Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

Julie Tvergrov

# Analyzing Automatic Pronunciation Assessment Performance on Norwegian Child Speech

Master's thesis in Electronics System Design and Innovation
Supervisor: Giampiero Salvi
Co-supervisor: Xinwei Cao
June 2024

NTNU
Norwegian University of
Science and Technology

Julie Tvergrov

# Analyzing Automatic Pronunciation Assessment Performance on Norwegian Child Speech

Master's thesis in Electronics System Design and Innovation
Supervisor: Giampiero Salvi
Co-supervisor: Xinwei Cao
June 2024

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

NTNU
Norwegian University of
Science and Technology

Julie Tvergrov

# Analyzing Automatic Pronunciation Assessment Performance on Norwegian Child Speech

Master's thesis in Electronics System Design and Innovation

Student: Julie Tvergrov
Supervisor: Giamperio Salvi
Co-Supervisor: Xinwei Cao
Spring 2024

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

NTNU

# ABSTRACT

This thesis explores pronunciation difficulties among native and Second Language (L2) child speakers of Norwegian and assesses the performance of Automatic Pronunciation Assessment (APA) systems in this context. A detailed theoretical framework on speech characteristics and elements of traditional and deep learning-based Automatic Speech Recognition (ASR) systems are presented. A diverse methodological approach is employed, analyzing linguistic, developmental, and acoustic influences on pronunciation, as well as performing detailed analysis on prediction errors of APA system, and training new multitask models for use in Computer-Assisted Language Learning (CALL) systems.

The Teflon dataset, which contains single-word utterances of both native and non-native child speakers, provides the basis for this work, and the wide range of annotations makes it possible to analyze both pronunciation difficulties and pronunciation errors related to speaker age, first language, speech recording quality, target words, and phoneme pronunciation. The main focus is how such variables affect APA prediction errors. The observed results demonstrate the significant influence of language background and environmental factors on pronunciation and prediction errors, as well as highlight the challenges of scarce datasets and biased data representation, recognizing the limitations of the APA system in handling diverse phonetic inputs. The APA model performed better for speakers whose first languages share phonetic similarities with Norwegian and where there were several other speakers with the same first language background.

A new multitask ASR and APA model was trained using the Combined Short model from the Scribe project. Improvements in several metrics were achieved, and one specific fold got results of 9.23% WER, 3.12% CER, 59.54% ACC and 41.05% UAR. This progress confirms the possibility of continued improvement of APA for foreign language learners.

The thesis concludes with a discussion regarding key challenges and findings and presents natural next steps following these results. Expansion of the dataset is recommended to represent each age group and first language better, allowing for a more detailed analysis of pronunciation and prediction error trends without being overshadowed by speaker-dependent proficiency. The continued development of Norwegian datasets and models is essential for developing more robust ASR and APA systems that can effectively aid language learning for a broader range of child speakers in increasingly multilingual environments. By enhancing comprehension of the factors affecting speaker pronunciation and prediction errors of APA systems, this research contributes to the field of CALL.

# SAMMENDRAG

Denne masteroppgaven utforsker uttalevansker blant barn med og uten Norsk som morsmål, og vurderer i denne sammenhengen ytelsen til systemer for automatisk uttalevurdering (engelsk: APA). Et detaljert teoretisk rammeverk om talekarakteristikker og elementer i tradisjonelle og dyp-læringsbaserte systemer for automatisk talegjenkjenning presenteres. En variert metodisk tilnærming benyttes, der lingvistiske, utviklingsmessige og akustiske påvirkninger på uttale analyseres. I tillegg utføres detaljerte analyser av prediksjonsfeil i APA-systemet og nye fleroppgavemodeller trenes til bruk i systemer for datamaskinassistert språklæring (engelsk: CALL).

Teflon-datasettet, som inneholder enkeltord-uttalelser fra både barn med og uten Norsk som morsmål, danner grunnlaget for dette arbeidet. Det brede spekteret av annotasjoner gjør det mulig å analysere både uttalevansker og prediksjonsfeil knyttet til talerens alder, morsmål, kvaliteten på taleopptaket, mål ord og fonemuttale. Hovedfokuset er på hvordan slike variabler påvirker APA-prediksjonsfeil. De observerte resultatene demonstrerer betydelig innflytelse fra språkbakgrunn og miljøfaktorer på uttale og prediksjonsfeil, samt fremhever utfordringene med sparsomme datasett og skjev datarepresentasjon, i tillegg til å anerkjenne begrensningene i APA-systemet i håndtering av forskjellige fonetiske inndata. APA-modellen presterte bedre for talere hvis morsmål deler fonetiske likheter med norsk, og der det var flere andre talere med samme morsmål.

En ny kombinert ASR- og APA-modell ble trent ved å bruke Combined Short-modellen fra Scribe-prosjektet. Forbedringer i flere målinger ble oppnådd, og en spesifikk iterasjon oppnådde resultater på 9.23% WER, 3.12% CER, 59.54% ACC og 41.05% UAR. Denne fremgangen bekrefter muligheten for fortsatt forbedring av APA for fremmedspråklige barn. Avhandlingen avsluttes med en diskusjon rundt hovedutfordringer og funn, samt presenterer naturlige neste skritt. Det anbefales å utvide datasettet for å få bedre representasjon av hver aldersgruppe og morsmål, noe som ville muliggjøre en mer detaljert analyse av trender innen uttale og prediksjonsfeil uten å bli overskygget av taleravhengig språkkompetanse. Den fortsatte utviklingen av norske datasett og modeller er avgjørende for å utvikle mer robuste ASR- og APA-systemer som effektivt kan støtte språklæring for et bredere spekter av barn i stadig mer flerspråklige miljøer. Ved å forbedre forståelsen av faktorene som påvirker talerens uttale og prediksjonsfeil i APA-systemer, bidrar denne forskningen til utvikling av CALL.

# PREFACE

This thesis is a continuation of my specialization project from the fall of 2023. It has been an incredibly exciting journey to delve deeply into speech technology, engaging with new technologies while also understanding their relationship with traditional methods.

I would like to extend my acknowledgment to my supervisor, Giampiero Salvi, for his guidance and the interesting conversations about the challenges and opportunities in speech technology research in Norway and the Nordic region.

I am also very grateful to my co-supervisor, Xinwei Cao, for his assistance in helping me navigate the technical challenges encountered during this research.

This work not only advanced my understanding of complex technological concepts but also solidified my passion for the field of artificial intelligence, and I look forward to learning more through my professional career closely related to AI.

I hope this thesis contributes to the field and inspires further research in the interesting area of speech technology.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACRONYMS

**ACC** Accuracy.

**AI** Artificial Intelligence.

**APA** Automatic Pronunciation Assessment.

**ASR** Automatic Speech Recognition.

**CALL** Computer-Assisted Language Learning.

**CER** Character Error Rate.

**CNN** Convolutional Neural Network.

**CTC** Connectionist Temporal Classification.

**CV** Cross Validation.

**DNN** Deep Neural Network.

**GMM** Gaussian Mixture Models.

**HMM** Hidden Markov Models.

**KB** National Library of Sweden.

**L2** Second Language.

**LLM** Large Language Models.

**MAE** Mean Absolute Error.

**MFCC** Mel-Frequency Cepstrum Coefficients.

**NbAiLab** National Library of Norway AI Lab.

**NLP** Natural Language Processing.

**NPSC** Norwegian Parliamentary Speech Corpus.

**NST** Nordisk Språkteknologi.

**NTNU** Norwegian University of Science and Technology.

**STFT** Short-Time Fourier Transform.

**Teflon** Technology-Enhanced Foreign and Second-Language Learning of Nordic Languages.

**UAR** Unweighted Average Recall.

**WER** Word Error Rate.

CHAPTER

# ONE

# INTRODUCTION

## 1.1 Motivation

The use of speech technology has advanced significantly in recent years and has become an integral part of everyday life for many people. Speech-based systems can now not only recognize and verify speech but also interpret meaning and take actions based on speech. Additionally, they provide crucial communication aids for individuals with disabilities.

The capabilities of these systems are ever-evolving, especially with the rise of new Artificial Intelligence (AI) models that can handle enormous amounts of data; the potential seen from Large Language Models (LLM) now propagates into other fields. This progress can be observed in the field of speech technology, where the same benefits used for text-based input can be leveraged for speech-based input. By employing models that can learn variations and generalizations, cluster languages, and understand pronunciation patterns, smarter speech technology systems with multilingual understanding can be developed. These systems can comprehend and distinguish spontaneous dialogues among friends, assess pronunciation in speech, and facilitate successful transfer learning to low-resource languages. Ongoing research and development in these areas are expected to bring about new discoveries that will fundamentally change the way we interact with technology and one another in the future.

Given the rapid technological advancements, it is important to acknowledge that access to them will not be evenly distributed among different populations. Therefore, research focusing on low-resource languages must keep pace with new innovations. These languages usually lack extensive datasets and are understudied. Despite recent advancements in speech technology and Natural Language Processing (NLP), there is still a need to develop systems specifically tailored for these languages.[1] Ultimately, technology should be used to address global inequalities and provide solutions that ensure greater accessibility, enabling more people to benefit from these systems.

## 1.2 Project description

This work was carried out in conjunction with the Technology-Enhanced Foreign and Second-Language Learning of Nordic Languages (Teflon) project. This initiative is a collaboration among universities in the Nordics, focusing on making Computer-Assisted Language Learning (CALL) systems accessible to immigrant children.[2] Using gamification to engage and motivate children, advanced speech recognition robust enough for Second Language (L2) learners in Nordic languages is used to assess the pronunciation of the speakers and provide helpful feedback. [3] Making this available digitally will make such tutoring available to more children, as tutoring is traditionally done one-on-one with language teachers and speech therapists, and allows for implementation in both classroom teaching and individual remote learning. [4, 5]

To develop this language learning game, the Teflon project conducted an extensive data collection of single-word utterances of both native and non-native children.[6] This was done for both Norwegian, Swedish, and Finnish and was annotated at a highly detailed level, containing an assessment of word pronunciation on both word level and phoneme level for Norwegian. The task of building such a dataset is challenging in and of itself, given the speaker group, and is scarce even for traditional Automatic Speech Recognition (ASR) tasks, not to mention for the use in Automatic Pronunciation Assessment (APA). In addition to the scores, the human annotators noted information on prosody, pre-speech noise, noise/disturbances, and repetition in all recordings. Background information on each speaker, such as age, what languages they are proficient in, and the amount of time they had lived in the respective Nordic country, was also collected. Resulting in a highly valuable dataset used for training multitask models to both transcribe utterances and perform pronunciation assessments within the game.

In this thesis, we use preliminary results from the Teflon project to take a deep dive into the results of an APA system and look for correlations between the dataset annotation and the model's performance. The existing analyses have only focused on word-level scoring used for ASR and APA multitask training. However, given the detailed dataset available, an extensive range of information has still not yet been explored. Doing a thorough analysis of these results and corresponding annotations will further help the development of an APA system and can give insight into the aspects that affect the performance of the system.

In addition, we inspect the transcribing performance of different Norwegian models that can serve as the base model in a new ASR/APA system and need further fine-tuning. The existing results from the Teflon project were produced using wav2vec2.0 [7] based models, fine-tuned on each Nordic language. For the Norwegian data, models from the AI lab at the National Library of Norway were used as a base model.[8]. In this work, we re-evaluate existing open-source wav2vec2.0 models from Facebook [9], Scribe project [10], and NbAiLab [11] on the ASR task, to assess if these new models could serve as a better base model for the APA task. Through this work, the difficulty of the ASR and APA task on children's speech and the importance of specifically trained models is highlighted.

This thesis will seek to answer the following three research questions:

- **What factors influence the pronunciation difficulties of native and L2 child speakers of Norwegian?**
- **What factors impact the prediction error of the APA system?**
- **Can the performance of the multitask ASR and APA system be improved by adopting another base model that has been fine-tuned on Norwegian data?**

The structure of the thesis is divided into six chapters. **Chapter 2** reiterates the theoretical concepts related to speech, human and automatic pronunciation assessment, and both traditional and deep learning-based ASR systems alongside the evaluation metrics used in ASR and APA. **Chapter 3** presents the methodology, introducing the Teflon dataset and describing the multitask training and implementation strategies for the conducted experiments. **Chapter 4** provides results from detailed analysis of the dataset, human assessments, automatic prediction errors, and evaluates new training of a multitask model. **Chapter 5** broadens the discussion, addressing the selection of models for APA, challenges in creating publicly accessible datasets, and evaluation methods used while also outlining directions for future research. The thesis concludes in **Chapter 6**, where the research questions are revisited, key findings are summarized, and the implications of these findings for future work in the field are discussed.

CHAPTER

# TWO

# THEORY

This chapter provides the background and theory serving as the basis for the work presented in this thesis. Parts of the theory given in this chapter were reported in [12].

## 2.1 Speech Communication

The speech communication process involves three main components: speech generation, recognition, and understanding. Figure 2.1.1 illustrates the fundamental elements of speech communication and how speech generation and perception are linked in a speech chain. To generate speech, a speaker first formulates a message and then uses their language knowledge to translate it into the physical production of speech. The speaker then uses their vocal tract system to produce a speech segment. On the other hand, when a listener, or the speaker themselves, hears the sound waves produced by speech, the cochlea in the ear converts these waves into motion and begins to extract key features of the sound. Combining these features with their language knowledge, they can understand what has been said.

Starting from the left-hand side of the speech chain in Figure 2.1.1, we can see that the speaker, or even a computer system, employs various contextual cues like semantics, phonemes, words, and intonation related to both the language and the topic at hand to produce the appropriate sound for conveying information. To produce sound, air-pressure waves are generated by the lungs and then propagated through the trachea and the oral and nasal cavities. An overview of the anatomy of the head and neck is shown

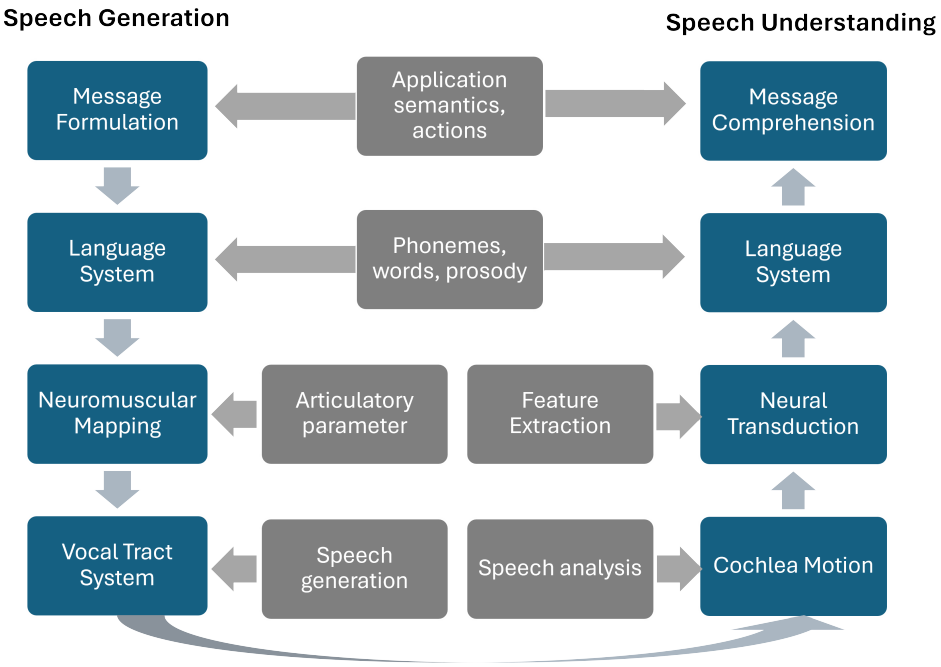



**Figure 2.1.1:** Speech chain showing components of speech generation and understanding, modified from [13].

in Figure 2.1.2. The breathing process involves the diaphragm contracting and relaxing, filling the lungs with air. This air creates air pressure within the trachea and on the vocal cords of the larynx. The vibration of the vocal cords and/or the exhalation of these airwaves serve as a sound source, while the shape of the vocal tract and nasal cavity act as a filter, causing the sound to differ based on the anatomy of the throat, mouth, and nose region.

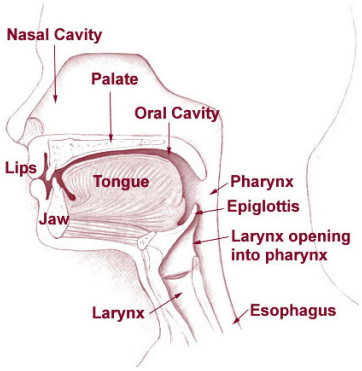



**Figure 2.1.2:** Anatomy of head and neck, showing elements used in speech production, from [14].

It is worth noting that when we describe phonemes as sounds, we must also acknowledge the significant influence of variability on all aspects of speech. Various factors can affect the pronunciation of speech sounds, such as physical attributes like the length of the vocal tract, age, gender, and illnesses, as well as environmental elements like noise in the surroundings, stress levels,

the person you're talking to, different accents, and more. Despite this variability, phonemes remain a group of sounds generally understood to represent the same underlying meaning.

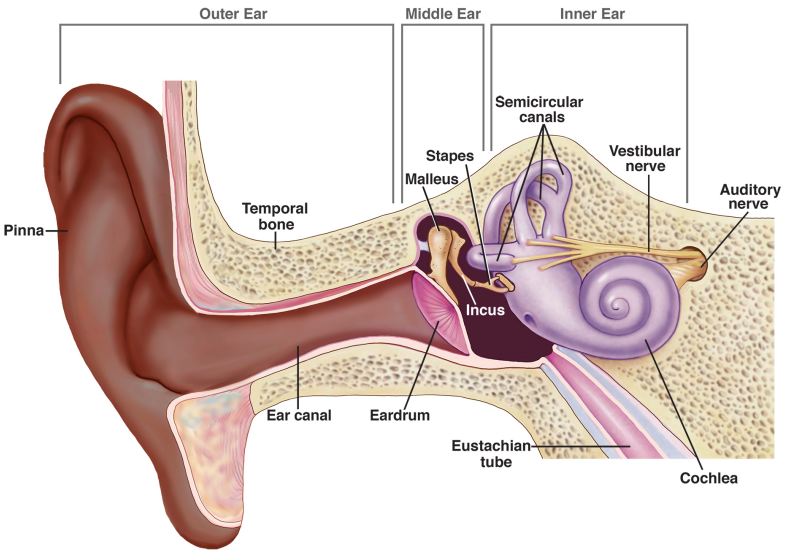



**Figure 2.1.3:** Overview of the peripheral auditory system with the outer, middle, and inner ear, from [15]

In the process of speech analysis, cochlea motion plays a crucial role. When humans hear sounds, it is the air pressure waves that reach the outer ear and cause the eardrum to vibrate accordingly. The peripheral auditory system, which includes the outer, middle, and inner ear, as depicted in Figure 2.1.3, is responsible for this process. The middle ear transmits sound from air into the inner ear's fluid through the oval window [16]. The inner ear is where the cochlea connects directly to the auditory nerve, which sends signals to the brain with features representing the sound [13]. The ear structure emphasizes and de-emphasizes certain sound frequencies, which helps determine the location of a sound source. Lastly, the listener's contextual knowledge of phonemes, words, and semantics is employed to comprehend the message they receive, as shown on the right-hand side of Figure 2.1.1.

