP, Vol. 37, No. 12, pp. 1923-1930, Dec. 1989.

Gauvain, J-L., Lamel, L.F. and Eskenazi, M., "Design Considerations and Text Selection for BREF, a large French read-speech corpus", Proc. ICSLP'90, pp. 24.6.1-4, 1990.

Gibbon, D., "Survey of prosodic labelling for EC languages", Report e.6 in SAM-UCL G002, 1990

Glass, J.R. and Zue, V.W., "Detection of Nasalized Vowels in American English", Proc. ICASSP-85, pp. 1569-1572, 1985.

Glass, J.R. and Zue, V.W., "Multi-Level Acoustic Segmentation of Continuous Speech", Proc. ICASSP-88, pp. 429-432, 1988.

Gray, M.A. and Markel, J.D., "Distance Measures for Speech processing", Trans. on ASSP, Vol. ASSP-24, no. 5, pp. 380-391, Oct. 1976.

Gray, R.M., Buzo, A. and Gray Jr., A.H., "Distortion measures for Speech Processing", Trans. on ASSP, Vol. ASSP-28, no. 4, pp. 367-375, Aug. 1980.

Gray, R.M., "Vector Quantization", IEEE ASSP Magazine, pp. 4-29, April 1984.

Grice, M., Barry, W.J., and Fourcin, A., "Specification of EUROM0 assessment", Appendix B: Part2, in *Support available from SAM-project for other ESPRIT speech and language work*, SAM-document G001/B/2, 1989.

Grice, M. and Barry, W.J., "Levels of transcription", handout at ETR meeting, January, 1990.

Gupta, V.N., Gowdy, J.N. and Bryan, J.K., "Evaluation of some Distance Measures for Speaker Independent Isolated Word Recognition", Proc. ICASSP-77, pp. 460-463, 1977.

Gupta, V.N., Bryan, J.K. and Gowdy, J.N., "Speaker-Independent Vowel Identification in Continuous Speech", Proc. ICASSP-78, pp. 546-548, 1978.

Gupta, V., Lennig, M., Marcus, J. and Mermelstein, P., "Syllable Network for Phonemic Decoding of Speech", Proc. Int. Symp. on Speech Recognition, pp. 45-46, Montreal, 1986.

Haltsonen, S., Jalanko, M., Bry, K.-J. and Kohonen, T., "Application of Novelty Filter to Segmentation of Speech", Proc. ICASSP-78, pp. 565-568, 1978.

Haltsonen, S. and Bry, K.-J., "Automatic Selection of Phonemes from a Equally Spaced Quasi-Phoneme String by the Entropy Principle", Proc. ICASSP-79, pp. 108-111, 1979.

Haltsonen, S., "Improvement and Comparison of Three Phonemic Segmentation Methods of Speech", Proc. ICASSP-81, pp. 1160-1163, 1981.

Haltsonen, S. and Ruusunen, P., "Collection of Phoneme Samples using Time Alignment and Spectral Stationary of Speech Signals", Proc. ICASSP-85, pp. 1561-1564, 1985.

Hampshire II, J.B. and Waibel, A.H., "The Meta-Pi Network: Connectionist Rapid Adaption for High-Performance Multi-Speaker Phoneme Recognition", Proc. ICASSP-90, pp. 165-168, 1990.

Hanson, B.A. and Wakita, H., "Spectral Slope Distance Measures for all-pole models of speech", Proc. ICASSP-86, pp.765-768, 1986.

Hanson, B.A. and Wakita, H., "Spectral Slope Distance Measures with Linear Prediction Analysis for Word Recognition in Noise", IEEE Trans. on ASSP, Vol. 35, no. 7, July, 1987.

Hanson, B.A. and Applebaum, T.H., "Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech", Proc. ICASSP-93, pp. II.79-82, 1993.

Harmegines, B. and Poch-Olive, D., "Some aspects of vowel reduction in Spanish spontaneous speech", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 31.1-5, Barcelona, 1991.

Harrison, T.D. and Fallside, F., "A Connctionist Model for Phoneme Recognition in Continuous Speech", Proc. ICASSP-89, pp. 417-420, 1989.

Hatazaki, K. et.al., "Phoneme segmentation by an expert system based on spectrogram reading knowledge", Proc. of Speech'88, pp. 927-934, Edinburgh, 1988.

Haton, J.-P. and Sanchez, C., "An Experimental System for Acoustic-Phonetic Decoding of Continuous Speech", Proc. ICASSP-79, pp. 105-107, 1979.

Haton, J.-P. and Morel, O., "Automatic Recognition of Connected Digits Sequences by means of Segmentation and Dynamic Programming", Proc. ICASSP-79, pp. 245-248, 1979.

Haton, J.-P. and Damestoy, J.-P., "A Frame Language for the Control of Phonetic Decoding in Continuous Speech Recognition", ICASSP-85, pp. 1565-1568, 1985.

Heggstad, K., *Norsk frekvensordbok - De 10000 vanligste ord fra norske aviser*, (in Norwegian), Universitetsforlaget, 1982.

Hemert, J.P.Van, "Automatic segmentation of speech into diphones", Philips Technical Review, Vol. 43, No. 9, pp. 233-242, Sept. 1987.

Hermansky, H., "Perceptual linear predictive (PLP) analysis for speech", Proc. JASA, pp. 1738-1752, 1990.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P., "Compensation for the effect of the communication channel in the auditory-like analysis of speech (RASTA-PLP)" Proc. EUROSPEECH'91, pp. 1367-1370, 1991.

Hermansky, H. and Morgan, N., "Towards handling the Acoustic environment in spoken language processing", Proc. ICSLP'92, pp. 85-88, 1992.

Hermansky, H., Morgan, N., and Hirsch, H-G., "Recognition of speech in additive and convolutional noise based on RASTA spectral processing", Proc. ICASSP'93, pp. II 83-86, 1993.