## 2.2 Phonemes

The sounds we produce when we speak can be grouped based on the shape of our vocal tract and, specifically, on the placement of our tongue. These sounds are called phonemes, and they are the basic unit of speech. If we replace one phoneme with another, we get a different word. Consonants and

vowels are the two main groups of phonemes. Consonants are articulated with constrictions in the vocal tract, while vowels are not. It is the shape and size of the resonating cavities, with the tongue and lips as primary articulators, that change to produce different vowels. All vowels and some consonants are classified as voiced sounds, meaning that they are produced with the vibration of the vocal cords, creating a fundamental frequency in the sound. Consonant sounds are produced by placing the tongue against the hard palate in the oral cavity, where different placements can produce a range of sounds. When the valve soft palate opens for airflow through the nasal cavity, we get nasal sounds like "m" and "n." The tongue, teeth, and lips are also articulators that help shape the oral cavity and produce different sounds. Phonemes in a word are often influenced by the sounds before and after themselves, which is why we have many different phonemes for the same letter in the alphabet. For example, the "a" in "apple" has a different pronunciation than the "a" in "many." The phonetic alphabet names each of these sounds to make them easier to distinguish in text format. For instance, the "a" in "apple" is written as "/æ/" and the "a" in "many" is written as "/e/." But even with the wide variety of phonemes, we don't account for all coarticulations. For speech to be fluid, we prepare or start producing the next sound in the position of the last sound, creating overlapping movements in the mouth so that phonemes are still influenced by neighboring sounds.

## 2.3 Variability

Speech communication entails a large amount of variation from speaker to speaker. Some variability comes from the physical differences of the speakers, while others are more unique to each speaker and environment. These factors affect the way we produce and understand speech, and it's important to be aware of the relevant variability factors to accommodate these in a speech communication system.

As described in Section 2.1, the physical form of the vocal tract and mouth highly affect speech production. As the vocal tract length is dependent on both age and gender, this is a source of variability in pronunciation.

Speech is greatly influenced by the environment in which it occurs. Suppose a speaker is presenting a prepared topic in a formal setting in front of a group of people. In that case, their speech will differ significantly in terms of prosody, which includes rhythm, intonation, and stress patterns, in addition to word choice and pronunciation, as opposed to when talking to a close friend in

private. One key difference between these settings is whether the speech is spontaneous or rehearsed. In spontaneous speech, there is a variation in prosody that reflects the speaker's emotions, and there is a higher use of filled pauses such as "um" and "uh" during hesitation or when filling in gaps in the conversation.

Additionally, the noise level in the room can significantly impact how the speaker projects their voice. For instance, a speaker will automatically raise their loudness if there is a high volume level in the room and will whisper in a quiet library. Whether or not the audience is familiar with the speaker can also affect the speed, choice of words, and clarity of articulation. When talking to a close friend, for example, a speaker may use slang words, speak faster, and use dialects and accents specific to their shared geographic region.

When speech is an element of a technological system, recording devices such as microphones and processing equipment can also affect the speech sample. Therefore, a dataset must be recorded in a uniform environment to prevent variability due to equipment or the immediate recording environment — unless, of course, this variability is wanted to create systems that should tolerate irregularity.

All these variability factors can cause challenges when working with speech in a technology-based system, especially if a model is to be used in a specific situation. If a speech recognition system is invented to transcribe a conversation between two speakers on the phone to each other, this will be a very different type of speech than if you are to transcribe a news broadcast on TV. This is a known issue in speech technology, as one of the most popular datasets used in training is **LibriSpeech** [17], which only contains read speech and not spontaneous speech. If a model is trained solely on read speech, it will not perform as well for spontaneous speech, so it is important to know which environment the system should be used in and develop adapted datasets. The same goes for systems that are intended to understand child speech. As they are still in a developmental state of height, giving a short vocal tract length and higher formant frequencies, and due to still being in a language acquisition state, their pronunciation differs from adults in the use of repetition, intonation, and simplification of words or phonemes. Child speech can have a high grade of variability within age groups due to different developmental stages. Children may also not be as familiar with how prosody can convey meaning beyond words. For example, a rising intonation at the end of a sentence can indicate a question, while a falling intonation can indicate a statement. Similarly, stress on certain syl-

lables can convey emphasis or importance. Such aspects of speech are often learned at different stages of language development.

### 2.3.1 Norwegian

Language is also a source of variability among speakers, and due to the large number of dialects and their range of inequality in the Norwegian language, this can cause additional challenges when building speech systems intended for Norwegian speakers. Not only is there an enormous range of pronunciation within dialects, but Norwegian also has two official written languages that, in essence, provide two languages in which a speech sample can be transcribed. Most native Norwegians are exposed to various dialects from a young age through family, TV shows, and music. Still, it can be challenging for non-native speakers or immigrants to learn the appropriate dialect in the local area. Dialects in the urban east part of Norway, close to its capital, Oslo, are most similar to the written language Bokmål, but in western parts, one would see more similarities with Nynorsk. Having speech technology systems that generalize across dialects but simultaneously can distinguish between dialects or written languages for transcription is still being researched. [18]

## 2.4 Human pronunciation assessment

Pronunciation assessment is known as the evaluation of how accurately and clearly a speaker pronounces the sounds of a specific language. That said, there is no clear definition of correct pronunciation; rather, a scale of pronunciation is used during the assessment. This scale can range from unintelligible speech to native-sounding speech (accentedness) and will change depending on the goal of the assessment. [5, 19] Pronunciation assessment can greatly benefit speech development and language learning, as it provides detailed feedback. In contexts where spoken language proficiency is critical, it can be used to assess the level of clarity and intelligibility, ensuring good comprehensibility for the listener.

To help unify pronunciation assessment among assessors, various types of pronunciation errors have been established. Using Figure 2.4.1, there are phonetic and prosodic errors, as well as errors related to accuracy, fluency, and completeness.

Phonetic errors account for mistakes in the production of individual sounds,

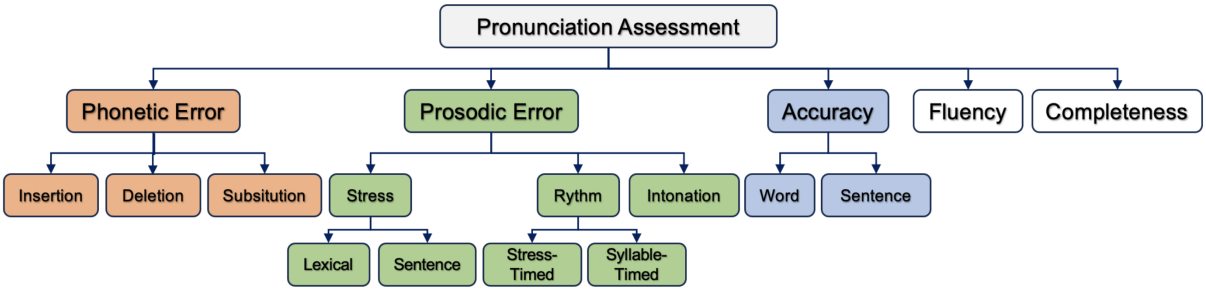



**Figure 2.4.1:** Types of Pronunciation Errors for Assessment, from [19] used under Creative Commons CC-BY license [20].

such as vowels or consonants, and include inserting, where the speaker adds a sound; deletion, when the speaker removes a sound; and substitution, where the speaker swaps out a sound. Prosodic errors are categorized into stress, rhythm, and intonation, focusing on elements that impact pronunciation at a higher level, either across words or sentences. The stress of a word is determined by the sound the speaker emphasizes. Emphasizing a syllable in a word or sentence by increasing loudness, duration, or pitch distinguishes between words or expresses emotion in an utterance. Rhythm explains the patterns of stress and pauses in a language. There are two categories of rhythm: stress-timed languages, such as Norwegian, maintain an almost constant period between each stressed syllable, while other languages, such as Finnish, have syllable-timed stress, with roughly equivalent syllable durations. Intonation encompasses the melodic pattern and pitch variations across phrases and sentences, as many languages use intonation to indicate a question. Tonal languages, like Vietnamese, use pitch to differentiate words, posing a challenge for non-tonal speakers to learn. Norwegian and Swedish have pitch accents in some dialects but are not tonal languages, despite intonation being an important part of the languages. [19, 21, 5]

In order to accurately assess the communication skills of a L2 learner, it is crucial to evaluate the intelligibility, accuracy of sounds, and completeness of their vocabulary. Additionally, fluency must be assessed to determine if the listener can easily perceive the speech. Accentedness, or the influence of the speaker's native language on their pronunciation, is a significant factor that can affect intelligibility. Even highly proficient L2 speakers often exhibit some degree of accentedness, but for most speakers, this will not impact their overall communication effectiveness. [19, 5]

## 2.5 Automatic Pronunciation Assessment - APA

Human speech therapists have traditionally performed pronunciation assessments, but with advancements in computing power, traditional machine learning, and neural networks, many systems have been developed to automate part of the pronunciation assessment. Fully encompassing all aspects of pronunciation assessment has been a challenging task, but by utilizing speech recognition based on deep learning and new datasets, APA is possible. [19, 5]

## 2.6 ASR systems

In the domain of spoken language systems, three fundamental subsystems collectively form a dialogue system: text-to-speech, speech-to-text, and speech understanding. Text-to-speech systems generate speech signals from written text, speech-to-text systems convert spoken words into text, and speech understanding systems map words to actions, known as a dialogue manager. ASR systems, using speech recordings as input and providing transcriptions as output, are the focus of this chapter.

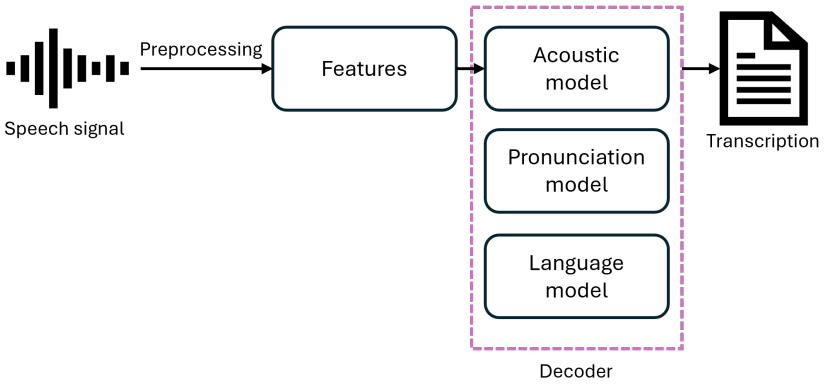



**Figure 2.6.1:** Elements of traditional ASR system.

A traditional ASR machine learning system, as shown in Figure 2.6.1, consists of the input speech signal and preprocessing, an acoustic model, a pronunciation model, and a language model. Each element provides the information needed to decide what the speech signal contains. The speech signal is processed, and spectrograms are produced using the Short-Time Fourier Transform (STFT). Different representations of the speech signal can be used to analyze the recording directly by studying periodic flow, harmonics created by the vibrating vocal folds, or formants created by the shape of the vocal tract.

In addition, Mel-Frequency Cepstrum Coefficients (MFCC) are used to extract important features from the recording. The MFCCs mimic the human non-linear auditory system using mel filterbanks and have been shown to work well for speech recognition.[22] The computed frequency spectrum is transformed to match the human perception of sounds using triangular bandpass filters in the Mel filterbank. Human ears have a logarithmic perception of loudness, so the log of the mel filterbank output is computed, similar to a logarithmic scale. Taking the inverse Fourier transform yields the cepstrum, which de-correlates the MFCC. The first 13 cepstrum coefficients typically contain the phonetic information of the speech signal, so they are the ones typically used for ASR.

The language model is a statistical model that captures grammatical structures and variation within a spoken language and gives the likelihood of a word sequence. It represents the language system of the speech chain, giving the overall language-specific information a speaker would know. The possible word sequences in a speech recording highly depend on the language model's probabilities. In continuation of the language model, the pronunciation model provides information on a more detailed level. Including a lexicon or dictionary of words and their phonetic transcription also includes information about how phonemes are based on context, their neighboring phonemes, and word boundaries.

The acoustic model provides a means to model a sequence of feature vectors given a sequence of phonemes. This representation combines the knowledge of acoustics, phonemes, and environmental variability. Using a Markov chain that describes a sequence of possible states, the states themselves are hidden and bound by the Markov assumption they are only dependent on the previous state. To model the transition probability between states, we use Hidden Markov Models (HMM), and for each state, there is an emission probability stating the probability of observing a set of features given the current state. This emission probability is modeled using Gaussian Mixture Models (GMM) that include the component's mean, covariance, and weighting of the represented features. When using HMM and GMM as the acoustic model in an ASR system, it is common to use the Forward algorithm in the evaluation of the likelihood of the observed data, the Viterbi algorithm to find the most likely sequence of states, and at last, the Baum-Welch algorithm is used to estimate and update HMM parameters.

These components result in an equation that, based on all information from features and models, outputs the sequence of words with the highest proba-

bility value. Traditional ASR systems operate by breaking down speech into smaller components, such as phonemes, and matching these to known patterns using statistical models. However, advances in research show that with the help of neural networks, it is possible to process speech input directly to output text without the need for these smaller speech units. Even though the traditional models are no longer used for ASR, they are still used for APA due to their fundamental phoneme-level insight.

## 2.7 Neural Networks End-to-End ASR

Inspired by the neurons in the human brain, a neural network consists of layers with nodes that connect to each other with associated weight and activation functions. Some systems model each element of a traditional ASR system with a neural network, but it is possible to model most of the input-output sequence with a Deep Neural Network (DNN); this is an End-to-End ASR system.

## 2.8 Transformer based ASR - wav2vec2.0

Facebook, recently rebranded as Meta, have developed an End-to-End model called wav2vec2.0.[7] This transformer-based model takes raw audio waveforms as input and gives a set of output tokens such as characters.

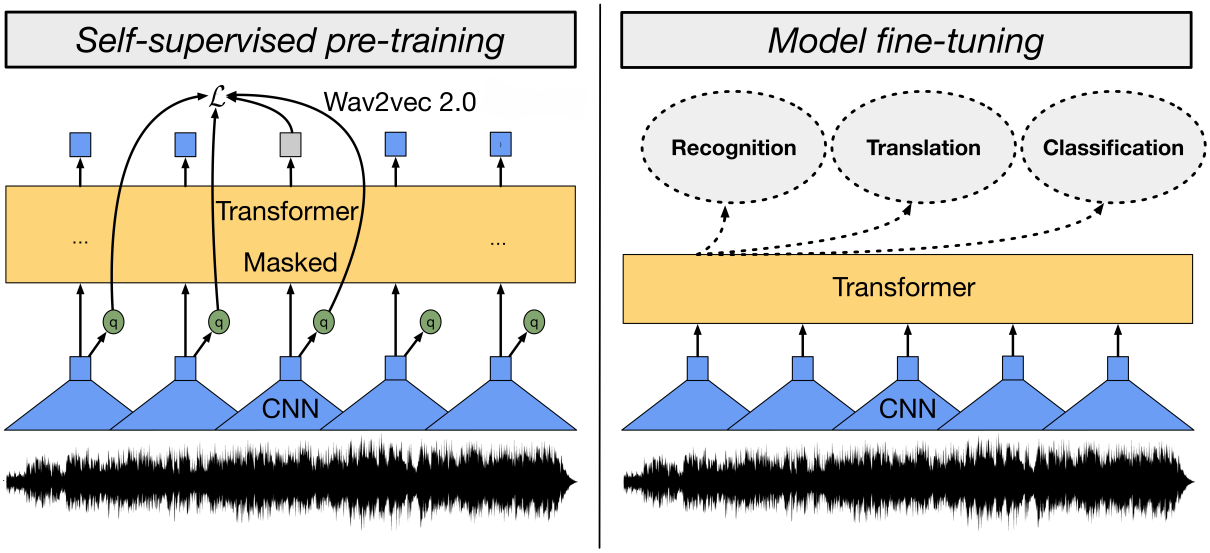



**Figure 2.8.1:** Wav2vec2.0 illustration of pre-training and fine-tuning, figure from [23, 24] with written permission of use.

Following the illustration on the left in Figure 2.8.1, during the self-supervised pre-training, raw audio input is encoded into latent speech representation through the feature encoder consisting of multilayer Convolutional Neural Network (CNN). The output of these networks are used in parallel, one is masked and serves as input to the transformer block, the other is the quantized representations serving as distractors later. The outputs of the transformer blocks are context representations. These, together with the quantized representations, give input to the contrastive task that will identify the correct quantized latent audio representation in a set of distractors for each masked time step. The contrastive task uses both a contrastive loss and a diversity loss. This pre-training is done with unlabeled speech using the LibriSpeech dataset. [7]

For the model fine-tuning for the downstream speech recognition task, the recognition box on the right of Figure 2.8.1 would consist of a randomly initialized linear projection layer added after the transformer block. Here, the Connectionist Temporal Classification (CTC) loss function minimizes the loss between a continuous times series and a target sequence. [7, 25, 26] SpecAugment, a data augmentation technique, is used to improve robustness by time warping features and masking blocks of frequency channels and time steps. [27]

Several datasets, including the **TIMIT** dataset [28], which consists of detailed phoneme-level labels, were used in specified fine-tuning of **wav2vec 2.0**, creating task-specific models capable of learning phoneme-level patterns. However, it's important to note that the quantized units in the standard model do not explicitly model phonemes. Instead, these units encompass more general acoustic features that can be adapted to phoneme recognition through fine-tuning.

There were also experiments using language models in the **wav2vec2.0** model, and both a 4-gram language model and a transformer-based model trained on the **LibriSpeech LM** [17] corpus were tested. The model provided state-of-the-art results on different datasets, but the transformer-based language model generally performed better.

The initial experiments using the **wav2vec2.0** model proved that the pre-trained model learned speech structure, so only a small amount of labeled data was required to fine-tune it for speech recognition. This is the basis for the further work explained in Chapter 3, where using the general speech knowledge of the **wav2vec2.0** model, one only needs a relatively

small dataset of labeled data to fine-tune the model on a specific speech recognition problem.

## 2.9 Evaluation metrics

Evaluation metrics are used to objectively assess the performance of ASR and APA systems. Given the sliding scale of pronunciation assessment from unintelligible to native-like, the need for objective metrics is helpful in aligning assessments across both human and digital systems. This section presents the benchmark evaluation metrics used in this field today. However, there is new research showing the need for more extensive evaluation methods. [29, 30]

### ASR evaluation

**Word Error Rate (WER)** is the main evaluation metric used for ASR systems. Here, the number of words correctly transcribed by the system, compared to the target text, gives an overview of the model's accuracy. It is important to note that each mistake is weighted equally, so even though the same meaning can be understood from the transcription, only the words identical to the target text will be correct. WER is calculated as in Equation 2.1,

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{2.1}$$

where $S$ is number of substitutions of full words, $D$ is number of deletions of words, $I$ is number of insertions of word, $N$ is number of words in the reference, and $C$ is the number of correct words.