Hirsch, H.G., Meyer, P. and Ruehl, H.W., "Improved speech recognition using high-pass filtering of subband envelopes", Proc. EUROSPEECH'91, pp.413-416, 1991.

Honda, M. and Shiraki, Y., "Very Low-Bit-Rate Speech Coding", in *Advances in Speech Signal Processing*, Edited by S. Furui and M.M. Sondi, Marcel Dekker, Inc., pp. 209-230, 1992.

Houben, C. and Hendriks, J., "Structure of and retrieval from an acoustic phonetic database", Proc. of Speech'88, pp. 1139-1146, Edinburgh, 1988.

Houben, C.G.J., "Automatic labelling of speech using an acoustic-phonetic knowledge base", Proc. EUROSPEECH'89, pp. 104-107, 1989.

Houben, C., "Semi-automatic labelling: A review", SAM document no. SAM-RNL-3, Dec. 1989.

Huang, X. and Lee, K.-F., "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition", Trans. on SAP, pp. 150-157, Vol.1, No. 2, April 1993.

Huber, D., "A Statistical Approach to the Segmentation and Broad Classification of Continuous Speech Into Phrase-Sized Information Units", Proc. ICASSP-89, pp. 600-603, 1989.

HucKvale, M., "Exploiting Speech Knowledge in Neural Nets for Recognition", Proc. in Speech Communication 9, Elsvier Science Publishers, pp. 1-13, 1990.

HucKvale, M., "The benefits of tired segmentation for the recognition of phonetic properties", Proc. EUROSPEECH'93, pp. 1473-1476, 1993.

Hunt, M.J., Lenning, M. and Mermelstein, P., "Experiments in syllable based recognition of continuous speech", Proc. ICASSP-80, pp. 880-883, Denver, Apr. 1980.

Hunt, M.J., "The speech signal", AGARD lecture series No. 170 on *Speech Analysis and Synthesis and Man-Machine Speech Communication for Air Operations*, 1990.

Husøy, P.O., *Forward Connected Artificial Neural Networks Applied To Automatic Speech Recognition*, Ph.D. Thesis, Norwegian Institute of Technology, 1991.

Huttenlocher, D.P. and Zue, V.W., "A Model of Lexical Access Based on Partial Phonetic Information", Proc. ICASSP 84, pp. 26.4.1-4, San Diego, 1984.

Huttenlocher, D.P. and Withgott, M., "On Acoustic versus Abstract Units of Representation", Proc. Int. Symp. on Speech Recognition, pp. 61-62, Montreal 1986.

Huttenlocher, D.P., "A Broad Phonetic Classifier", Proc. ICASSP-86, pp. 42.9.1-4, 1986.

Høhne, H.D., Coker,C., Levinson, S.E. and Rabiner, L.R., "On Temporal Alignment of Sentences of Natural and Synthetic Speech", Trans. on ASSP, Vol. ASSP-31, No. 4. pp. 807-813, Aug. 1983.

Høyland, A., *Sannsynlighetsberegning og statistisk metodelære*, Tapir, 1985.

IPA, "The International Phonetic Alphabet", Journal of the Phonetic Association, Vol. 19, no. 2, Dec. 1989.

Itakura, F. and Saito, S., "Analysis synthesis telephony based on the maximum likelihood method", Proc. 6th Int. Cong. Acoust., C-5-5, 1968.

Itakura, F. and Saito, S., "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies", Trans. Inst. Electron. Commun. Eng. (Japan), pp. 36-43, 1970.

Jacovitti, G., Pierucci, P. and Falaschi, A., "Speech segmentation and classification using higher order moments", Proc. EUROSPEECH'91, pp. 1335-1338, 1991.

Jaschul, J., "An Approach to Speaker Normalization for Automatic Speech Recognition", Proc. ICASSP-79, pp. 235-238, 1979.

Jayant, N.S. and Noll, P., *Digital Coding of waveforms - Principles and Applications to Speech and Video*, Prentice-Hall, 1984.

Jayant, N., "High-quality coding of telephone speech and wideband audio", pp. 85-108 in *Advances in Speech Signal Processing,* Edited by S. Furui and M.M. Sondi, Marcel Dekker, Inc. 1992.

Johnson, R.A. and Wichern D.W., *Applied multivariate statistical analysis*, Prentice-Hall International Editions, 1988.

Johnston, J.D., "Transform coding of audio signals using perceptual noise criteria", Proc. in IEEE Journal on selected areas in communications, Vol. 6, No. 2. pp. 314-323, 1988.

Johnston, J.D. and Brandenburg, K., "Wideband Coding - Perceptual Considerations for Speech and Music", pp. 109-140, in *Advances in Speech Signal Processing* Edited by S. Furui and M.M. Sondi, Marcel Dekker, Inc. 1992.

Juang, B.H. Rabiner, L.R., Levinson, S.E. and Sondhi, M.M., "Recent Developments in the Application of Hidden Markov Models to Speaker-Independent Isolated Word Recognition", Proc. ICASSP-85, pp. 9-12, 1985.

Juang, B.H. Rabiner, L.R. and Wilpon, J.G., "On the Use of Bandpass Liftering in Speech Recognition", Proc. ICASSP-86, pp. 14.18.1-4, 1986.

Juang, B.H. Rabiner, L.R. and Wilpon, J.G., "On the Use of Bandpass Liftering in Speech Recognition", Proc. on ASSP, Vol. 35, No.7, July, 1987.

Junqua, J.C., Wakita, H., and Hermansky, H., "Evaluation and optimization of the perceptually-based ASR front-end", Trans. on SAP, pp. 39-47, Vol.1, No. 1, Jan. 1993.