When working with one-word utterances, the **Character Error Rate (CER)** is more useful, giving character-level accuracy between transcription and target text. CER has the same equation as WER, shown in Equation 2.2, with the adjustment that $C$ represents the number of correct characters, $N$ is the number of characters in the reference, and substitutions, deletions, and insertions are on character-level.

$$CER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{2.2}$$

### APA evaluation

For evaluating APA systems, performance is measured based on how accurately the system assigns a score on pronunciation. Classic **Accuracy (ACC)** is used to measure correct predictions among the total number of cases processed. It is calculated as in Equation 2.3, with inputs of predictions and corresponding references,

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

where $TP$ stands for True positive, $TN$ is True negative, $FP$ is False positive, and $FN$ is False negative [31]. This formula explains ACC for binary classification problems. In multiclass classification, accuracy is calculated by comparing the predicted and true class labels for each class, and the total accuracy will then be the total number of correct classifications divided by all classifications, as shown in Equation 2.4.

$$ACC_{multi} = \frac{\text{correct classifications}}{\text{all classifications}} \tag{2.4}$$

The datasets used for APA training are usually unbalanced between different classes, especially if the same dataset is used for both ASR and APA training; the best ratings are over-represented. **Unweighted Average Recall (UAR)** is therefore used to give a better understanding of the performance without weighting by class frequency, as it represents the fraction of positive examples correctly labeled as positive by the model. It's computed as in Equation 2.5, where $TP$ is True positive, $FN$ is False negatives, and $N$ is the number of classification classes.[32, 33]

$$UAR = \frac{1}{N} \sum_{i=1}^{N} \frac{TP}{TP + FN} \tag{2.5}$$

Lastly, the **Mean Absolute Error (MAE)** is a metric used for measuring the average of the absolute differences between the predicted and the observed values. [34] It's calculated as shown in Equation 2.6,

$$MAE = \frac{1}{n} \sum_{i=n}^{n} |y_i - \hat{y}_i| \tag{2.6}$$

where $n$ is the number of observations, $y_i$ is the actual value of the $i$-th observation, and $\hat{y}_i$ is the predicted value from the model.

CHAPTER

# THREE

# METHOD

This section presents the datasets, methods, and models used as a basis for my experiments. The method will explain the general implementation done under the Teflon project, described in [35, 6], and the author's implementation of this method and further use is described in Section 3.3.

## 3.1 Teflon Dataset

The basis of this work is the results of a new dataset containing single-word utterances of child speech in Nordic languages.[6] This dataset was originally made for use in the gamification of a CALL system, but it also provides a valuable dataset for further advancement in both ASR and APA for child speech in general. [3]

The full Teflon dataset consists of speech in both Swedish, Finnish, and Norwegian for both L2 learners and children with speech sound disorder (SSD). Information on all L2 -related datasets, as that is the focus of this thesis, are given in Table 3.1.1. This figure shows the general information of the datasets and offers a comparison between the Swedish, Finnish, and Norwegian L2 corpus. The **TeflonNorL2** dataset is unique because it includes both native Norwegian speakers and L2 speakers. It offers orthographic, global 1-5, and phoneme-level scoring, and it is the only publicly available dataset of the three. Further, it has a higher number of speakers and more unique words, resulting in six times more speech recording minutes than the Swedish corpus. The speakers of the Norwegian dataset also have a wider age range than the others.

**Table 3.1.1:** Collection of facts regarding Teflon L2 corpus, from

| Corpus | language | speaker kind | # speakers (L2 ) | ages | # utterances (minutes) | # words | annotations | public availability |
|---|---|---|---|---|---|---|---|---|
| TeflonNorL2 | nor | Native/L2 | 52 (33) | 5-12 | 9443 (544) | 205 | orth, glob, phon | Yes |
| TeflonSweL2 | swe | L2 | 20 (20) | 7-11 | 2384 (90) | 121 | orth, glob | No |
| TeflonFinL2 | fin | L2 | 24 (24) | 7-11 | 2124 (83) | 90 | orth, glob | No |

The annotations of each utterance on a global word level range from 1 to 5 and the guiding descriptions of each score were given to the assessors as stated in Table 3.1.2. In addition, binary phoneme level scores were annotated, where 0 equals incorrect pronunciation, and 1 equals correct pronunciation of the phoneme. 30% of the utterances were annotated by two human assessors, and the rest were only annotated by one human assessor.

**Table 3.1.2:** Overview of labels corresponding to global 1-5 scores used by human assessors, from [6].

| Score | Label |
|---|---|
| 1 | Not at all identifiable as the target word |
| 2 | Difficult to identify as the target word |
| 3 | Slight phonemic error(s) |
| 4 | Subphonemic error(s) or ”unexpected variants” |
| 5 | Prototypical, adult-like |

The distribution of the annotated global scores showed in Figure 3.1.1 is skewed, with a larger representation of level 3, 4, and 5 scores, but this division is more even than for **TeflonSweL2** and **TeflonFinL2**.

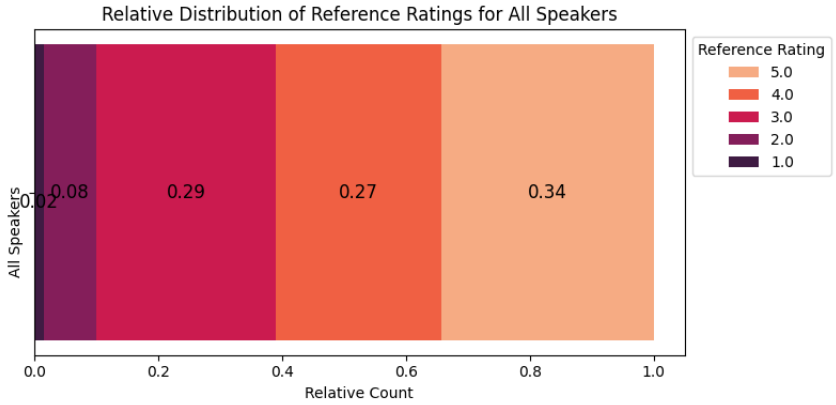



**Figure 3.1.1:** Distribution of human assessed reference rating for Norwegian used part of Teflon dataset.

In addition to global and phone-level scoring, the annotators have also marked for the presence of additional features: prosody, noise/disruption, pre-speech noise, and repetition. A marking for prosody, as introduced in Section 2.3, is used to emphasize that even though all phoneme sounds are

correct, there is some mispronunciation overall, for example in rhythm or intonation. Such annotations can help explain why only 33.1% of utterances were scored at a level 5, even though around 44% had no phoneme-level errors. [6] The level of noise in a recording can influence how well annotators are able to hear the actual pronunciation, so markings for noise/disruption in the full recording, as well as pre-speech noise, have also been annotated. Lastly, repetition of sounds, such as the full word or re-starting pronunciations, is marked as repetition. In this work, we use these features to evaluate how they affect human and automatic annotations.

As additional data, extensive information regarding each speaker's primary languages and levels of proficiency were given to use in this thesis as part of the Teflon project. Information on each speaker, beyond only native or non-native status, allows further analysis based on first language, proficiency levels, duration of time they have lived in Norway, and what languages are the main ones used at home. Some of this information, and their effect on prediction error, is presented in Chapter 4.

The varying interpretations of speech samples as personal data in different countries have led to the **TeflonNorL2** corpus being the only dataset publicly available [36]. This highlights the evolving ethical considerations around speech. While the full dataset required informed consent from parents and anonymization of speakers, in Finland, speech recordings themself are considered personal information. When involving a large group of children's speech in a technological system, it's crucial to fully consider ethical aspects. However, since this dataset comprises single-word recordings, the possibility of capturing a large amount of sensitive information is low. Additionally, children's voices undergo extensive changes, as discussed in Section 2.3, making their voices unrecognizable in a few months or years. Some voices may already be unidentifiable due to the passage of time since the recordings took place.

The data collection process is explained in detail in [6], so the choice of sources, collection methods, and dataset contents was predetermined. However, given the research question regarding what affects pronunciation errors in an APA system, the level of detailed annotations in this dataset is more than valid and reliable for this task.

## 3.2 Multitask method

Wav2vec2.0 models have already been shown to be suitable for pronunciation assessment.[33] Meanwhile, with the use of multitask learning, even better APA results have been achieved, all the while keeping the convenience of ASR transcriptions.[35] By taking advantage of the similarity between the ASR and APA tasks, one can share components of the deep learning architecture, improving generalization and efficiency and reducing computational cost.

In the work done by Aalto University for the Teflon project, the two tasks share layers and have respective ASR and classification heads, as shown in Figure 3.2.1. Intermediate layers in the model embed a higher degree of phonetic information compared to the last encoder layer. Therefore, the linear ASR head layer is placed on top of the last layer, and the classification head used for APA is placed a number of layers before. The layer the classification head is connected to, here referred to as X, is model and dataset-dependent but usually ranges from levels 15 to 20. Transformer layers after layer X are optimized by CTC loss, while the rest of the layers, as well as the CNN network of the wav2vec2.0 model, are trained with a combined gradient of the CTC and cross-entropy loss. The CTC loss calculates the probability of a target sequence, and cross-entropy measures the difference between the predicted probability distribution and the true distribution. For the APA task, a decision tree trained on CER s from the ASR component is used to adjust the classification head, finally merging them into output assessment levels. [35, 6]

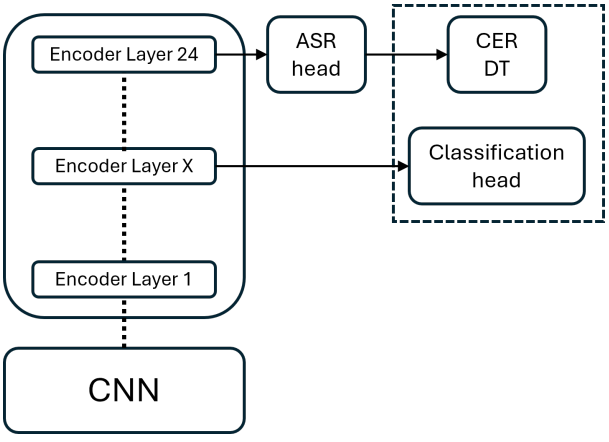



**Figure 3.2.1:** Overview of multitask training, modified from [35]

As the base model used by Aalto University was originally fine-tuned for adult ASR in the target language, the main learning aspect of the model is child speech and all the variations that it entails, as well as the pronunciation

assessments. The continued fine-tuning they implemented works well for the limited dataset and is necessary due to the current scarcity of models trained on child speech. Only the utterances rated as 4 or 5 were used in training the ASR model, further reducing the amount of training data.

Due to the limited dataset, the Aalto University group used 6-fold Cross Validation (CV). The main concept of cross-validation is to be able to train models on the full dataset, all while avoiding overfitting. In practice, this is done by training several models on only one subset of the dataset and taking the average of the evaluation of each model, as illustrated in Figure 3.2.2. This method also uses one subset of the data as an intermediate development set, and the rest is used for training.

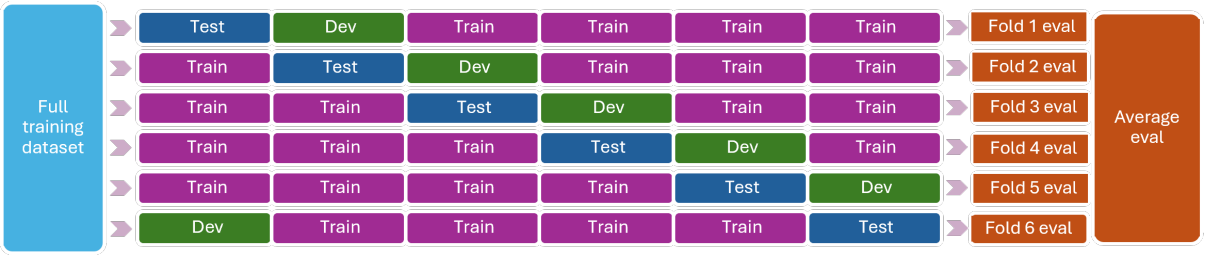



**Figure 3.2.2:** Illustration of training using 6-fold cross-validation.

## 3.3 Experiments

As only general examples of multitask wav2vec2.0 ASR APA training are available open-source [37], the first part of this works experiment is implemented based on code used in the development of CALL system, as partners in the Teflon project.[3] First, their code was re-created to get utterance-level prediction results, as well as evaluation of these, and then the code was used as a template for further fine-tuning of a wav2vec2.0 model on child speech. A Norwegian fine-tuned base model was selected for this through CER analysis of models created by the Scribe project and the National Library of Norway AI Lab (NbAiLab) [38, 8].

### 3.3.1 Evaluation of Teflon results

As part of the Aalto University work [3], the *nb-wav2vec2-300m-bokmaal* model from NbAiLab, [8] previously fine-tuned on Norwegian adult speech, is further fine-tuned for the multitask ASR /APA task on child speech. Training is done using a 6-fold CV, with no fold or speaker overlap, giving 4 folds with 9 speakers and 2 folds with 8 speakers. As around 30% of

the utterances have two annotations, this poses a challenge for what score should be set as the reference rating for the model. As a solution, in cases where two annotations exist for one utterance, the two scores are averaged and rounded up, so each utterance has only one reference rating. For the NbAiLab 300m bokmaal model, layer 20 of the encoder blocks is used as input to the classification head layer while keeping the ASR head connected to the last transformer block. All evaluation methods are implemented using *evaluate.load* from **Hugging Face** [39].

While the existing evaluation of the multitask model is presented in [6], in this work, the training checkpoints are used to get both ASR transcriptions and APA predictions for each utterance in the dataset. The relevant L2 evaluation results are repeated in Table 3.3.1, and further analysis of how APA prediction errors are distributed across different groups in the dataset are presented in Chapter 4, and is the main focus of this work.

**Table 3.3.1:** Evaluation results of Teflon L2 datasets, replicated from [6].

| | Measures ASR Performance | | Measures APA Performance | | |
|---|---|---|---|---|---|
| Data | WER [%] (↓) | CER [%] (↓) | ACC [%] (↑) | UAR [%] (↑) | MAE (↓) |
| TeflonNorL2 | 10.74 | 4.21 | 55.18 | 39.83 | 0.53 |
| TeflonSweL2 | 9.95 | 4.04 | 48.24 | 35.12 | 0.70 |
| TeflonFinL2 | 6.30 | 2.13 | 72.08 | 43.07 | 0.37 |

In order to analyze the distribution of prediction errors made by the APA system, we performed a thorough group-based error analysis. This process included dividing the dataset into different groups based on various demographic characteristics such as speaker type (native/non-native), age, target word, features, and first language. We then created a set of visual plots for each group to illustrate the error distribution within these subgroups.

This approach allowed for a systematic examination of the impact of these demographic factors on the performance of the APA system. By analyzing the variance in error rates across different groups, we aimed to identify any biases or discrepancies in system accuracy that could influence the effectiveness of the pronunciation assessment for diverse learner populations. The visualization of error distributions was facilitated by using **Python** programming language [40], **Matplotlib** plotting library [41], **Seaborn** data visualization library [42], which enabled detailed comparative analysis and insightful interpretation of the results.

The selection of demographic groups for error analysis for this work was based on strategic considerations of the available data, informed by ex-

ploratory research. The initial selection process involved a trial-and-error method to evaluate various demographic combinations and determine the most informative groupings. This practical method optimized the analysis considering the limitations of the dataset's variety and size.

To gain a solid foundation for understanding prediction errors, reference ratings within different demographic groups were reviewed prior to analyzing prediction errors. This initial analysis offered important insights into the unique challenges and differences within each subgroup, highlighting how the complexity and subjectivity of the pronunciation assessment task could impact reference ratings. Identifying variations in these ratings is important to show that higher prediction errors in specific groups may not only result from shortcomings in the APA system but also due to the intricate nature of the task for those particular groups.

### 3.3.2 Fine-tuning new model

As a result of the rapid advancements in the deep learning field, new wav2vec2.0-based models fine-tuned for Norwegian have been released after the preliminary experiments at Aalto University took place. [35] A comparative analysis was conducted to identify the possibility of a more effective base model for a multitask ASR and APA system. The primary metric used for evaluation was CER , which is relevant for measuring the accuracy of model transcription at a character level and crucial for both ASR and APA tasks. Multiple models from NbAiLab and the Scribe project were assessed against original Facebook models.

Note that the implementation of CER was done using *evaluate.load* from **Hugging Face** [39], if implementation is done using *Char Error Rate* from **PyTorch**, mismatched results will occur due to different policies regarding double spacing. When using wav2vec2.0 models to transcribe, double spacing between words can arise; even though the dataset consists of single-word utterances, the transcriptions will sometimes mistakenly output several words.

As a result of transcribing the entire directory of utterances for the Teflon project, it was observed that some utterances exhibited an extremely high CER . Further investigation into these high-CER utterances revealed issues such as additional speech segments and significant background noise. To address this, a list of outliers was compiled based on their CER scores, identifying recordings that either contained additional speech or suffered from

poor audio quality. These outliers can either be edited to ensure they contain only the intended single-word utterances or excluded from the dataset altogether to enhance the model's accuracy.

Moreover, the dataset available for this work in the Teflon directory differs from that utilized by Aalto University for cross-validation training of a multitask model. To evaluate the impact of these differences, various Norwegian fine-tuned models were ranked based on their CER using different subsets of the data: one with the complete dataset from the directory, one excluding the data omitted by Aalto University, one with the outliers identified in this work removed, and another comprising only utterances rated with a score of 4 or 5 by human assessors. However, the dataset used by Aalto university is what serves as the "full dataset" for all other parts of this work.

Additionally, it was discovered that some recordings had been rated with a score of 0 by human annotators, indicating they should not be used for model training due to various quality issues. However, discrepancies were noted where one annotator assigned a score of 0 while another provided a substantive score. This necessitates the development of clear guidelines on how to handle such recordings to ensure consistency and reliability in training data quality.

#### Facebook

The Facebook **Base 960h** and **Large 960h LV60 Self** wav2vec 2.0 models served as baselines for the CER rankings. The base model is pre-trained and fine-tuned on 960 hours of LibriSpeech ASR data. The Large model is first pre-trained on 53k hours of un-labelled audio data from the LibriSpeech and LibriVox corpora and then fine-tuned on 960 hours of LibriSpeech ASR data similar to the base model. [7, 9] Due to the content of these training datasets, these models are essentially trained on English adult read speech, so their performance on Norwegian child one-word speech is expected to be inadequate. Still, they provide a valuable benchmark to demonstrate the importance of fine-tuning for Norwegian speech.

#### NbAiLab

As previously introduced, the NbAiLab has developed models fine-tuned on Norwegian speech, one with 300 million parameters (300m) and another with 1 billion parameters (1b). There were models focusing on nynorsk as well, but for the scope of this project, only the bokmål related models are discussed.

The initial release, version one of the NbAiLab models, utilized only the Norwegian Parliamentary Speech Corpus (NPSC) dataset, which contains recordings of meetings from the Norwegian parliament (Stortinget) and was the first publically available dataset containing unscripted Norwegian speech. [43] The following overview of the models is based on presented information in [8].

This initial **300m** model is based on the Swedish VoxRex model, which was trained on the P4-10k corpus—a comprehensive collection of 10,000 hours of Swedish public service radio broadcasts along with 1,500 hours of audiobooks and other speech materials from the National Library of Sweden (KB). The decision to base the Norwegian model on the Swedish VoxRex was influenced by the linguistic similarities between Swedish and Norwegian, both stemming from the North Germanic language family and sharing numerous phonetic and lexical characteristics. Lastly, it is fine-tuned for Norwegian bokmål with the NPSC dataset.

Moreover, the first version also featured a larger scale **1b** model adapted from the multilingual XLS-R models. These extensive models were trained on a massive dataset of 436,000 hours of publicly available speech from diverse sources, including parliamentary proceedings and audiobooks, covering 128 different languages, aiming to empower the model with a broad phonetic landscape and multilingual versatility. Similarly, this is also fine-tuned for Norwegian bokmål with the NPSC dataset.

In the second version, the models were improved by integrating the Nordisk Språkteknologi (NST) dataset with the NPSC, broadening their training resources. The NST contains diverse speech data, mostly manuscript-read speech in Bokmål and some repeated words and numbers. The readers had different Norwegian dialects, but since the speech is not spontaneous, the pronunciation leans closer to bokmål. Still, models trained on the NST dataset are more likely to generalize across dialects. By combining these datasets, a comprehensive representation of the phonetic and contextual variations in the Norwegian language ultimately leads to improved performance across different Norwegian speech scenarios.

#### Scribe

A collaboration between the National Library of Norway, Telenor, and NTNU under the Scribe project released four Norwegian fine-tuned models. The following overview of the models is based on presented information in [38].

First, the **Radio** model is trained on the Bokmål elements from the Rundkast training set, which does not have an open license but consists of transcribed radio news and TV shows such as interviews and debates and was developed by NTNU. [44]

Then, a **Stortinget** model is trained on the Bokmål part of the NPSC training set, with segment lengths between 1-15 seconds, resulting in approximately 80 % of the original Bokmål training set.

Two combined models are also trained. One named **Combined Short** trained on random samples of the **Radio** and **Stortinget** training sets, with around 70 hours combined. The second, named **Combined Long**, is trained on the full combination of **Radio** and **Stortinget** datasets, which is 114 hours in total.

#### 3.3.2.1 Training new multitask model

After the best Norwegian-finetuned model was selected, training of a new multitask model was started to see if both ASR and APA results could improved due to the enhanced Norwegian language performance, selected based on CER .

Due to a power outage, the training was interrupted after three folds were completed, giving implementation as shown in Figure 3.3.1. However, it was decided that this served as a good indicator of performance. Therefore, the training was not restarted or continued. One underlying argument for this decision was the substantial energy consumption over several days needed to run this type of code on servers, as well as freeing up resources internally at NTNU during a critical period of usage.


**Figure 3.3.1:** Illustration of training using 6-fold cross-validation method, with three folds completed.

CHAPTER

# FOUR

# RESULTS AND RELATED DISCUSSION

This chapter will present results and discuss relative potential causes and effects throughout the section.

Section 4.1 presents detailed information about the TeflonNorL2 dataset and Section 4.2 give insight into pronunciation difficulties of native and L2 child speakers of Norwegian by presenting human assessments and how these reference ratings correlate to the groupings of the dataset.

Section 4.3 presents and discusses the performance of the APA system on different categories of the data, and shine light on what factors influence the prediction error of this APA system.

Section 4.4 regards the training of a new multitask ASR and APA model, including the CER comparison of models used as a method for selecting a base model for this training and accompanying findings of distinct recordings.

## 4.1 Analysis of dataset

To give a more detailed insight into the contents of the TeflonNorL2 dataset, beyond what is presented in [6] and Section 3.1, this section presents newly compiled information from the dataset on the speaker, age, target word, and first language level.

### 4.1.1 Speaker ID

There are 52 speakers and 205 target words in the dataset, but there is only a subset of speakers that have recorded utterances for each target word.

Therefore, there is a varying number of utterances per speaker ID, with around 40% that have 205 utterances and the rest having mostly around 167 recordings, as shown in Figure 4.1.1. There are multiple reasons for this; one is that due to the two rounds of data collecting, some non-native speakers recorded utterances for additional words that were not pronounced by the rest of the speakers, meaning that all native speakers have 167 or lower amount of recordings, none with 205. In addition, individual recordings have been excluded from the dataset due to, for example, extensive noise.

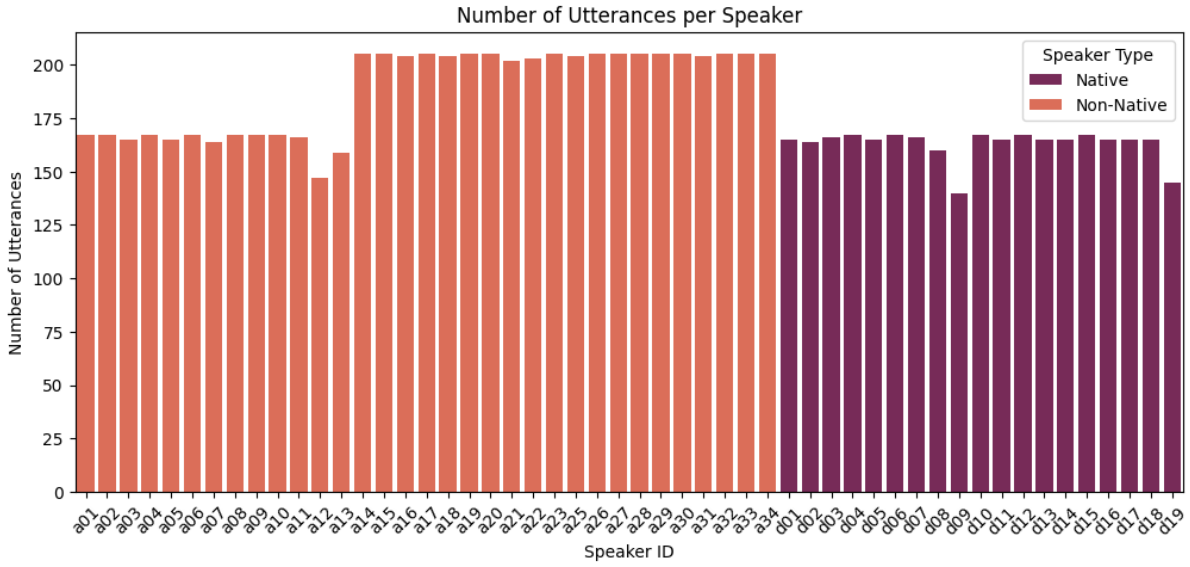



**Figure 4.1.1:** Number of utterances per speaker.

### 4.1.2 Speaker Age

The 52 speakers range from age 4 to 12, but as Figure 4.1.2 shows, both the number of speakers per age group and the distribution of native versus non-native speakers are not even. For the lowest age groups of 4-8, there is a higher number of native speakers, with group 4 consisting of only one native speaker. The older groups from 10-12 consist of only non-native speakers, with 12 being the largest age group in terms of the number of speakers. Age group 9 is the only group that has an equal amount of native and non-native speakers. This unbalanced distribution will largely affect all analyses done on an age group level. Therefore, having this overview is very informative for further work.

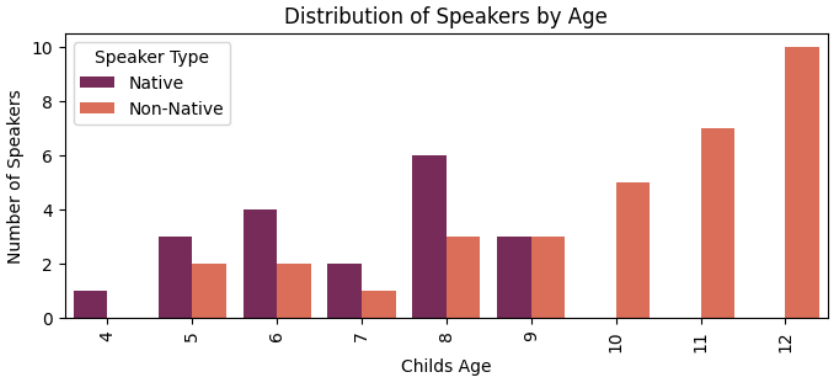



**Figure 4.1.2:** Distribution of Speakers by Age and Speaker Type.

### 4.1.3 Target words

As some words were only subject to pronunciation recordings for one of the data collection rounds, consequently, the number of utterances per target word varies. Figure 4.1.3 shows that most target words were uttered around 50 times, and they have around 20 utterances from native speakers and the rest from non-native speakers. However, there are 38 additional target words that only have around 20 recordings of non-native speakers. The takeaway here is that most words are represented by both native and non-native utterances, and there are many recordings from each group, but for approximately 18% of the target words, there are only non-native speech samples.

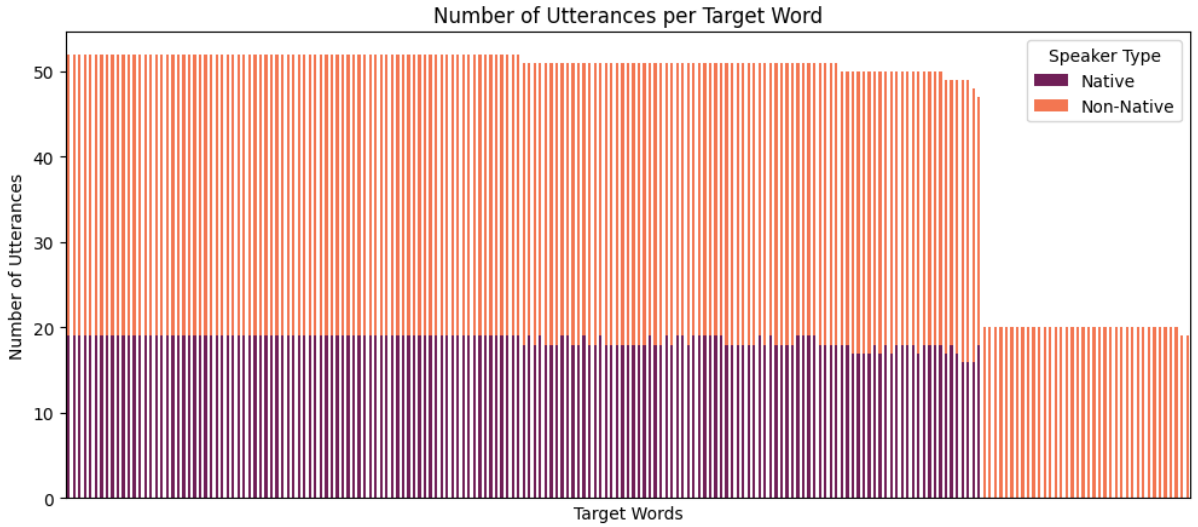



**Figure 4.1.3:** Number of utterances per target word.

### 4.1.4 First Language

From the additional information noted for each speaker, the distribution of first languages is shown in Figure 4.1.4. There are 7 first languages that only have one speaker, so due to the possibility of one speaker's inherent proficiency overshadowing all other results, doing analysis on the first language level will not be possible for these languages. It should be marked that even though there are 33 non-native speakers, compared to the 19 native speakers, they are highly spread out in regards to first language background, therefore each language group have limited representation so one can not expect sufficient sample diversity.

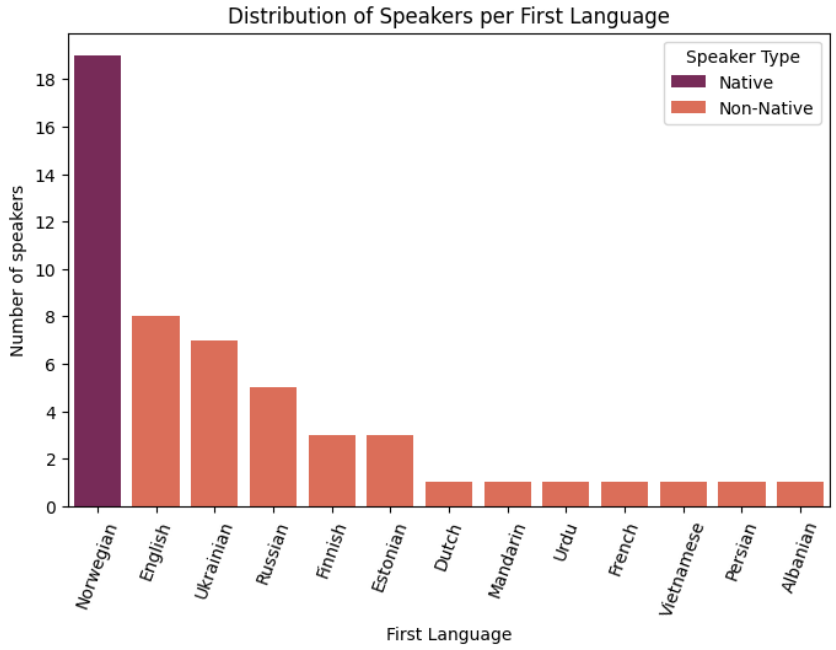


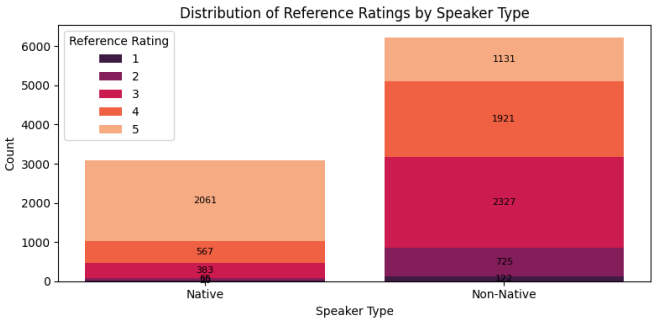

**Figure 4.1.4:** Distribution of speakers per first language.

## 4.2 Analysis of human assessment

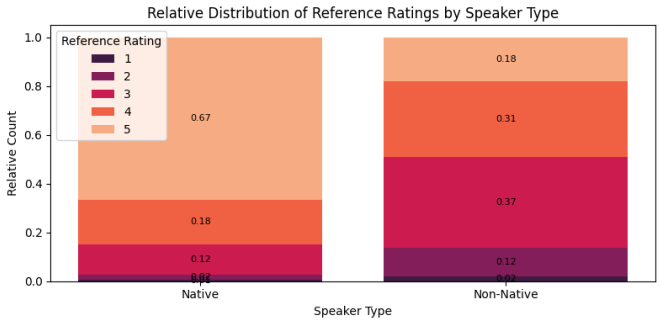
This section presents the human assessments in the dataset, providing valuable information on the pronunciation difficulties of native and L2 child speakers. These annotations are stated as reference ratings throughout this chapter and subsequently used as reference ratings in the multitask ASR and APA model training.

### 4.2.1 Reference rating - all data

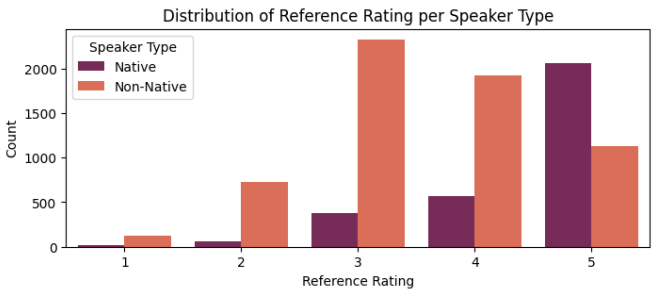
First, Figure 4.2.1 shows the complete overview of all reference ratings. As there is an irregular distribution between the number of native and non-native speech samples, the relative distribution is also shown. Looking at the absolute count, the majority of the speech samples were rated 3-5 by the human assessors. Both ratings 1 and 2 have very few utterances, but rating level 1 is especially underrepresented. From the relative count, it is apparent that most of the reference rating 5 scores are native speakers and that all other ratings have a higher number of non-native speakers.



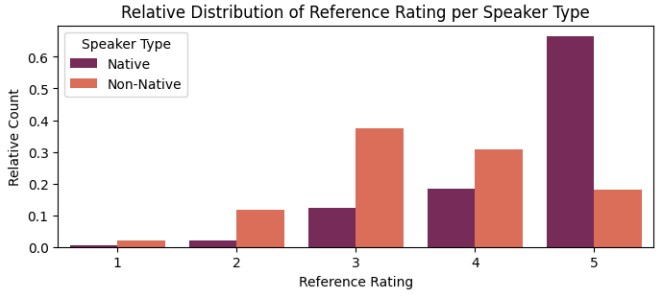

**(a)** Absolute count - stacked




**(b)** Relative count - stacked




**(c)** Absolute count - separated




**(d)** Relative count - separated

**Figure 4.2.1:** Absolute count and relative distribution of reference rating levels showing native and non-native speakers.

### 4.2.2 Speakers

The contrast between native and non-native speakers is noticeable in Figure 4.2.2, where the mean reference rating for non-native speakers is generally lower than for native speakers. However, there are exceptions where some native speakers have a lower mean reference rating than others. Speakers d08, d09, and d19 stand out, which will be recurring in these results. For native speakers d01-d06, the standard deviation is largely smaller than that of most non-native speakers.

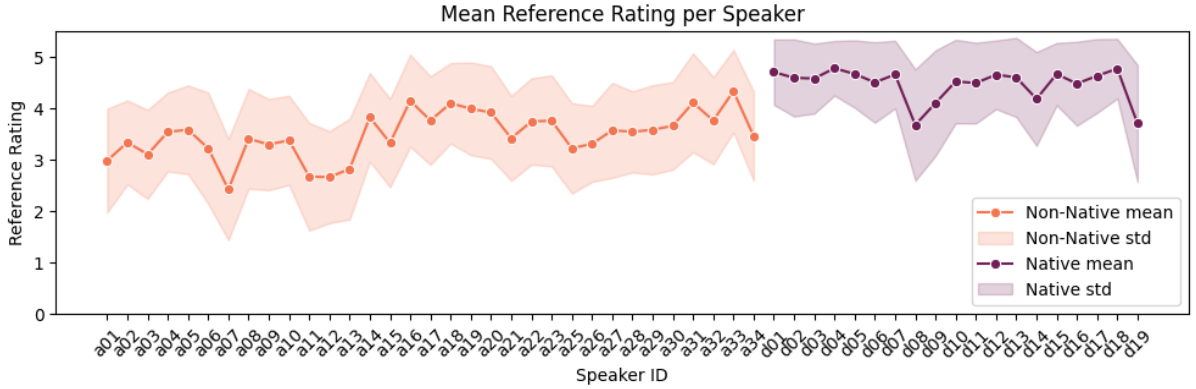



**Figure 4.2.2:** Mean reference rating for all speakers, showing standard deviation and speaker type.

### 4.2.3 Speaker age

Looking at the mean reference rating per age group in Figure 4.2.3, the distinction between native and non-native speakers is more apparent. For the native speakers, despite some variations, there is no clear correlation between age group and mean referenced rating. For non-native speakers, however, there is a clear rising trend where non-native speakers in age group 5 have the lowest mean reference rating and age group 12 have the highest. These results are consistent with the fact that speaker proficiency improves with age as the development of speech sounds advances and that proficiency in the mother tongue can affect proficiency in foreign languages as well. [45] For native speakers, the trend is more unclear, but this could simply be a result of the irregular distribution of speakers per age group.

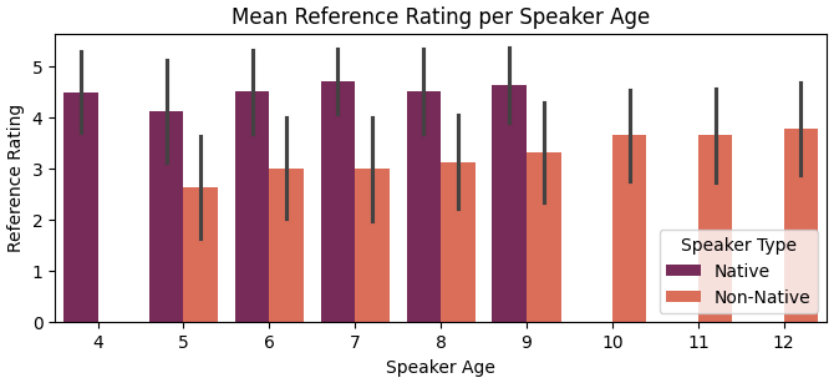



**Figure 4.2.3:** Mean reference rating for each age group, including standard deviation.

### 4.2.4 Features

For the additional annotations regarding prosdy, noise/disrution, pre-speech noise and repetition, one can compare the absolute and relative number of occurrences across each reference rating level. These features are marked 1 if there is a mistake or something not fully correct regarding this feature or to mark the presence of noise before or during the pronunciation of the target word.