Kahn, D., "Syllable-Based Phonological rules and their Implications for Speech Recognition", Proc. Int. Symp. on Speech Recognition, pp. 43-44, Montreal 1986.

Kamm, C.A, Streeter, L.A., Kane-Esrig, Y and Burr, D., "Comparing performance of spectral distance measures and neural network methods for vowel recognition", in Computer, Speech and Language, pp. 21-34, 1989.

Kasuya, H. and Wakita, H., "Automatic Detection of Syllable Nuclei as Applied to Segmentation of Speech", Proc. ICASSP-77, pp. 652-655, 1977.

Kido, K., Miwa, J., Makino, S. and Niitsu, Y., "Spoken Word Recognition System for Unlimited Speakers", Proc. ICASSP-78, pp. 735-738, 1978.

Kimber, D.G., Bush, M.A. and Tajchman, G.N., "Speaker-Independent Vowel Classification Using Hidden Markov Models and LVQ2", Proc. ICASSP-90, pp. 497-500, 1990.

Kimberly, B.P. and Searle, C.L., "Automatic Discrimination of Fricative Consonants Based on Human Audition", Proc. ICASSP-79, pp. 89-92, 1979.

Klatt, D.H., "Prediction of perceived phonetic distance from critical band spectra: A first step", Proc. ICASSP-82, pp. 1278-1281, 1982.

Klatt, D.H., "Models of Phonetic recognition I: Issues that arise in Attempting to Specify a Feature-Based Strategy for Speech Recognition", Proc. Int. Symp. on Speech Recognition, pp. 63-66, Montreal, 1986.

Kohonen, T., Nemeth, G., Bry, K.-J., Jalanko, M., and Riittinen, H., "Spectral Classification of Phonemes by Learning Subspaces", Proc. ICASSP-79, pp. 97-100, 1979.

Kohonen, T. and Torkkola, K., "Using Self-Organizing Maps and Multi-Layered Feed-Forward Nets to Obtain Phonemic Transcriptions of Spoken Utterances", Proc. EUROSPEECH'89, pp- 561-564, 1989.

Kohonen, T., "The Self-Organizing Map", Proc. of the IEEE, Vol. 78, No. 9, pp. 1464-1479, Sept. 1990.

Komori, Y., Hatazaki, K., Tanaka, T., Kawabata, T.and Shikano, K., "Phoneme Recognition Expert System Using Spectrogram Reading Knowledge and Neural Networks", Proc. EUROSPEECH '89, pp. 549-552, Paris, 1989.

Komori, Y., Hatazaki, K., Tanaka, T. and Kawabata, T., "Combining Phoneme Identification Neural Networks Into an Expert System Using Spectrogram Reading Knowledge", Proc. ICASSP-90, pp. 505-508, 1990.

Kopp, J. and Lane, H.L., "Hue discrimination related to linguistic habits",
Proc. Psychonomic Science, 11, pp. 61-62, 1968.

Kuhl, P.K. and Padden, D.M., "Enhanced discriminability at the phonetic boundaries for the place of feature in macaques", JASA, 73, pp. 1003-1010, 1983.

Kvale, K., "Speech recognition based on acoustic subword units",
M.Sc. thesis, (in Norwegian), Norwegian Institute of Technology, 1987.

Kvale, K. and Foldvik, A.K., "Manual Segmentation and Labelling of Continuous Speech", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 37.1-5, Barcelona, 1991.

Kvale, K. and Foldvik, A.K., "The multifarious r-sound",
Proc. ICSLP'92, pp. 1259-1262, Banff, Canada, 1992.

Ladefoged, P., *A Course in Phonetics*, Harcourt Brace Jovanovich, Publishers, 1982.

Lagger, H. and Waibel, A., "A Coarse Phonetic Knowledge Source for Template Independent Large Vocabulary Word Recognition", Proc. ICASSP-85, pp. 862-865, 1985.

Lamel, L., Hassel, R.H. and Seneff, S., "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus",
Proc. DARPA Speech Recognition Workshop, pp. 100-109, 1989.

Lea, W.A. and Shoup, J.E., "Gaps in the Technology of Speech Understanding",
Proc. ICASSP-78, pp. 405-408, 1978.

Lea, W.A. and Shoup, J.E., "Specific contributions of the ARPA SUR project",
in W.A. Lea (ed.), *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, 1980.

Lee, C.-H., Soong, F.K. and Juang, B.H., "A Segment Model Based Approach to Speech Recognition", Proc. ICASSP-88, pp. 501-504, 1988.

Lee, C.-H., "Applications of Dynamic Programming to Speech and Language Processing",
AT&T Technical Journal, pp. 114-130, May/June, 1989.

Lee, C.-H., Juang, B.H., Soong, F.K., and Rabiner, L.R. "Word recognition using whole word and subword models", Proc. ICASSP-89, pp. 683-686, 1989.

Lee, C.-H., Giachin, E., Rabiner, L.R., Pieraccini, R., and Rosenberg, A.E.,
"Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition", Proc. ICASSP-91, pp. 161-164, 1991.

Lee, K.-F., *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPINX System*, Ph.D. Thesis, Carnegie Mellon University, Pittsburg Pennsylvania,
Kluwer Academic Publisher, 1988.

Lee, K.-F., "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition",
Trans. on ASSP, Vol.38, No.4, pp. 599-609, April, 1990.

Lee, Y.-T., "Information-Theoretic Distortion Measures for Speech Recognition",
Trans. on ASSP, Vol. 39, No. 2, pp. 330-335, Feb. 1991.

Lennig, M., Sharp, D., Kenny, P., Gupta, V., and Precoda, K., "Flexible vocabulary recognition of speech", Proc. ICSLP'92, pp. 93-96, Banff, Canada, 1992.

Leung, H.C. and Zue, V.W., "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech", Proc. ICASSP-84, pp. 2.7.1-4, 1984.