In Figure 4.2.4, the distribution of reference ratings with the presence (1.0) or absence (0.0) of mistakes regarding prosody is presented. An important observation is that almost no utterances with a reference rating equal to 5 have been marked for flaws in prosody. On the other hand, for reference rating level 4, there is a high amount of speech samples with annotations for prosody mistakes. This observation could help understand why around 44% of utterances have no phoneme errors, but still, only 33.1% is given reference rating 5 as stated in [6]. The annotation of flaws in prosody could be used to mark that there are some overall pronunciations of rhythm or intonation that are not completely native-like, even though the individual phoneme sounds could be correct.

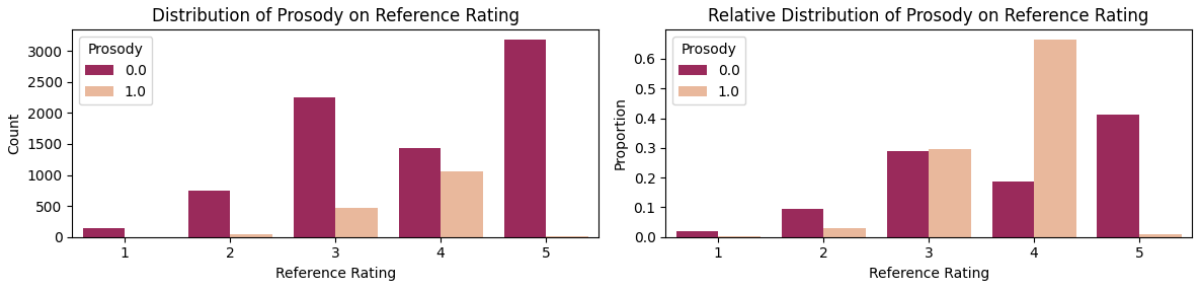



**Figure 4.2.4:** Distribution of reference rating, showing presence (1.0) or absence (0.0) of prosody.

In cases where there is noise or disruption throughout the recording, the annotators mark it. From Figure 4.2.5, it is clear that there are rather few utterances that contain such noise compared to the rest of the dataset, and when looking at the relative count, each reference rating level is approximately equal in regard to the presence or absence of noise. This shows that the human annotators are not highly affected by the overall noise in the recording when assessing each utterance, so one can assume that the rating is based on pronunciation or other factors than noise.

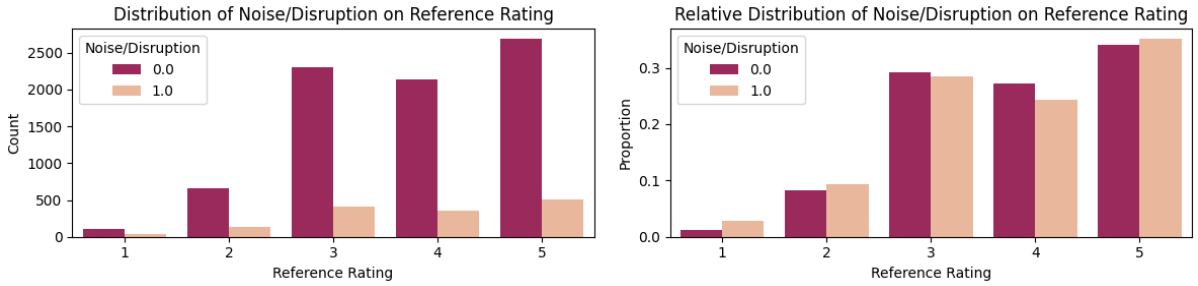



**Figure 4.2.5:** Distribution of reference rating, showing presence (1.0) or absence (0.0) of noise/disruption.

In general, there is a higher number of utterances that contain pre-speech noise, as seen in Figure 4.2.6, than for general noise throughout the full recording. For reference rating levels 3 and 4, the recordings are divided between the presence and absence of pre-speech noise. For reference rating level 5, there is a lower occurrence of pre-speech noise, so one can assume that the presence of such noise influenced the score of the human annotators. If the pre-speech noise was produced by the child speaker itself, such as additional speech, fill words, or laughing, this could affect the pronunciation of the first phonemes in the word, further resulting in a lower reference rating.

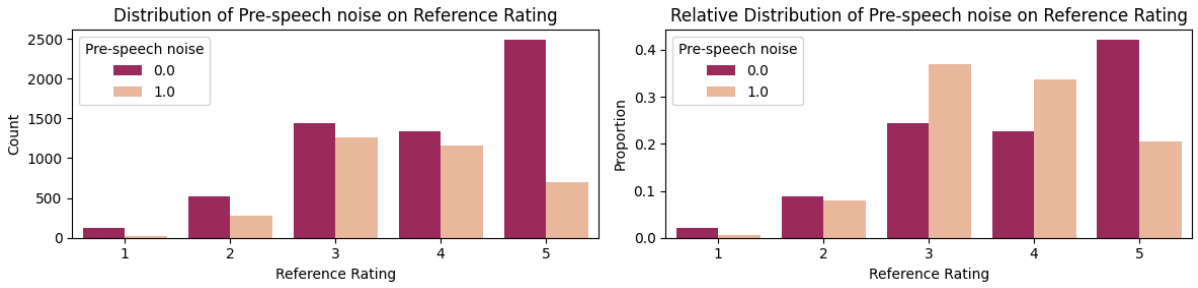



**Figure 4.2.6:** Distribution of reference rating, showing presence (1.0) or absence (0.0) of pre-speech noise.

The feature that no doubt has the lowest number of occurrences is repetition, where the absolute count is almost non-existent compared to the other recordings in Figure 4.2.7. However, looking at the relative distribution,

for reference ratings 2 and 3, the proportion of recordings with repetition is higher than those that do not. For levels 4 and 5, there is a higher proportion of repetition absence but still a significant amount for presence. One can, therefore, observe that repetition in a recording does not affect the human assessment rating as much as prosody and that one can still achieve a high rating with repetition present.

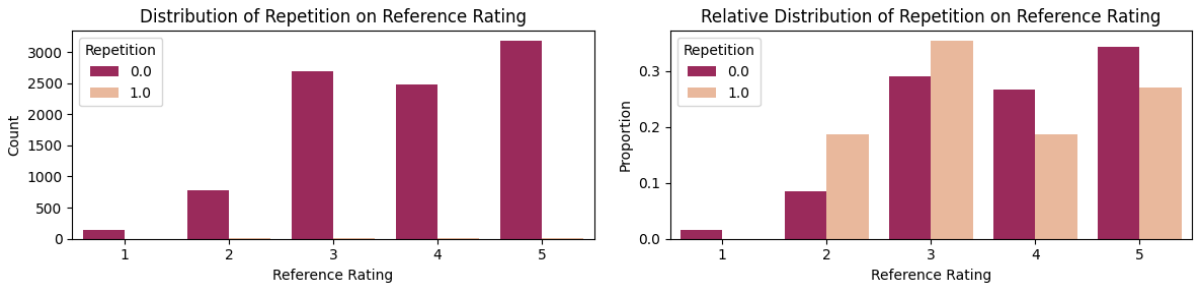



**Figure 4.2.7:** Distribution of reference rating, showing presence (1.0) or absence (0.0) of repetition.

### 4.2.5 Target words

Figure 4.2.8 shows that across all target words, the mean reference rating is higher for native speakers than for non-native speakers. In addition to this general trend, the mean reference rating for each target word is generally even. There are some variations, and the standard deviation is high, but for the majority of target words, the mean reference rating is between 3 and 4.5. This shows that there are both high-level and lower-level pronunciations for each target word, providing good representation, which is helpful for achieving good generalization when training the APA model on these words.

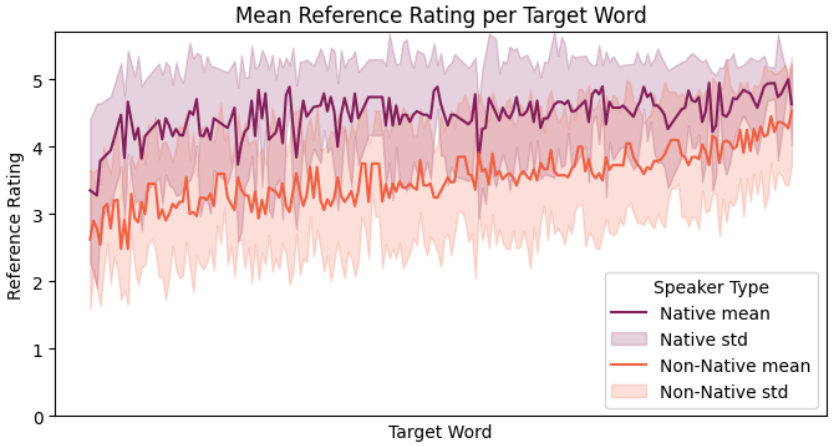



**Figure 4.2.8:** Mean reference rating for each target word, showing standard deviation and the difference between native and non-native speakers.

### 4.2.6 Phonemes

Each utterance is annotated at the phoneme level with a binary score signifying correct or incorrect phoneme pronunciation. Looking at the mean phoneme score from the human annotations in Figure 4.2.9, the majority of the phonemes have a mean rating between 0.7 and 0.9. Similar to the target words, this shows that there is a good representation of both correct and incorrect pronunciation of each phoneme, but also that they are mostly correct.

The phoneme with the lowest mean score is /æʉ/ with a mean of 0.35. This phoneme only occurs in one target word, namely *fortau*, and therefore has only 51 occurrences in the dataset. The low mean error rate can be caused by the limited amount of diverse utterances. Comparing to the mean reference rating for the word *fortau*, which is 3.2, and inspecting all phoneme ratings for this target word. There are examples that indicate that the pronunciations for all phonemes are not correct, but there are also several examples where only this phoneme is marked as incorrect. If these results arise due to the difficulty of combinations of phonemes in this word or that this phoneme is especially hard in itself, it is hard to determine.

For the second lowest mean phoneme score, /ç/ with 0.41 mean, the case is rather different. This phoneme is present in 156 utterances, as the start of words such as *tjern*, *kjerne* and *kikkert*, so there is a much more diverse representation of this phoneme. These target words have a mean reference rating of 3.0, 3.5, and 3.6, respectively. Studying phoneme level scoring for these words, it is likely that the /ç/ phoneme itself is hard to pronounce, consequently affecting the mean reference rating of the word rather than the other way around.

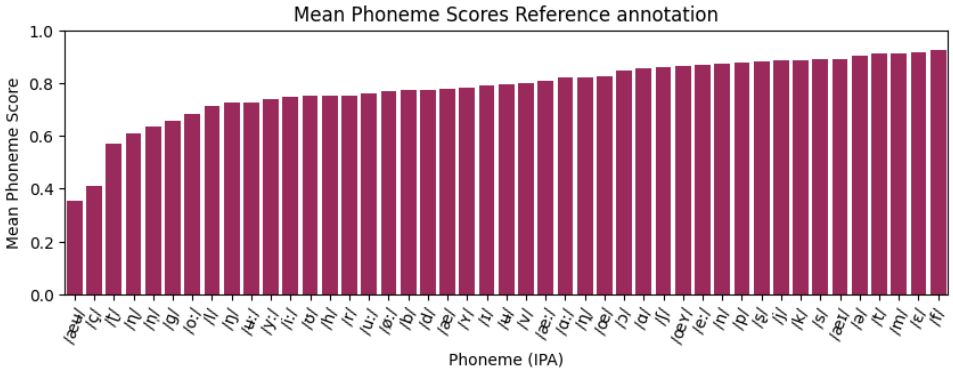



**Figure 4.2.9:** Mean phoneme score (binary) based on phoneme level reference annotations.

### 4.2.7 First Language

Grouping the utterances by the first language of speakers, the mean reference rating gives the plot in Figure 4.2.10. Here, only languages with multiple speakers are shown. Native Norwegian speakers have a mean reference rating of around 4.4, and English is the language with the second-highest mean reference rating. The fact that speakers with English and Estonian as first languages have the best pronunciation is expected, as these languages have a lot of shared sounds with the Norwegian language, and research has shown that L2 learners have increased proficiency in languages that contain the same sounds as their mother tongue. [4, 46, 45] This also corresponds to Russian being the first language with the greatest pronunciation difficulty of Norwegian, as that is the language with the fewest shared sounds with the Norwegian language. The speaker's proficiency in their mother tongue could also be a misleading source of errors, as their age is diverse, but this will be further discussed in Section 5.2.

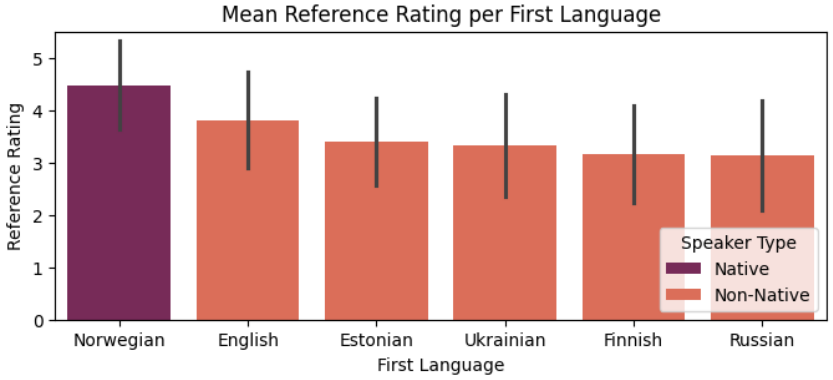



**Figure 4.2.10:** Mean reference rating per first language with standard deviation, only showing languages with multiple speakers.

## 4.3 Analysis of prediction error from APA

This section will analyze the prediction error for the APA results from the multitask ASR and APA work done by Aalto University for the Teflon project. To get an overview of the predicted ratings, some confusion matrices between reference and predicted ratings will be presented. However, the main focus of this chapter is to analyze where the APA model does not correctly predict global pronunciation ratings. Therefore, the prediction error, being the deviation between the reference rating and the predicted rating, will be used for the majority of the plots. The results are structured around the same groups within the dataset, such as speaker type, ID, age, features, target words, phonemes, and first languages, similar to the structure of results regarding the human assessments.

### 4.3.1 All data

Dividing the dataset by speaker type into native and non-native speakers and comparing the reference ratings and predicted ratings give a good overview of the performance of the APA system. Figure 4.3.1 shows confusion matrices for native speakers; on the left is the absolute count for each rating level, and on the right are the normalized values. The overrepresentation of rating 5 utterances is apparent, and the APA model correctly predicts most of these speech samples. For all rating levels, the predicted scores are biased towards giving higher scores. There are more recordings that get a higher predicted rating than their reference rating, compared to the reverse happening. For example, most level 2 reference ratings are predicted to rating level 3 by the model. This trend is intentionally designed as the APA model is used in the gamification of a CALL system. Here, it is important to motivate the children to continue playing and learning; if a speaker constantly gets a lower predicted rating even though the pronunciation might be good, they can get frustrated and not want to play the game. It is important to keep this design in mind when analyzing predicted ratings. The level 3 reference ratings are the most scattered in regard to prediction ratings, with ratings dispersed between prediction levels 3 and 5. Hence, the native speech samples do have mismatches between the reference rating and the predicted ratings. But overall, there are most utterances with reference rating 5, which are well estimated by the APA model.

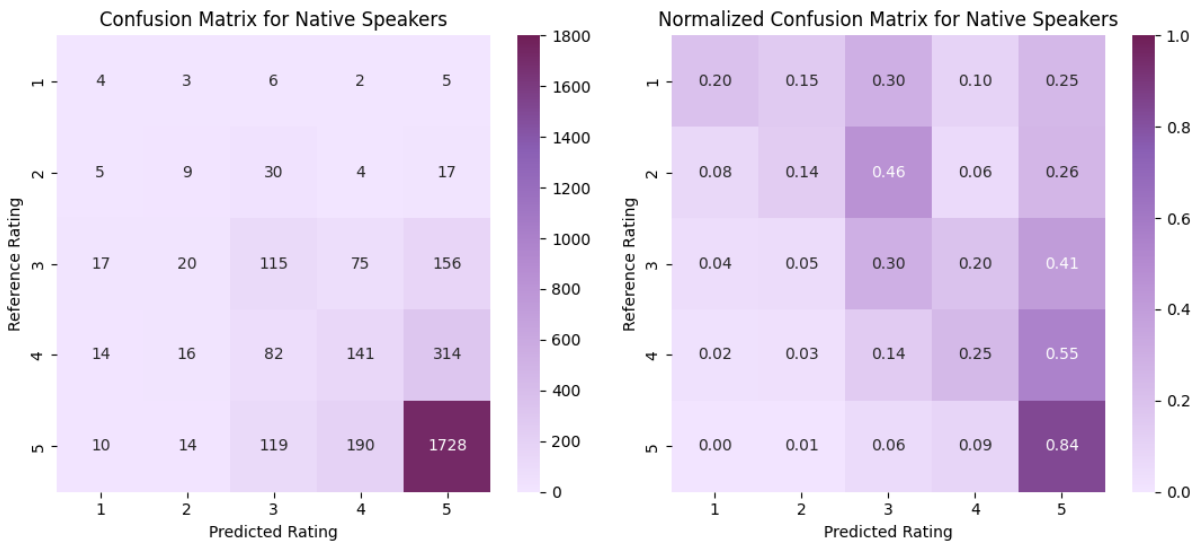



**Figure 4.3.1:** Confusion matrix between reference and predicted ratings for native speakers, both absolute count and normalized values.

For non-native speakers, 37% of the reference ratings are level 3, and we see more distributed predicted ratings in Figure 4.3.2. The APA system still tends to overrate all levels. For each reference rating, 53% to 60% are correctly predicted. As reference ratings 3 to 5 are better represented among non-native speakers compared to native speakers, the model also predicts these levels more accurately than native speech samples.

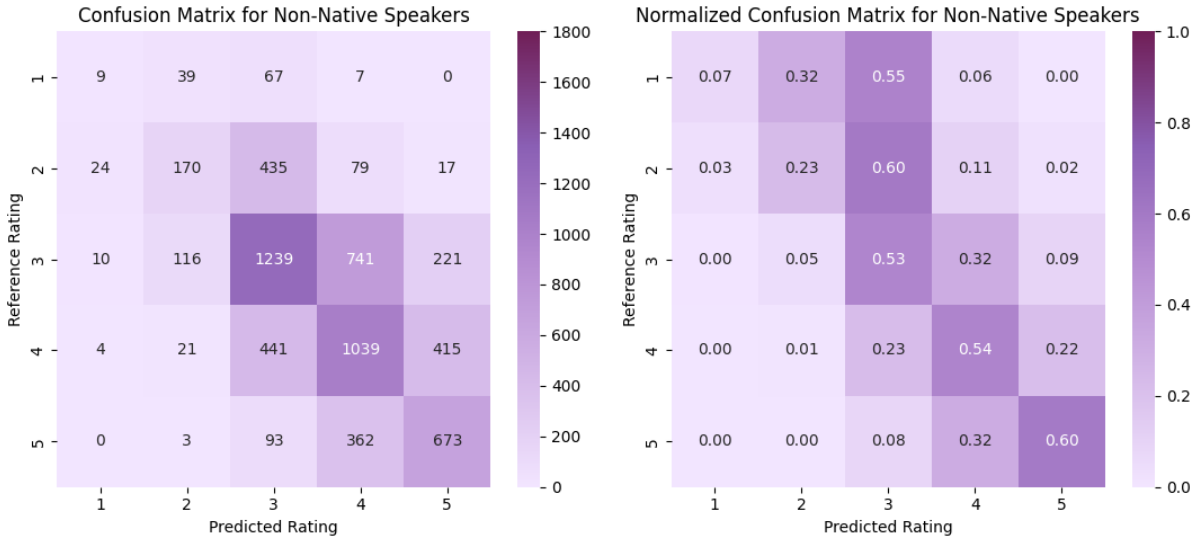



**Figure 4.3.2:** Confusion matrix between reference and predicted ratings for non-native speakers, both absolute count and normalized values.

Figure 4.3.3 shows the relative distribution of the prediction error levels. As the reference ratings range from 1-5, the prediction error ranges from -4 to 4, where a positive prediction error indicates a higher prediction rating than the reference rating, and negative errors are the opposite. The APA model performs better on native speakers than on non-native speakers. This could be due to the high amount of level 5 ratings, which have a better articulation of Norwegian sounds and give the model a larger amount of training data for that level. Given that the model is already fine-tuned for Norwegian speech, recognition of the correctly pronunciated sounds might be easier for the model. Where first languages with similar sounds to Norwegian had a better mean reference rating, the opposite could affect the prediction error. If a non-native speaker utters sounds from their first language that is not present in Norwegian, the model will not be able to recognize these. The prediction errors are centered around 0, with mostly only -2 to +2 errors. There are only a few cases of -4,-3 or +4, +3 prediction errors, some of these will be discussed later in this chapter.

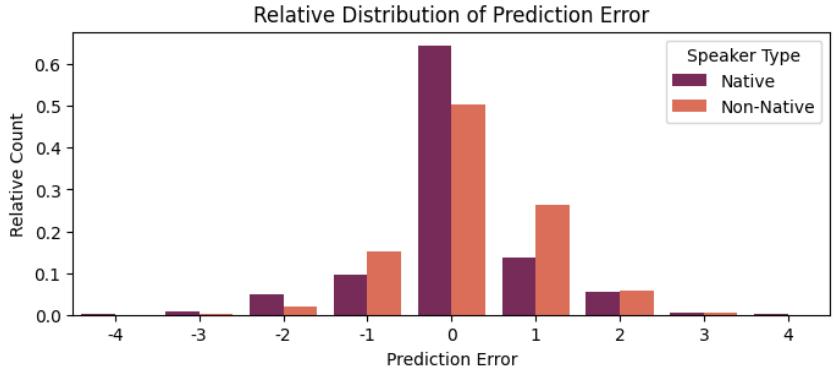



**Figure 4.3.3:** Distribution of prediction error levels for all data, showing relative distribution between native and non-native speakers.

### 4.3.2 Speaker

To understand whether the APA model performs better on selected speakers, Figure 4.3.4 shows mean prediction error per speaker ID. An interesting result is that two native speakers, d08 and d09, have a significantly low mean prediction error. These two speakers also had low mean reference ratings, and by further inspection, several of their speech recordings included additional speech, laughing, or additional noise. It seems like these speech samples are so dissimilar to the majority of the dataset that the prediction error is very negative. For the human reference ratings, some of these disturbances in the recordings might have been disregarded, and only the pronunciation of the target word has been assessed. The APA model on the other hand, can not recognise what is the target word and what is additional speech or background noise, therefore the recording is rated on the whole, and the pronunciation will not be similar to other recordings of that target word.

Despite these two cases, the prediction error is quite even for the rest of the speakers. There are some variations of zero mean prediction error and some up to 0.5, but these are the same for both native and non-native speakers. There is no clear trend showing that the model performs significantly better on selected speakers. On the speaker level, there is also a clear trend of more positive mean prediction errors rather than negative ones.

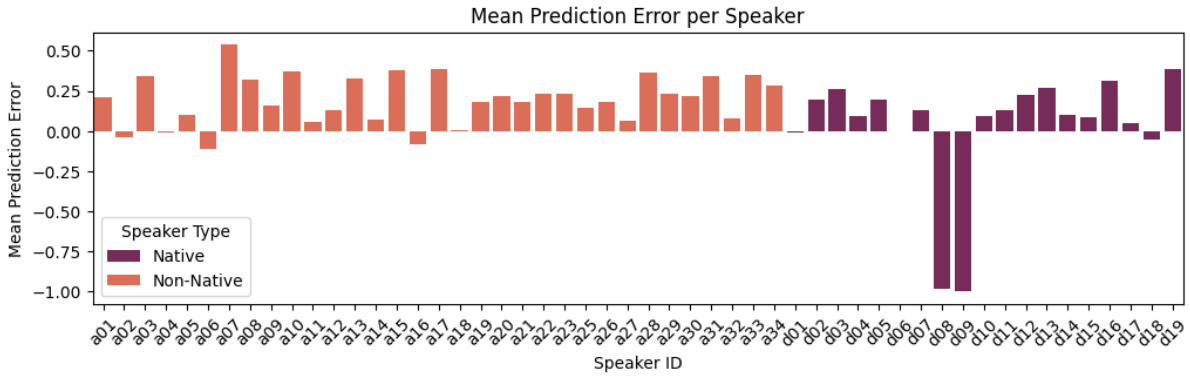



**Figure 4.3.4:** Mean prediction error for each speaker.

### 4.3.3 Age

For all age groups, the prediction error is equal to zero for the majority of the utterances, as shown in Figure 4.3.5. Positive prediction errors are more prominent than negatives.

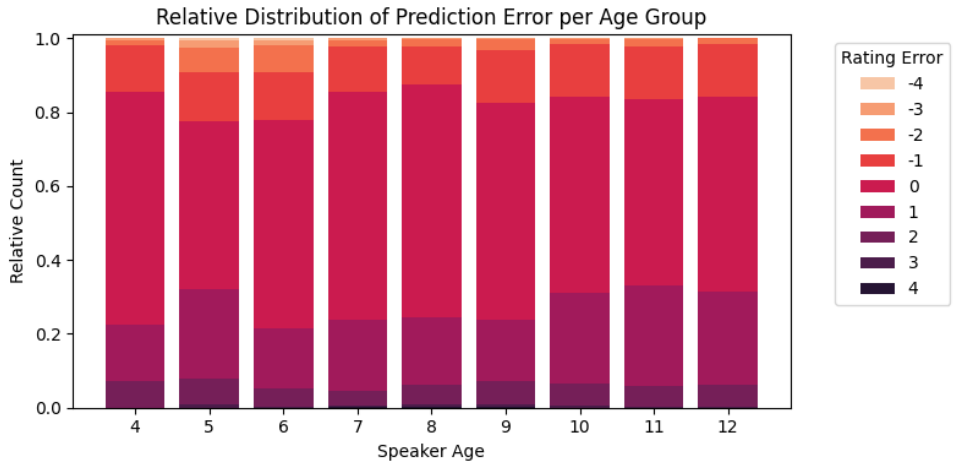



**Figure 4.3.5:** Relative distribution of prediction error for each age group, showing all levels of error.

Studying Figure 4.3.6, the lowest age groups 5-7, excluding age 4 as this is only one speaker, the mean prediction error is higher. For age groups 8-9, it is lower, but for the oldest age groups 10-12, the mean prediction error rises again. The correlation here could be a combination of both speaker age and speaker type. The lower age groups have more native speakers, but as they are still quite young, there is a higher variability in pronunciation, making the training data more in-concise in what pronunciation correlates to these groups, causing prediction errors. For the middle age groups, there is a good representation of both native and native speakers, and as they are older, one can assume good representation in the dataset, with both good pronunciation from the native speakers but also examples of pronunciation mistakes, providing the model with valuable training data to learn these rating levels.

For the oldest groups, there are only non-native speakers. Therefore, one can assume a greater proficiency in their first languages, increasing the possibility of their pronunciation being affected by sounds not present in the Norwegian language. The model is not trained on these sounds, which could increase confusion for the APA model and consequently give higher prediction errors. It is important to note that the prediction error is not reliant on the goodness of pronunciation of the speakers, rather the performance of the model on the

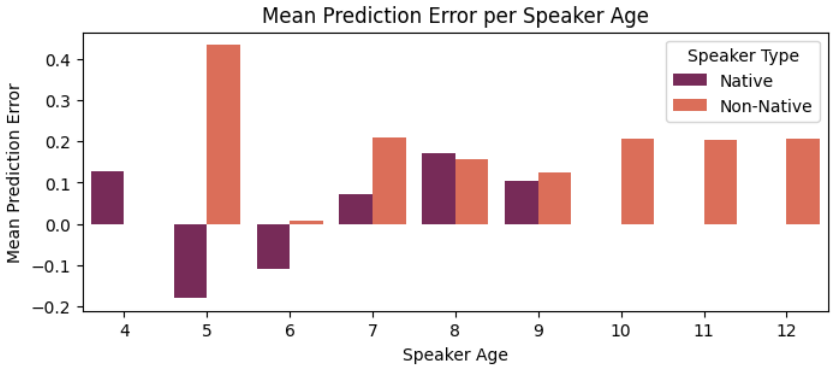
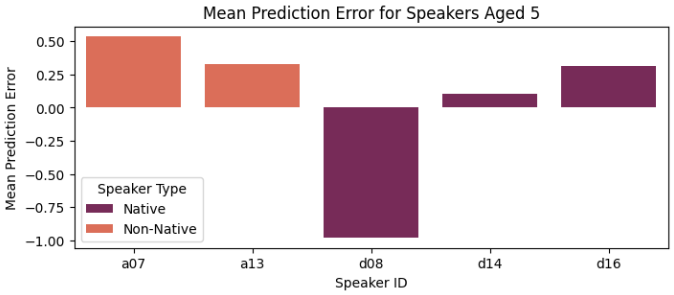


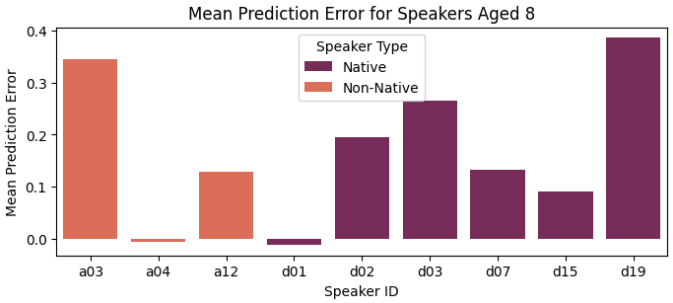

**Figure 4.3.6:** Mean prediction error per speaker age group, showing both native and non-native speakers.

different speaker groups can be affected by the representation of variation in the dataset.

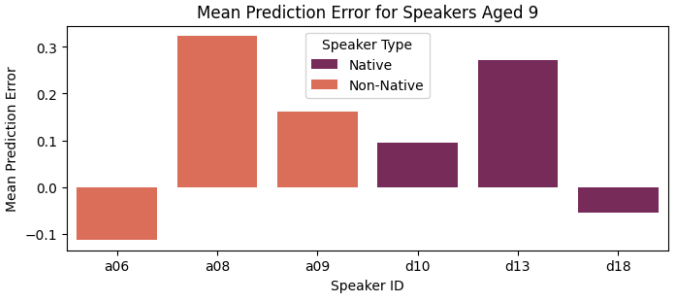
Looking at the distribution between speaker types for each age group, there is a high difference between native and non-native speakers in age groups 5 and 6. These age groups include speakers d08 and d09, respectively, which is largely contributing to the negative mean prediction error. Figure 4.3.7 depicting detailed mean prediction error for age groups 5, 8, and 9. These plots show that age groups 5 and 8 have native speakers that largely contribute to a higher positive or negative mean prediction error, while age group 9 has a more even distribution of mean prediction error among both native and non-native speakers.




**(a)** Speakers aged 5, showing that speaker d08 has the highest negative mean rating error, even though this is a native speaker.




**(b)** Speakers aged 8, showing that speaker d19 has the highest positive mean rating error, even though this is a native speaker.




**(c)** Speakers aged 9, showing a more expected distribution where the non-native speakers have higher mean prediction error, but still variation within the group.

**Figure 4.3.7:** Mean prediction errors for age groups 5, 8, and 9.

To further analyze performance the intragroup performance per age group, normalized confusion matrices in Figure 4.3.8 are presented. For these plots, it is important to remember that age group 4 only has one native speaker, and age groups 10-12, while having several speakers, are only non-native. The APA system performs best for level 5 ratings in age groups 7-9. Ages 10-12 have a better performance on levels 3-5 compared to the lower ages 5 and 6.

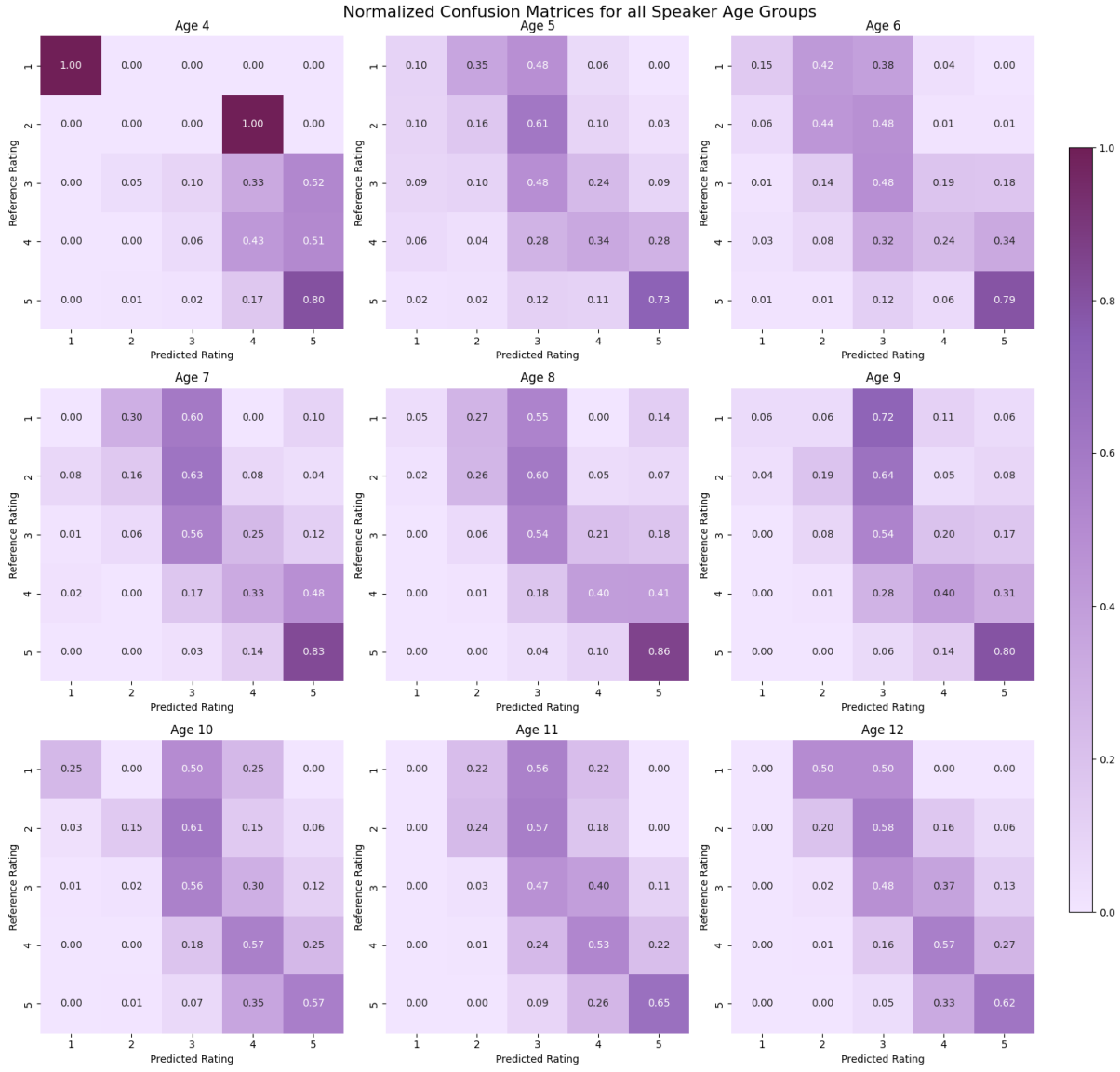



**Figure 4.3.8:** Normalized confusion matrixes for reference and predicted ratings, separated by age groups.

### 4.3.4 Features

In addition to reference ratings, the features annotated by human assessors can also provide insight into the performance of the APA predictions. Figure 4.3.9 shows the probability of prediction error levels given the presence (1.0) or absence (0.0) of each feature. In these plots, the change in probability of prediction error equal to zero, and the spread across prediction error levels provides valuable insight. If all features are marked as present, this results in more prediction errors, as seen from the lower probability of prediction error equal to zero.

For noise/disruption and repetition, the probability of presence is spread out across prediction error levels, and the probability of high positive or negative errors is higher if noise or repetition is present. For prosody mistakes or pre-speech noise in the recordings, the distribution is more concentrated around low prediction errors, but the presence of these features still increases prediction errors. This means that noise and repetition in the recording affect the performance of the automatic assessments more than prosody and pre-speech noise. These results are valuable because they are different from how features affected the human reference ratings, where prosody and pre-speech noise affected the rating the most. Again, we are analyzing the prediction error and not the prediction rating, but it still seems as though human annotators disregarded noise and repetition in the recording to a higher degree. However, these factors disturb the performance of automatic assessments. Ensuring quiet recording environments and editing out repetition in the recordings where possible could, therefore, be worthwhile to improve APA performance.