Leung, H.C., Hetherington, I.L. and Zue, V.W., "Speech recognition using stochastic explicit-segment modeling", Proc. EUROSPEECH'91, pp. 931-934, 1991.

Levinson, S.E., "Continuously variable duration hidden Markov models for speech recognition", in Computer, Speech, Language, vol.1, no.1, pp. 29-46, 1986.

Levinson, S.E., Ljolje, A. and Miller, L.G., "Large Vocabulary Speech Recognition Using a Hidden Markov Model for Acoustic/Phonetic Classification",
Proc. Speech Technology, pp. 26-32, Apr/May, 1989.

Levinson, S.E. and Roe, D.B., "A Perspective on Speech Recognition",
IEEE Communications Magazine, pp. 28-34, Jan. 1990.

Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C.,
"The discrimination of speech sounds within and across phoneme boundaries",
Journal of Experimental Psychology, 54, pp. 358-368, 1958.

Liberman, A.M. and Mattingly, I.G., "The motor theory of speech-perception revisited",
Cognition, 21, pp. 1-36, 1985.

Lindberg, B. and Danielsen S.W., "Specification of SESAM Assessment Workstation",
Appendix B in *Support available from SAM project for other ESPRIT Speech and Language Work*, Document: SAM-UCL-G001, 1989.

Linde, Y, Buzo, A and Gray, R.M., "An Algorithm for Vector Quantizer Design",
Proc. IEEE Trans. on Communications, Vol. com-28, No.1, pp. 84-95, Jan. 1980.

Lippmann, R.P., "An Introduction to Computing with Neural Nets",
Proc. IEEE ASSP Magazine, pp. 4-22, April 1987.

Lippmann, R.P., "Pattern Classification Using Neural Networks",
Proc. IEEE Communications Magazine, pp. 47-64, Nov. 1989.

Ljolje, A. and Riley, M.D., "Automatic segmentation and labelling of speech",
Proc. ICASSP'91, pp. 473-476, 1991.

Ljolje, A. and Riley, M.D., "Optimal speech recognition using phone recognition and lexical access", Proc. ICSLP'92, pp. 313-316, Banff, Canada, 1992.

Ljolje, A. and Riley, M.D., "Automatic segmentation of speech for TTS", Proc. EUROSPEECH'93, pp. 1445-1448, 1993.

Lourdes, A.C. et. al., "Analysis of the Spanish sequence "de" in content words and in function words in continuous speech", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 7.1-5, Barcelona, 1991.

Macchi, M.J., "A Phonetic Dictionary for Demisyllabic Speech Synthesis", Proc. ICASSP-80, pp. 565-567, 1980.

Malmberg, B. (ed.), *Manual of Phonetics*, North-Holland publishing company, 1968.

Maragos, P., "Nonlinear systems for speech signal processing", Lecture at the Norwegian Institute of Technology, Trondheim 1992.

Marchal, A., Hardcastle, W.J., Nicilaidis, N'guyen, N., and Gibbon, F., "Non-linear annotation of multi-channel speech data", Proc. ICSLP'92, pp. 787-790, Banff, Canada, 1992.

Marcus, S.M. and van Lieshout, R.A.J.M., "Temporal decomposition of speech", IPO Annual progress report, pp. 25-31, 1984.

Mariani, J.J. and Lienard, J.S., "Acoustic-Phonetic Recognition of Connected Speech using Transient Information", Proc. ICASSP-77, pp. 667-670, 1977.

Markel, J.D. and Gray, A.H., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

Mauturi, P., "Spontaneous speech and phonological models: the role of signal-independent information in linguistic communication", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 40.1-5, Barcelona, 1991.

Matsumoto, H. and Wakita, H., "Frequency Warping for Nonuniform Talker Normalization", Proc. ICASSP-79, pp. 566-569, 1979.

McCullough, D.P., "Variations on Itakura's Spectral Match Score", Proc. ICASSP-81, pp. 732-735, 1981.

Mergel, D. and Ney, H., "Phonetically Guided Clustering for Isolated Word Recognition", Proc. ICASSP-85, pp. 854-857, 1985.

Merialdo, B., "Speech Recognition with Very Large Size Dictionary", Proc. on ICASSP-87, pp. 364-367, 1987.

Mermelstein, P., "Automatic segmentation of speech into syllabic units", Proc. JASA, 58:4, pp. 880-883, 1975.

Morgan, N. and Bourlard, H., "Continuous Speech Recognition using Multilayer Perceptrons with Hidden Markov Models", Proc. ICASSP-90, pp. 413-416, 1990.

Myers, C.S., Rabiner, L.R. and Rosenberg, A.E., "An Investigation of the Use of Dynamic Time Warping for Word Spotting and Connected Speech Recognition", Proc. ICASSP-80, pp. 173-177, 1980.

Myers, C.S. and Rabiner, L.R., "Connected Word Recognition Using a Level Building Dynamic Time Warping Algorithm", Proc. ICASSP-81, pp. 951-955, 1981.

Myers, C.S. and Levinson, S.E., "Connected Word Recognition using a Syntax-Directed Dynamic Programming Temporal Alignment Procedure", Proc. ICASSP-81, pp. 956-959, 1981.

Nadas, A., Mercer, R.L., Bahl, L.R, Bakis, R., Cohen, P.S., Cole, A.G., Jelinek, F. and Lewis, B.L., "Continuous Speech Recognition with Automatically Selected Acoustic Prototypes Obtained by Either Bootstrapping or Clustering", Proc. ICASSP-81, pp. 1153-1155, 1981.

Nakatsui, M., "Half-Syllabic Units for Speech Processing - An automatic Segmentation", Proc. Int. Symp. on Speech Recognition, pp. 51-52, Montreal, July 1986.