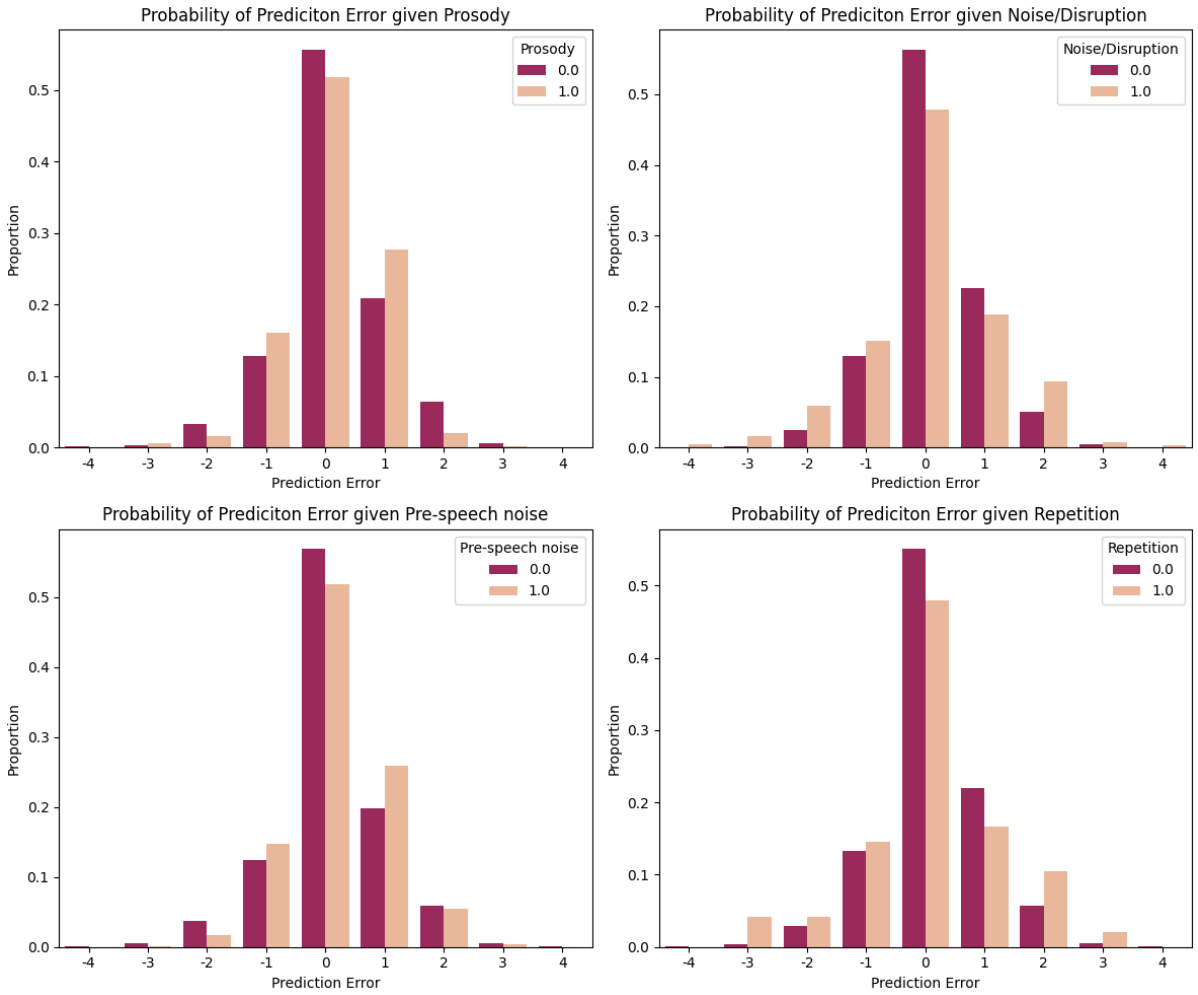
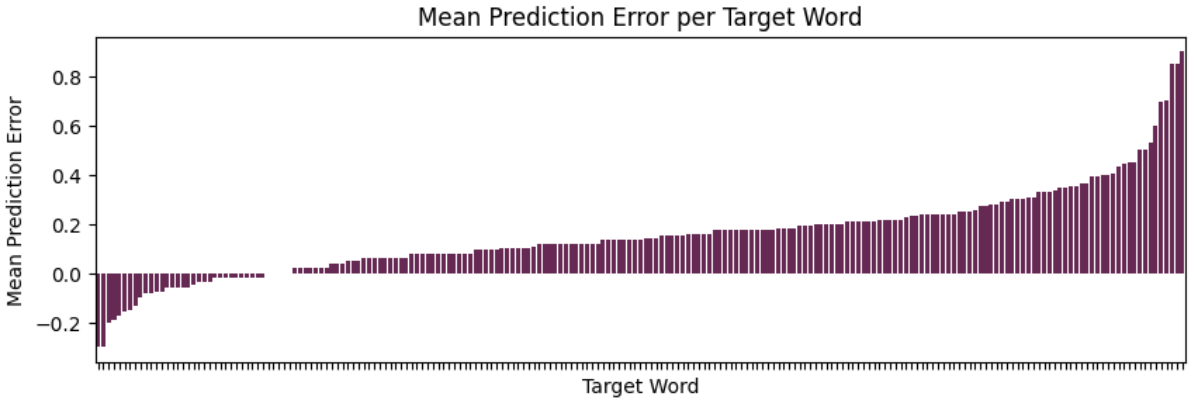



**Figure 4.3.9:** Probability of prediction error given presence (1.0) or absence (0.0) of features; prosody, noise/disruption, pre-speech noise, and repetition on prediction error.

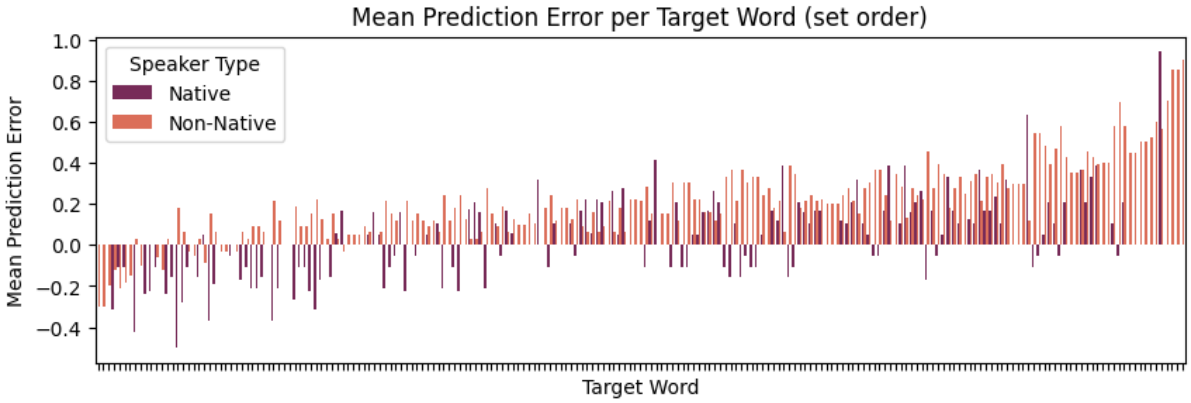
### 4.3.5 Target word

The mean reference ratings per target word were rather even, and this is also the case when analyzing prediction error per target word. Figure 4.3.10**(a)** shows an expected trend for mean prediction error per target word, where most words have positive prediction error, as wanted for the field of application. Keeping the order of the target word on the x-axis the same, 4.3.10**(b)** shows mean prediction error for both native and non-native speakers within each target word. The details of this plot are not easily visible, but the general trend is interesting. We see that most of the negative mean prediction errors are from native speakers, and the majority of the high positive mean prediction errors are from non-native speakers. From this plot, it seems that the APA underestimates native speakers and overestimates non-native speakers, but in actuality, a simple practicality is most likely the cause of this trend. Because most native utterances are rated high, with many reference ratings at 5, if the automatic assessment makes a mistake, that mistake has to be lower than 5 because the system is limited to only giving scores between 1 and 5. It is the opposite for non-native speakers, as most of them have lower ratings, and the possibility of the system overestimating the score is higher, giving positive prediction errors.

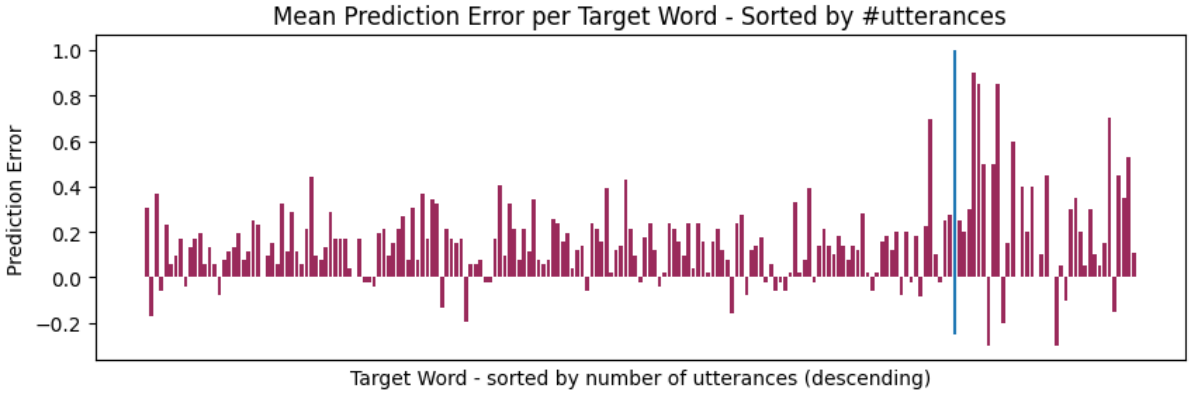
In addition, we know that several target words only have around 20 utterances from non-native speakers. Studying 4.3.10**(c)** where the target words are sorted by number of utterances, making all target words with 50 utterances, including both native and non-native speakers to the left of the blue line, and all target words with only 20 utterances of non-native speakers, there is a clear distinction between the mean prediction error. This could mean that the system performs worse on target words that only contain non-native speakers, which would emphasize the need for a more inclusive dataset. On the other hand, this could just be a result of the number of utterances per target word, where the model has not learned the target words with fewer utterances well enough, which argues for the importance of high quantities of speech recordings for each target word.




**(a)** Mean prediction error for all target words, showing a higher degree of positive prediction error, rather than negative.




**(b)** Mean prediction error for all target words, showing speaker type. Target words with mostly native speakers have a higher negative mean prediction error, and target words with mostly non-native speakers have a higher positive mean prediction error

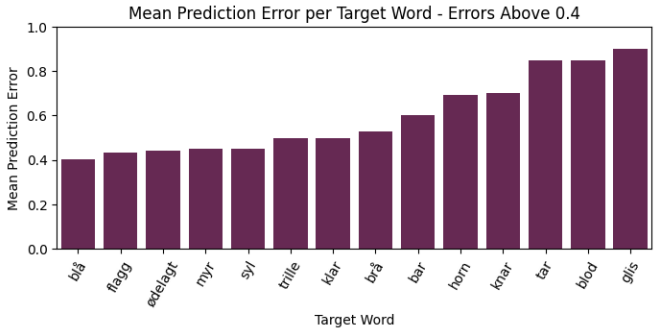



**(c)** Mean prediction error for all target words, sorted by number of utterances per target word. The blue line distinguishes between target words with around 50 utterances to the left and words with only around 20 non-native samples to the right.

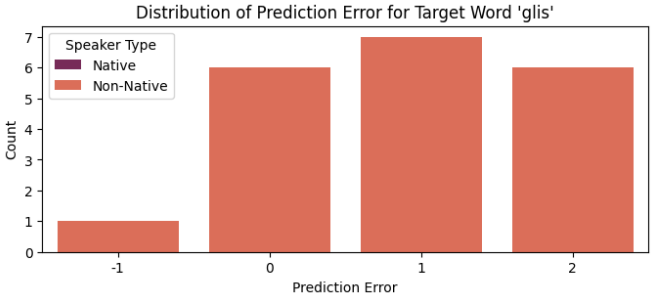
**Figure 4.3.10:** Mean prediction error for all target words.

To further corroborate these observations, detailed plots of the target words with the highest positive and negative mean prediction errors are introduced. Figure 4.3.11**(a)** presents all target words with a mean prediction error above

0.4. Of these words, all but four have only 20 utterances of non-native speakers in age groups 10-12. The target word *glis* has the highest mean prediction error of 0.9, and 4.3.11**(b)** shows the distribution of prediction error levels, where the highest prediction error is plus 2.



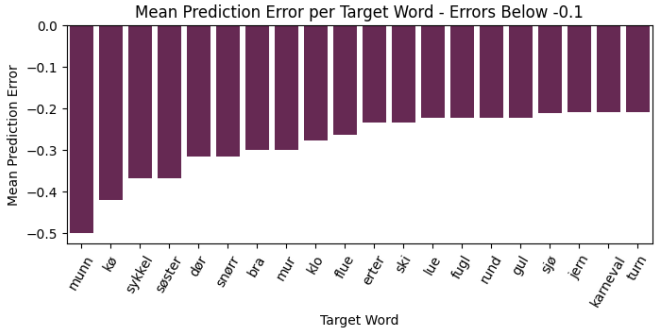

**(a)** Showing target words with highest positive mean prediction error (above 0.4).



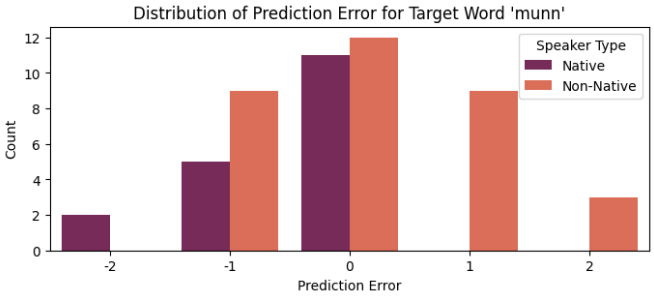

**(b)** Prediction error levels for the target word with highest positive mean prediction error; 'glis'. Showing that level +2 of prediction error is the highest for this target word, and only non-native speakers

**Figure 4.3.11:** High positive mean prediction error detailed plots.

Figure 4.3.12**(a)** gives the same detailed information for the highest negative mean prediction errors below -0.1. For these utterances, most have 52 recordings, meaning that most have both native and non-native speakers with an even distribution of age groups. The target word *munn* has the lowest mean prediction error at -0.5, 4.3.12**(b)** shows a bias towards negative prediction errors, but this is likely due to the number of native speech samples that have a high reference rating, resulting in negative prediction errors where the APA model is not correct.




**(a)** Showing target words with highest negative mean prediction error (below -0.1).




**(b)** Prediction error levels for the target word with highest negative mean prediction error; 'munn'. This shows that -2 is the lowest prediction error with only native speakers, and the highest prediction error is +2 with only non-native speakers.

**Figure 4.3.12:** High negative mean prediction error detailed plots.

### 4.3.6 Phonemes

The APA system does not predict phoneme-level annotations. Therefore, it is challenging to analyze the correlation between phoneme scores and prediction errors. Figure 4.3.13 shows how the number of correct phonemes in a target word, normalized by word length, relates to the prediction error levels. In this plot, the highest and lowest pronunciation error levels (+4 and -4) only have native speakers, and there is an asymmetrical distribution where all negative prediction error levels have a high amount of correct phonemes, whereas positive prediction errors have a descending amount of correct phonemes. This effect is likely due to the already discussed fact that most negative pronunciation errors come from native speakers, where the reference rating is generally high, with most phoneme scores being correct, so there are other factors in the pronunciation, such as noise or repetition that cause the model to underpredict.

For the positive pronunciation errors, however, performance difficulty increases when the number of correct phonemes decreases. This could be because these recordings might contain foreign sounds that the model has not been trained on or because the number of recordings with a low percentage of correct phonemes is small. As we know, recordings with reference ratings 1 and 2 are scarce, making it hard for the model to learn the traits of these scores.

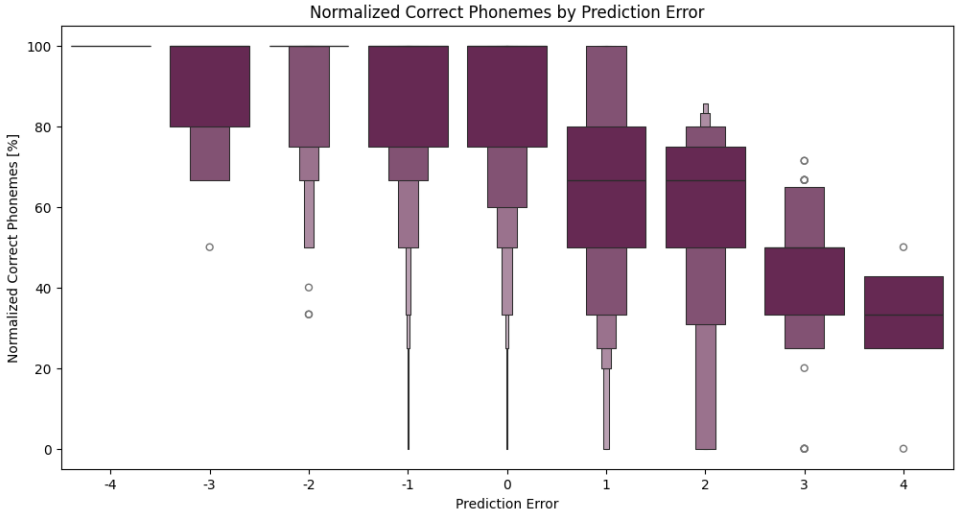
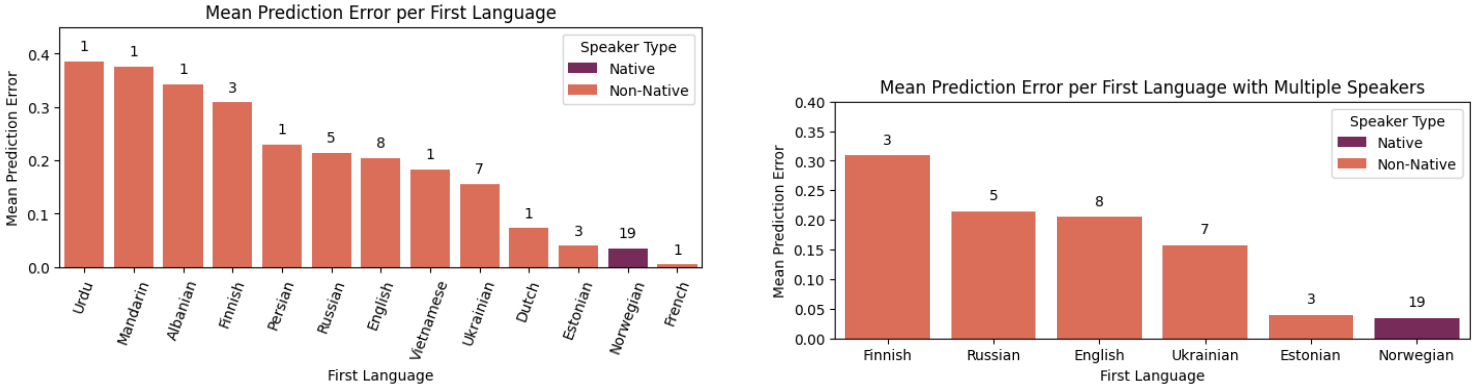



**Figure 4.3.13:** Showing how the number of correct phonemes, normalized by word length, affect prediction error.

### 4.3.7 First language

Figure 4.3.14**(a)** presents the mean prediction error per first language, where the number above each bar states the number of speakers for the corresponding language. As analysis on first languages with only one speaker can be misleading due to speaker-dependent proficiency, 4.3.14**(b)** shows only languages with multiple speakers. This figure shows that Finnish is the language with the highest mean prediction error at 0.3, and Estonian has close to the same mean prediction error as Norwegian at 0.04.




**(a)** All languages

**(b)** Languages with multiple speakers

**Figure 4.3.14:** Mean prediction error for each first language, with numbers indicating how many speakers per language.

There are likely two reasons for these results. The first is the number of speakers for each language. When one language is underrepresented in the training data, the performance of the model will be inadequate. This could explain why Finnish has a higher mean prediction error, but it does not explain why the model performs well for Estonian first-language speakers and poorly for Russian. This leads to the second reason, which is the number of similar sounds between Norwegian and the other languages. The Russian language has the least number of sounds similar to Norwegian, so there is a higher chance of having recordings where non-native speakers are pronouncing sounds that the model has not been trained on. For Estonian, there are many sounds that are similar to Norwegian, which the model can easily recognize. Figure 4.3.15 supports this theory by showing that languages such as Estonian and English, which have the highest amount of similar sounds to Norwegian, have the highest number of prediction errors equal to zero. Meanwhile, other languages have more distributed prediction errors.