Ney, H., "An Optimization Algorithm for Determining the Endpoints of Isolated Utterances", Proc. ICASSP-81, pp. 720-723, 1981.

Ney, H., "A Script-Guided Algorithm for the Automatic Segmentation of Continuous Speech", Proc. ICASSP-85, pp. 1209-1212, 1985.

Niles, L.T. and Silverman, H.F., "Combining Hidden Markov and Neural Network Classifiers", Proc. ICASSP-90, pp. 417-420, 1990.

Nocerino, N., Soong, F.K., Rabiner, L.R. and Klatt, D.H., "Comparative Study of Several Distortion Measures for Speech Recognition", Proc. ICASSP-85, pp. 25-28, 1985.

Nord, L., "Acoustic-phonetic studies in a Swedish speech data bank", Proc. of Speech'88, pp. 1147-1154, Edinburgh, Scotland, 1988.

Nord, L., "Report on manual labelling criteria used on the Swedish EUROM0 material", ETR-document, Sept. 1990.

Nordli, I.C., *Variasjon i artikulasjon av /k/ i norsk*, Master thesis at the University of Trondheim (in Norwegian), Dec. 1991.

Ohala, J.J., *What's Cognitive, What's Not in Sound Change*, Duisburg: L.A.U.D., 1990.

Olive, J.P, "A Scheme for Concatenating Units for Speech Synthesis", Proc. ICASSP-80, pp. 568-571, 1980.

Olive, J.P,, Roe, D.B. and Tschirgi, J.E., "Speech processing systems that listen, too", in "AT&T Technology, products, systems and services", vol. 6, no. 4, pp. 26-31, 1991.

Oppenheim, A.V. and Schafer, R.W., *Digital signal processing*, Prentice-Hall International Editions, 1975.

Orfanidis, S.J., *Optimum Signal Processing, An Introduction*, 2.nd Edition, McGraw-Hill Publishing Company, 1988.

O'Shaughnessy, D., Barbeau, L., Bernardi, D. and Archambault, D., "Diphone Speech Synthesis", Proc. Speech Communication 7, Elsevier Science Publishers B.V., pp. 55-65, 1988.

O'Shaughnessy, D., *Speech Communication, Human and Machine*, Addison-Wesley Publishing Company, 1990.

Ostendorf, M. and Roukos, S., "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition", Trans. on ASSP, Vol. 37, No. 12, pp. 1857-1869, Dec. 1989.

Ottesen, G.E., "An Automatic Diphone Segmentation System", Proc. EUROSPEECH'91, pp. 713-716, 1991.

Pallett, S.P., "Speech corpora and performance assessment in the DARPA SLS program", Proc. ICSLP'90, pp. 24.3.1-4, 1990.

Pan, K.C., Soong, F.K., Rabiner, L.R. and Bergh, A.F., "An Efficient Vector-Quantization Preprocessor for Speaker Independent Isolated Word Recognition", Proc. ICASSP-85, pp.874-877, 1985.

Perennou, G., et. al. "Phonetic-String Alignment for an automatic labelling of speech corpora", ESCA Workshop on Speech Input/Output Assessment and Speech Databases, sept. 1989.

Philips, M., Glass, J. and Zue, V., "Automatic learning of lexical representations for sub-word unit based speech recognition systems", Proc. EUROSPEECH'91, pp. 577-580, 1991.

Picone, J., "Continuous Speech Recognition Using Hidden Markov Models", Proc. IEEE ASSP Magazine, pp. 27-41, July 1990.

Pieraccini, R. and Rosenberg, A.E., "Automatic Generation of Phonetic Units for Continuous Speech Recognition", Proc. ICASSP-89, Vol. 1, pp.623-626 (S12.7), Glasgow, 1989.

Plotkin, E., Plotkin, N. and Polevoi, E.Y., "Recognition of Spoken Digits by Joint Segmentation of Envelopes of Two-Signal Transforms", Proc. ICASSP-77, pp. 456-459, 1977.

Poddar, P. and Rao, P.V.S., "Neural network based segmentation of continuous speech", Proc. ICSLP'90, pp. 1365-1368, Kobe, 1990.

Pols, L.C.W. and Olive, J.P., "Intelligibility if Consonants in CVC Utterances Produced by Dyadic Rule Synthesis", Proc. Speech Communication 2, Elsvier Science Publishers, pp. 3-13, 1983.

Pols, L.C.W., "How useful are speech databases for rule synthesis development and assessment?", Proc. ICSLP'90, pp. 28.3.1-4, 1990.

Rabiner, L.R. and Sambur, M.R., "Voiced-Unvoiced-Silence Detection Using Itakura LPC Distance Measure", Proc. ICASSP-77, pp. 323-326, 1977.

Rabiner, L.R. and Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

Rabiner, L.R., Levinson, S.E., Rosenberg, A.E. and Wilpon, J.G., "Speaker Independent Recognition of Isolated Words Using Clustering Techniques", Proc. ICASSP-79, pp. 574-577, 1979.

Rabiner, L.R., Wilpon, J.G. and Rosenberg, A.E., "Application of Isolated Word Recognition to a Voice Controlled Repertory Dialer System", Proc. ICASSP-80, pp. 182-185, 1980.

Rabiner, L.R. and Schmidt, C.E., "A Connected Digit Recognizer Based on Dynamic Time Warping and Isolated Digit Templates", Proc. ICASSP-80, pp. 194-198, 1980.

Rabiner, L.R. and Wilpon, J.G., "Isolated Word Recognition Using A Two-Pass Pattern Recognition Approach", Proc. ICASSP-81, pp. 724-727, 1981.

Rabiner, L.R., Rosenberg, A.E., Wilpon, J.G. and Zampini, T.M., "A bootstrapping training technique for obtaining demisyllable reference patterns", JASA, Vol. 71, No. 6, pp. 1588-1595, 1982.

Rabiner, L.R., Wilpon, J.G. and Juang, B-L., "A Segmental K-Means Training Procedure for Connected Word Recognition", AT&T Tech. Journal, Vol. 65, Issue 3, pp. 21-31, May/June, 1986.

Rabiner, L.R., Wilpon, J.G. and Soong, F.K., "High performance connected digit recognition using hidden Markov models", Proc. ICASSP-88, pp. 119-122, Apr. 1988.

Rabiner, L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. of the IEEE, Vol.77, No. 2, pp. 257-285, Feb. 1989.

Rabiner, L.R., "Large vocabulary speech recognition", Lectures at the Norwegian Institute of Technology, Trondheim, May 1989.

Rabiner, L.R. and Juang, B.-H., *Fundamentals of speech recognition*, Prentice-Hall, 1993.

Reddy, R. and Watkins, R., "Use of Segmentation and Labelling in Analysis-Synthesis of Speech", Proc. ICASSP-77, pp. 28-32, 1977.

Repp, B.H., "Categorical perception: issues, methods, findings", in N.J. Lass (ed.), *Speech and language: advances in basic research and practice*, (vol 10), New York: Academic Press, 1984.

Roach, P., Roach, H., Dew, A. and Rowlands, P., "Phonetic analysis and the automatic segmentation and labelling of speech sounds", Working Papers in Linguistics and Phonetics 5, 54-62, Dept. of Linguistics and Phonetics, Leeds, 1990.

Rosenberg, A.E., Rabiner, L.R, Levinson, S.E. and Wilpon, J.G., "A Preliminary Study on the Use of Demisyllables in Automatic Speech Recognition", Proc. ICASSP-81, pp. 967-970, 1981.

Rosenberg, A.E., Rabiner, L.R, Wilpon, J.G. and Kahn, D., "Demisyllable-based Isolated Word Recognition", IEEE Trans. ASSP-31, 3, pp. 713-725, Jun. 1983

Rosenberg, A.E., "Recognition Error Measurements from Parameterized Distance Distributions", Proc. ICASSP-85, pp. 870-873, 1985.

Roucos, S., Schwartz, R. and Makhoul, J., "Segment Quantization for Very Low Rate Speech Coding", Proc. ICASSP'82, pp. 1565-1568, 1982.

Roucos, S., Ostendorf, M., Gish, H., and Derr, A., "Stochastic segment modeling using the estimate-maximize algorithm", Proc. ICASSP-88, pp. 127-130, 1988.

Ruske, G. and Schotola, T., "An approach to speech recognition using syllabic decision units", Proc. ICASSP-78, pp. 722-725, 1978.

Ruske, G. and Schotola, T., "The Efficiency of Demisyllable Segmentation in the Recognition of Spoken Words", Proc. ICASSP-81, pp. 971-974, 1981.

Ruske, G., "Automatic recognition of syllabic speech segments using spectral and temporal features", Proc. ICASSP-82, pp. 550-553, 1982.

Ruske, G. "Demisyllables as processing units for automatic speech recognition and lexical access", in R. DeMori and C.Y. Suen (eds.), *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, NATO ASI Series, Vol. F16, pp. 593-611, Springer-Verlag, 1985.

Ruske, G., "Experiments on the Use of Demisyllable for Automatic Speech Recognition", Proc. Int. Symp. on Speech Recognition, pp. 49-50, Montreal 1986.

Sagayama, S., "Phoneme Environment Clustering for Speech Recognition", Proc. ICASSP-89, (S8.3), pp. 397-400, 1989.

Sagisaki, Y., Takeda, K., Abe, M., Katagiri, S., Umeda, T. and Kuwabara, H., "A large-scale Japanese Speech Database", Proc. ICSLP'90, pp. 24.4.1-4, 1990.

Sauter, L.C., "Isolated Word Recognition using a Segmental Approach", Proc. ICASSP-85, pp. 850-853, 1985.

Scagliola, C., et.al., "Iterative Optimization of sub-word templates for speech recognition", Proc. EUROSPEECH'89, pp. 79-83, (S.5), Paris, 1989.

Schafer, R.W. and Markel, J.D. (Editors), "*Speech Analysis*", IEEE, John Wiley & sons, Inc. 1979.

Schmidt, M.S. and Watson, G.S., "The evaluation and labelling of automatic speech segmentation", Proc. EUROSPEECH'91, Vol. II, pp. 701-704, Genova, 1991.

Schotola, T., "On the use of demisyllables in automatic speech recognition", Proc. in Speech Communication no 3, Elsevier Science Publishers, pp. 63-87, 1984

Schwarts, R., Klovstad, J., Makhoul, J. and Sorensen, J., "A Preliminary Design of a Phonetic Vocoder Based on a Diphone Model", Proc. ICASSP-80, Denver, pp. 32-35, 1980.

Schwartz, R. and Roucos, S., "A comparison for 300-400 b/s vocoders", Proc. ICASSP-83, pp. 69-72, 1983.

Schwartz, R., Chow, Y.L, Roucos, S., Krasner, M., and Makhoul, J., "Improved hidden Markov modelling phonemes for continuous speech recognition", Proc. ICASSP-84, 1984.

Schwartz, R., Chow, Y.L, Kimball, O., Roucos, S., Krasner, M., and Makhoul, J., "Context-dependent modeling for acoustic-phonetic recognition of continuous speech", Proc. ICASSP-85, Apr. 1985.

Seneff, S., "Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model", Proc. ICASSP-84, pp. 36.2.1-4, 1984.

Seneff, S. and Zue, V.W., "Transcription and Alignment of the TIMIT Database", in *Getting Started With The DARPA TIMIT CD-ROM*, chap.4, (Unpublished manuscript: "To be distributed with the TIMIT database by NBS"), 1988.

Sherwood, T. and Fuller, H., *Guide to EUROM.1 Speech Database*, SAM-NPL-102, April, 1992.

Shiraki, Y. and Honda, M., "LPC speech coding based on variable-length segment quantization", Trans. on ASSP, Vol.36, no.9, pp. 1437-1444, 1988.

Shiraki, Y. and Honda, M., "Speaker adaption algorithms based on piece-wise moving adaptive segment quantization method", Proc. ICASSP'90, pp. 657-660, 1990.

Shoup, J.E., "Phonological aspects of speech recognition", pp. 125-138, in W.A. Lea (Editor), *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, 1980.

Silverman, H.F. and Morgan, D.P., "The Application of Dynamic Programming to Connected Speech Recognition", IEEE ASSP Magazine, pp.7-25, July 1990.

Silverman, K. et. al., "TOBI: A standard for labelling English prosody", Proc. ICSLP'92, pp. 867-870, Banff, Canada, 1992.

Sivertsen, E., *Fonologi*, Universitetsforlaget, 1988 (in Norwegian).

Socolf, M. and Zue, V.W., "Collection and Analysis of Spontaneous and Read Corpora for Spoken Language System Development", Proc. ICSLP'90, pp. 24.8.1-4, 1990.

Soong, F.K. and Rosenberg, A.E., "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", Proc. ICASSP-86, pp. 17.5.1-4, 1986.

Soong, F.K., "A Training Procedure for a Segment-Based-Network Approach to Isolated Word Recognition", Proc. ICASSP-87, pp. 17.1.1-4, 1987.

Soong, F.K., Rosenberg, A.E. and Juang, B.-H., "A Vector Quantization Approach to Speaker Recognition", AT&T Tech. Journal, Vol. 66, Issue 2, pp. 14-26, March/April, 1987.

Soong, F.K. and Sondhi, M.M., "A Frequency-Weighted Itakura Spectral Distortion Measure and Its Application to Speech Recognition in Noise",
IEEE Trans. on ASSP, Vol.36, No. 1, Jan. 1988.

Soong, F.K., "A Phonetically Labelled Acoustic Segment (PLAS) Approach to Speech Analysis-Synthesis", Proc. ICASSP-89, pp. 584-587, 1989.

Stanton, B.J., Jamieson, L.H., and Allen, G.D., "Robust recognition of loud and Lombard speech in the fighter cockpit environment", Proc. ICASSP-89, pp. 675-678, 1989.

Stevens, K.N., "Models of Phonetic Recognition II: An Approach to Feature-Based Recognition", Proc. Int. Symp. on Speech Recognition, pp. 67-68, Montreal, 1986.

Sugiyama, M., "Overview of ATR Interpreting Telephony Research Labs - Speech research activities and recent topics in neural network studies",
Lecture given at the Norwegian Institute of Technology, Trondheim, Sept. 1992.

Svendsen, T. and Soong, F.K., "On the Automatic Segmentation of Speech",
Proc. ICASSP-87, pp. 3.4.1-4, 1987.

Svendsen, T., Paliwal, K.K., Harborg, E. and Husøy, P.O., "An Improved Sub-Word Based Speech Recognizer", Proc. ICASSP-89, pp. 108-111, 1989.

Svendsen, T. and Kvale, K., "Automatic alignment of phonemic labels with continuous speech". Proc. ICSLP-90, Vol. II, pp. 997-1000, Kobe, Japan, Nov. 1990.

Svendsen, T., "Efficient quantization of speech spectral information",
Proc. EUROSPEECH'93, pp. 1143-1146, 1993.

Sørensen, H.B.D. and Dalsgaard, P., "Multi-level segmentation of natural continuous speech using different auditory front-ends", Proc. EUROSPEECH'89, pp. 79-82, 1989.

Tanaka, K., Hayamizu, S. and Otha, K., "The ETL speech database for speech analysis and recognition research", Proc. ICSLP'90, pp. 24.7.1-4, 1990.

Tappert, C.C., Suen, C.Y. and Wakahara, T., "The State of the Art in On-Line Handwriting Recognition", Trans. on ASSP, Vol. 12, No. 8, pp. 787-808, Aug. 1990.

Taylor, I., *Psycholinguistics - Learning and using language*,
Prentice-Hall International,Inc., 1990.

Taylor, P.A. and Isard, S.D., "Automatic Diphone Segmentation",
Proc. EUROSPEECH'91, pp. 709-711, 1991.

Tokhura, Y., "A Weighted Cepstral Distance Measure for Speech Recognition",
Trans. on ASSP, Vol. 35, pp. 1414-1422, 1987.

Thompsen, H.S., "Hill climbing to improve the performance of rule-based segmentation and labelling", Proc. EUROSPEECH'89, pp. 92-95, 1989.

Torkkola, K., "Automatic alignment of speech with phonetic transcriptions in real time",
Proc. ICASSP-88, pp. 611-614, 1988.

Tribolet, J.M., Rabiner, L.R. and Sondhi, M.M., "Statistical Properties of an LPC Distance Measure", Proc. ICASSP-79, pp. 739-743, 1979.

Tsopanoglou, A., Mourjopoulos, J. and Kokkinakis, G., "Continuous Speech Phoneme Segmentation Method Based on the Instantaneous Frequency",
Proc. EUROSPEECH'89, pp. 67-70, (S.29), 1989.

Vernooij, G.J., Bloothhooft, G. and van Holsteijn, Y., "A Simulation Study on the Usefulness of Broad Phonetic Classification in Automatic Speech Recognition",
Proc. ICASSP-89, pp. 85-88, 1989.

Vicenzi, C. and Sciarra, D., "Using Diphones in Large Vocabulary Word Recognition",
Proc. Int. Symp. on Speech Recognition, pp. 47-48, Montreal 1986

Vidal, E. and Marzal, A., "A review and new approaches for automatic segmentation of speech signals", Proc. in Signal processing V: Theories and Applications, Elsevies Science Publishers B.V., pp. 43-53, 1990.

Vieregge, W.H., "Evaluating the transcription process",
Proc. Speech'88, pp. 73-80, Edinburgh, 1988.

Vigououx, N. et.al., "Initial labelling experiments on the BDSON corpora",
Proc. of Speech'88, pp. 1155-1162, Edinburgh, Scotland, 1988.

Viterbi, A.J., "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm" in IEEE Trans. on Information Theory, Vol, IT-13, No. 2, pp. 260-269, 1967.

Wagner, M., "Automatic Labelling of Continuous Speech with a given Phonetic Transcription Using Dynamic Programming Algorithms", Proc. ICASSP-81, pp. 1156-1159, 1981.

Weigel, W., and Ruske, G., "Continuous speech recognition using syllabic segmentation and demisyllable Hidden Markov Models", Proc. EUROSPEECH'89, pp 17-20, (S.1), Vol I, Paris, 1989.

Weinstein, J.W., "Opportunities for Advanced Speech Processing in Military Computer-Based Systems", Proc. IEEE, Vol.79, No. 11, Nov. 1991.

Wells, J.C., Barry, W. and Fourcin, A.J., "Methods of Transcription and Labelling and Normative Reference", pp. 176-183, in *ESPRIT PROJECT 1541, SAM. Final Report; Definition Phase: 2.2.87-31.1.88*, 1988.

Wells, J.C., et al, "Specification of SAM Phonetic Alphabet (SAMPA)", Appendix A in *ESPRIT PROJECT 1541/2589 (SAM): Support available from SAM project for other ESPRIT Speech and Language work*, July, 1989.

Wells, J.C., et al, "Standard Computer-Compatible Transcription", SAM stage report sen. 3, SAM-UCL-037, in *ESPRIT PROJECT 2589 (SAM): Final Report; Year Three; 1.3.91-28.2.92*, 1992.

White, G.M., "Dynamic Programming, Viterbi Algorithm, and Low Cost Speech Recognition", Proc. ICASSP-78, pp. 413-417, 1978.

Wilpon, J.G. and Rabiner, L.R., "A Modified K-Means Clustering Algorithm for use in Isolated Word Recognition", IEEE Trans. on ASSP, Vol. 33, pp. 587-594, Jun. 1985.

Wilpon, J.G., Huang, B.H. and Rabiner, L.R., "An Investigation on the Use of Acoustic Sub-Word Units for Automatic Speech Recognition", Proc. ICASSP-87, pp. 20.7.1-4, 1987.

Wilpon, J.G., Rabiner, L.R., Lee, C.-H, and Goldman, E.R., "Automatic recognition in unconstrained speech using Hidden Markov Models", IEEE Trans. on ASSP, Vol. 38, no. 11, Nov. 1990.

Wilpon, J.G., Lee, C.-H., and Rabiner, L.R., "Improvements in connected digit recognition using higher order spectral and energy features", Proc. ICASSP-91, pp. 349-352, 1991.

Young, S.J., "HTK: Hidden Markov Model Toolkit v1.4, User Manual" and "HTK: Hidden Markov Model Toolkit v1.4, Reference Manual", Cambridge University Engineering Department, Speech Group, Sept. 1992.

Yegnanarayana, B. and Reddy, D.R., "A Distance Measure Based on the Derivative of the Linear Prediction Phase Spectrum", Proc. ICASSP-79, pp. 744-747, 1979.

Yuschik, M. and Martens, H., "An Evaluation of Hierarchical Features in Speech Recognition", Proc. ICASSP-81, pp. 979-982, 1981.

Zahorian, S.A. and Jagharghi, A.J., "Speaker normalization of static and dynamic vowel spectral features", Proc. JASA, Vol. 90, No.1, pp. 67-75, July 1991.

Zeiliger, J. and Serignat, J.F., "EUROPEC software v4.0, User's guide", App. R8 in ESPRIT project 2589 (SAM), Interim report 1990-1991, Document: SAM-UCL-G003, 1991.

Zelinski, R. and Class, F., "A Segmentation Procedure for Connected Word Recognition based on Estimation Principles", Proc. ICASSP-81, pp. 960-963, 1981.

Zue, V.W., "The Use of Speech Knowledge In Automatic Speech Recognition", Proc. of the IEEE. Vol 73., No. 11, pp. 1602-1615, Nov. 1985.

Zue, V.W., "Models of Phonetic Recognition III: The Role of Analysis-by-Synthesis in Phonetic Recognition", Proc. Int. Symp. on Speech Recognition, pp. 69-70, Montreal, 1986.

Zue, V.W., Glass, J., Phillips, M. and Seneff, S., "Acoustic segmentation and phonetic classification in the SUMMIT system", Proc. ICASSP'88, pp. 389-392, 1988.

Zue, V.W., *Speech Spectrogram Reading - An Acoustic Study of English Words and Sentences*, Course at University of Edinburgh, 29 May - 2 June, 1989.

Zue, V.W., Glass, J., Goodine, D., Philips, M. and Seneff, S., "The SUMMIT Speech Recognition System: Phonological Modeling and Lexical Access", Proc. ICASSP'90, pp. 49-52, 1990.

Zue, V.W., Glass, J., Goodine, D., Hirshman, L., Leung, H., Philips, M., Joseph, P. and Seneff, S., "The MIT ATIS system: Preliminary development, spontaneous speech data collection, and performance evaluation", Proc. EUROSPEECH'91, pp. 537-540, 1991.

Zwicker, E., Terhardt, E. and Paulus, E., "Automatic speech recognition using psychoacoustic models", JASA, 65:2, 487-498, 1979.