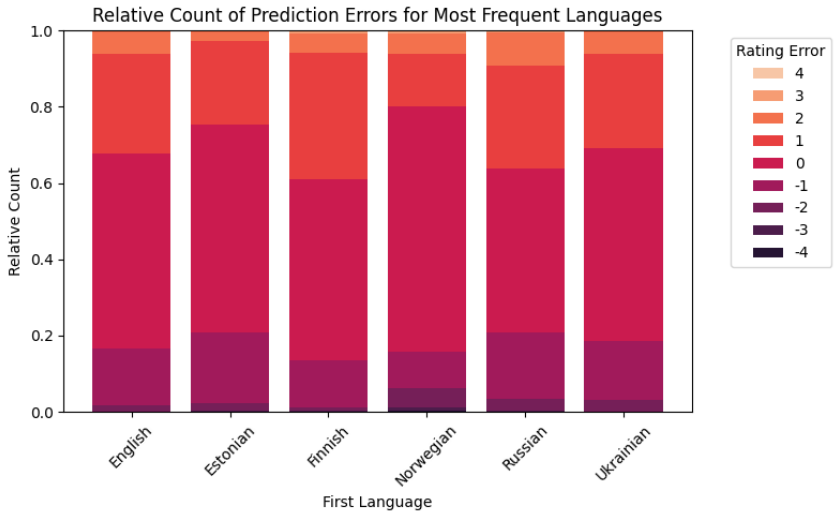
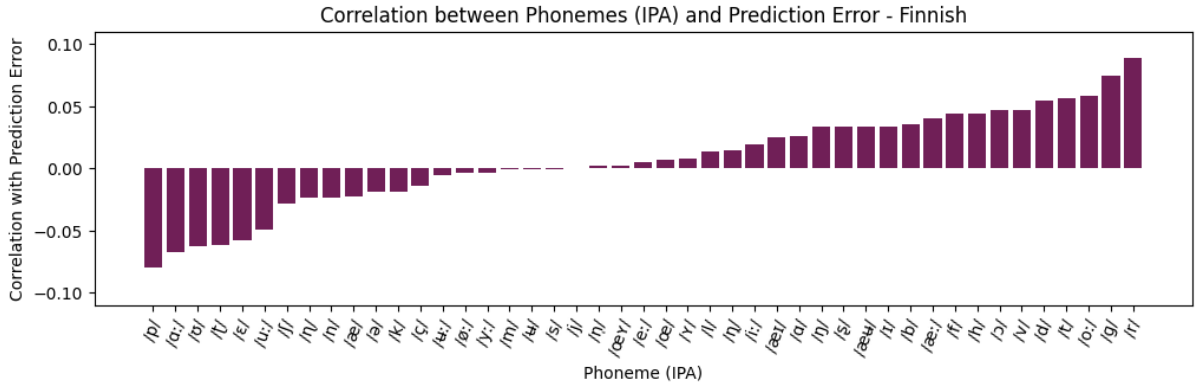


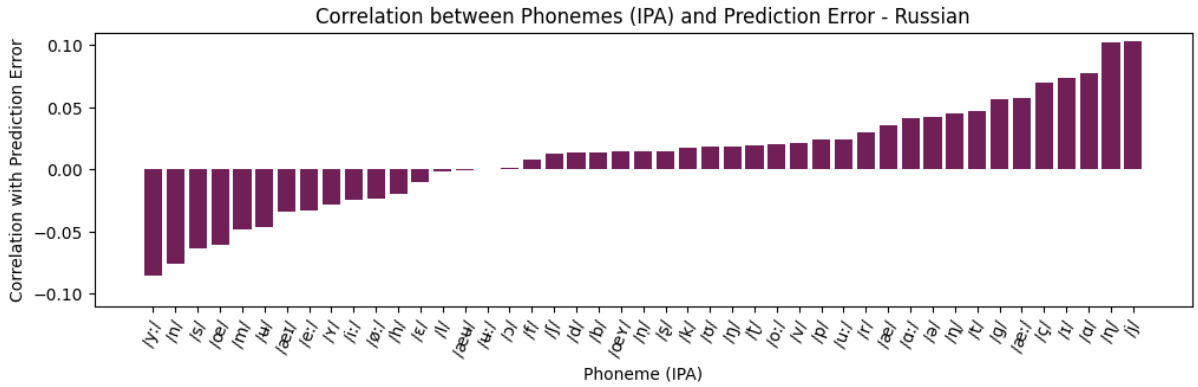

**Figure 4.3.15:** Relative count of each prediction error level for most frequent first languages.

The three languages with the highest mean prediction error, Finnish, Russian, and English, have been studied closely in terms of phonemes. Figure 4.3.16 shows the correlation between phoneme presence in a target word and the prediction error. A high positive correlation means that if this phoneme is present in the correct pronunciation of the word, it is more likely to have a positive prediction error. If the correlation is negative, the presence of the phoneme in the target word is more likely to result in a negative prediction error. Consequently, if the correlation is around zero, it means that the prediction error is more likely to be zero. Note that this does not give information about the pronunciation of the phoneme, only the performance of the APA model. All information about the similarity in sounds between languages used in this section is from [4] or provided by the author Koreman directly.

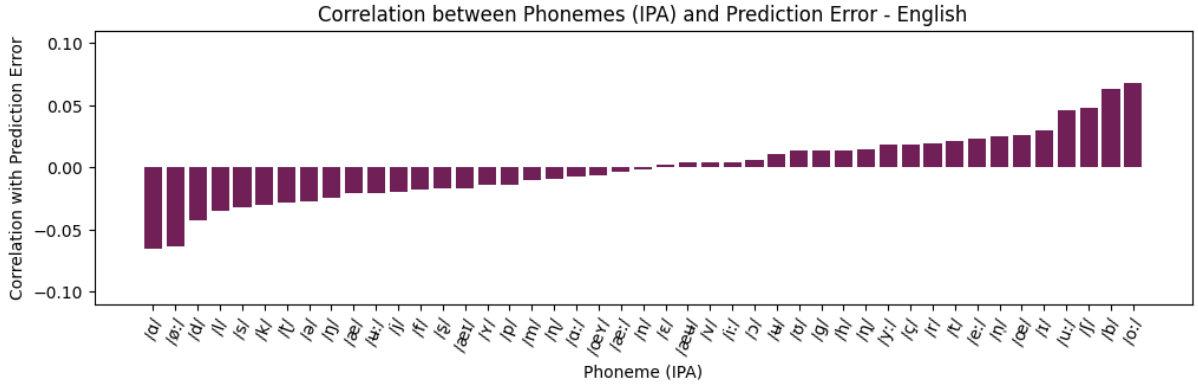
For the Finnish language, in Figure 4.3.16**(a)**, the phonemes /m/, /s/ and /j/, the correlation to prediction error is around zero, and these are all similar sounds between Norwegian and Finnish, so the model has been trained on these sounds through the Norwegian data, so will perform good predictions on these sounds. Phonemes /p/, /ʈ/ and /ʊ/ give a negative correlation and do not have similar sounds in the Finnish language, making the possibility of the speaker pronouncing an unknown sound for the APA system higher.




**(a)** Finnish




**(b)** Russian




**(c)** English

**Figure 4.3.16:** Correlation between presence of phoneme in target word and prediction error for languages with highest mean prediction error.

However, there are contradictions to this idea, where for example /g/ and /d/, which appear both in the Norwegian and Finnish languages, have a positive correlation to the prediction error. Surrounding phonemes in an utterance can still affect the pronunciation of the speaker, resulting in unknown sounds for the model, but the reason for this positive correlation could also simply be due to sparse data containing these phonemes or other factors in the recording affecting the prediction error.

Similar observations are found for Russian in 4.3.16**(b)** and English in 4.3.16**(c)**. The Russian language does not have a corresponding sound for the Norwegian phonemes /yː/ and /n/, so the pronunciation of these sounds from Russian speakers could result in sounds that are unknown to the APA model. However, sounds like /n/ and /j/ are present in both languages, so the model should have learned these. Why these sounds have a higher negative and positive correlation to the prediction error is uncertain but is most likely due to other factors in the respective speech samples.

For English, all phonemes have a lower correlation to prediction error. This is likely due to the high similarity between Norwegian and English sounds, so the model has been sufficiently trained on these sounds.

## 4.4 Training new multitask ASR/APA model

The second part of this work consists of training a new multitask ASR and APA model to explore the possibility of improving performance by using another base model. This section gives the results of the CER comparison of several wav2vec2.0 [7] based models, which will be the basis of selecting the base model for the multitask ASR APA training. During this work, several speech sample outliers were found based on ASR transcriptions and CER results. A list of these outliers is provided to help improve future training by either omitting these samples or to help explain unexpected results.

### 4.4.1 Findings outliers

This list presents a subset of outliers found based on extremely high CER and unusual ASR transcription. Most of these recordings contain additional speech, laughter, high noise, or completely wrong intonation.

Some examples are d08_bror where the child says the target word *bror* pretty good, but then there is additional speech where he says he will be a big brother *"jeg blir storebror"*. Recordings like these are unwanted. Not only because they create confusion when using them for training, as the model will rate the entire recording rather than only the pronunciation of the target word, but also because of privacy reasons. Just because of this little additional text, we learned information about both gender and family relations for this child. For this work, genders are already provided in the speaker information, but this is not publicly available as of now. Other examples include d08_dør, where there is additional speech from the adult supervisor in which he instructs the child to pronunciate in normal fashion *"si det på vanlig måte"*. Other recordings include laughter, and many have poor quality, either with disturbing noise or just very low loudness. a34_bygge.wav is mistakenly annotated with a phoneme score of 4, so this was excluded in the phoneme analysis done in this work.

- a02_snoerr.wav
- a02_gloer.wav
- a22_ny.wav
- a29_doer.wav
- a34_bygge.wav
- a34_flue.wav
- a34_lue.wav
- d02_bleie.wav
- d02_koe.wav
- d02_kvern.wav

- d02_snoe.wav
- d05_prins.wav
- d08_bart.wav
- d08_bleie.wav
- d08_bror.wav
- d08_doer.wav
- d08_flagg.wav
- d08_kylling.wav
- d08_oere.wav
- d09_bjoern.wav
- d09_kran.wav
- d15_loeve.wav
- d19_soester.wav

### 4.4.2 Comparison of models

Models from the Scribe project and NbAiLab were compared based on global CER scores. These results, as well as results from Facebook models, which have not been fine-tuned on Norwegian, are presented in Table 4.4.1. The models were initially deployed on the full Teflon data directory of Norwegian speech, only excluding recordings that were annotated 0 by human assessors. Based on these preliminary results, the best models were redeployed on different subsets of the data. "Without Aalto excluded" means only the subset that was used by Aalto University during multitask training. "Without outliers" means that the speech sample outliers that were found during this work are excluded. Lastly, "with only 4-5 scores" excludes all recordings that got a reference rating lower than 4; this is because, in the multitask training, the ASR part is only trained on speech samples that got a score of 4 or higher.

**Table 4.4.1:** Global CER results, used to compare several models.

| | Scribe models | | | | Facebook models | | NbAiLab models | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Data | Combined long | Combined short | Radio | Stortinget | base 960 | large 960h lv60 self | 300m bokmaal v1 | 300m bokmaal v2 | 1b bokmaal v1 | 1b bokmaal v2 |
| CER without zeros | 46.50 | 45.83 | 48.81 | 55.36 | 79.91 | 72.63 | 57.68 | **43.78** | 62.36 | 46.97 |
| CER without Aalto excluded | | **45.54** | | | | | | 47.76 | | |
| CER without outliers | 46.05 | **37.19** | | | | | | 43.44 | | 46.64 |
| CER with only 4-5 score | 37.12 | **36.64** | | | | | | 36.96 | | 41.08 |

The combined short model from the Scribe project got the best CER for all subsets of the data, while NbAiLab 's 300m bokmaal v2 model got the best score for all data without zeros. A CER of 36.64% is still very high, but this is to be expected as these models have only been trained on adult Norwegian speech and also perform better on longer-form speech rather than single-word utterances such as the Teflon datasets.

### 4.4.3 Results of training with new base model

The combined short model from the Scribe project was used as a base model for multitask ASR and APA training as explained in Section 3.3. Table 4.4.2 shows the resulting evaluation of each of the three completed folds.

Fold 2 outperformed the existing multitask Aalto model across WER, CER, ACC, and UAR. Specifically, WER decreased by 1.51 percentage points, CER improved by 1.09 percentage points, ACC increased by 4.36 percentage points, and UAR increased by 1.22 percentage points. These results also outperform the existing results on the Swedish L2 dataset, while results for the Finnish L2 are still better. Differences in the number of target words and speaker age groups could also be contributing factors to this disparity between the languages.

Fold 3 improves UAR by 3.41 percentage points. Using the existing evaluation code from Aalto University, the F1 score, rather than MAE, is calculated for each fold, so this is not directly comparable. As Fold 1 does not perform well, the average evaluation does not give a good picture of the results; therefore, Fold 2 or the mean serves as a better representation.

These results indicate that the performance of the multitask ASR and APA system can be improved by adopting another Norwegian fine-tuned base model. The continued development of Norwegian models that are trained on larger amounts of data, with good representations of diverse pronunciation and speech patterns, will help further develop ASR and APA for Norwegian.

**Table 4.4.2:** Evaluation of multitask training after three completed folds.

| | Measures ASR performance | | Measures APA Performance | | |
|---|---|---|---|---|---|
| | WER [%] (↓) | CER [%] (↓) | ACC [%] (↑) | UAR [%] (↑) | F1 (↑) |
| Fold 1 | 25.73 | 11.63 | 49.49 | 41.83 | 0.39 |
| Fold 2 | **9.23** | **3.12** | **59.54** | 41.05 | 0.42 |
| Fold 3 | 10.09 | 3.61 | 54.15 | **43.24** | **0.43** |
| Average | 15.02 | 6.12 | 54.39 | 42.04 | 0.41 |
| Mean | 10.09 | 3.61 | 54.15 | 41.83 | 0.42 |

CHAPTER

# FIVE

# GENERAL DISCUSSION

## 5.1 Choice of model

Selecting the right model for automatic pronunciation assessment in the context of ASR systems is a complex task. The ASR model that is best suited for accurately converting speech to text may not be the most effective for identifying pronunciation errors. Models trained for ASR are usually designed to generalize across speech variations, such as dialects and accents from native and non-native speakers, which means they can accept pronunciation mistakes. While this flexibility is beneficial for general ASR use, it is counterproductive for APA tasks that aim to pinpoint and rectify these errors.

In APA , the focus is on identifying and diagnosing mispronunciations, which demands the capacity to detect subtle phonetic variations made by speakers. The system needs to balance between being fine-tuned enough in Norwegian to effectively learn and identify specific sounds, but not so generalized that it automatically rectifies mispronunciations, thereby producing apparently accurate text from poorly articulated speech.

## 5.2 Build publicly available dataset

Despite significant progress within speech recognition and deep learning methods, the development of CALL systems for children faces a major challenge due to the scarcity of datasets containing child speech. Even for children's ASR systems, which are generally well-developed, the amount of avail-

able data is challenging. As a result, there is a pressing need for child speech datasets that have detailed annotations required for APA systems. One of the reasons for the scarcity of such datasets is the practical challenges of recording child speech, particularly when it comes to capturing pre-defined one-word utterances. In the context of training APA systems for children, the recording environment needs to be consistent, with minimal background noise and high-quality recording devices, unlike the more flexible approach used for systems designed to recognize spontaneous conversational speech, where variability is welcomed. Making the recording sessions engaging and having simultaneous activities like drawing can be used to extend the children's attention span and make the process more fun.

In a language learning system, it is necessary to include not only non-native speakers but also native speakers, especially when dealing with children's speech. [19] For adult speech, one can assume a fully developed sound system of their native language, but this is not the case for children. Therefore, a child's difficulty in pronouncing foreign sounds might not be due to their inability to do so, but rather because their sound system is not yet fully developed. Consequently, the need for a diverse representation of age groups and multiple speakers for each group is important to accurately assess the performance of the system, in contrast to having the result be influenced by individual speaker proficiency.

The existing research on child speech APA systems shows that the majority of studies used private data, with only a few publicly available datasets [19]. This trend could be attributed to varying interpretations of privacy laws. For example, in the Teflon project, only the Norwegian data was made publicly available, despite the Swedish and Finnish datasets being recorded and handled exactly the same way in terms of anonymization and data privacy. The difference in data availability can be traced back to the authorities' interpretation of the law, where derivatives of the Swedish data were permitted for publication, while in Finland, the speech recordings themselves are considered personal information and could, therefore, not be published.

The debate on data privacy also raises questions about the age of the children involved and the possibility of obtaining informed consent. While comprehensive data privacy laws and guidelines are crucial for further technological development in our society, one might argue that the speech recordings of children pose a minimal risk of identity theft or speaker identification due to the significant changes in speech parameters as children grow older. Factors such as vocal tract length, associated with speaker height, directly affect

speech characteristics, and formant frequencies change over time. Therefore, most children participating in the Teflon dataset collection become unidentifiable just a few months later.

The ethical considerations surrounding the construction of a dataset also encompass the crucial element of obtaining consent for its intended use and ensuring that the guardians of minor subjects have confidence that the data will not be repurposed for other uses in the future. Verifying the credibility of data sources and upholding the protocols and permissions associated with them presents a growing challenge within the field of AI. There are concerns regarding the legitimacy of data sources used to train the large language models utilized by major corporations such as Facebook, with rumors suggesting that the data may have been obtained through web scraping rather than legal acquisition. This raises issues such as authors having their written work utilized for language model training without appropriate licenses, and public figures like Scarlett Johansson might have had their voice incorporated into speech-generative AI systems without their explicit consent, despite voicing their opposition to such use. [47, 48] These issues underscore the complexity and significance of ethical considerations in the development and utilization of AI technologies.

## 5.3 Evaluation

The task of annotating speech samples on a detailed level, especially for short or one-word utterances, is very difficult, even for human experts. In the Teflon project, some of the data was re-annotated by human assessors sometime after the initial assessments. The results showed that even though the human results outperformed the automatic model, the experts did not fully agree with themselves. [35] In other research, permissive accuracy, where mistakes between the top two levels (score 4 and 5) are discarded, has been used. This is because, by definition, they are extremely hard to be separated. [33] The difficulty is further emphasized when looking at how phoneme scores relate to global scores, where even where all phonemes are marked as correct, the highest score might not be achieved. This proposes a need for multiple annotations per utterance to get a more balanced assessment. For the Teflon project, 30% of the Norwegian dataset was annotated by two separate human assessors, but when using the dataset for multitask ASR and APA training, the two scores were averaged and rounded up, resulting in only one reference rating for each recording. By doing this, one

avoids the challenges of division into folds and datasets dependent on utterance or assessor, but it also limits the training and analysis possibilities. Preferably, the full dataset should have been annotated multiple times so that the full annotations beyond just global rating could have been used to better explain model performance. This would also provide more comprehensive training, with reduced possibility of assessor bias and increased consistency in automatic annotations.

The issue of using WER or CER is also discussed in the field, where other evaluation metrics have been presented. The problem is that only words or characters that are an exact match to the transcriptions are accepted as correct. So, even words that are semantically correct but phonetically varied will be penalized. For short-form speech, this recession might be needed, but in a longer form, the general meaning of the speech sample, where stuttering or pause phrases are omitted, is more convenient. As a result, CER can not be used alone to determine pronunciation proficiency, thus highlighting the need for specially trained models.

## 5.4 Choice of analysis method

This section discusses the chosen methodology for doing the broad scope of analysis undertaken in this study. Rather than only comparing reference ratings to predicted ratings, as done in [6]. This work focused on exploring the majority of the additional data available per speaker and recording, maximizing the use of information within the dataset. As the analysis has had a more broad approach, this limits how detailed one can explain each factor. The depth of the analysis was also been limited by the amount of available data. Even though there are many speech recordings, when separating them into several subgroups, the risk of speaker-dependent proficiency overshadowing higher-level group trends quickly arises. However, this analysis helps to better understand what affects both human and automatic pronunciation assessments and provides valuable insight into what speaker groups should be considered when possibly doing new rounds of data collection. For example, one could focus on broadening the number of recordings within each age group or for all first languages.

## 5.5 Future work

Building on the findings in this work, several avenues for future research are suggested to deepen the understanding and improve the effectiveness of APA systems. Mostly, the need for larger amounts of data is emphasized. By expanding the quantity of the recordings within age groups or across first languages, the possibility of verifying intra-group trends increases, and the risk of speaker-specific proficiency disturbing the model performance diminishes. However, as the collection of an extensive dataset is one of the main challenges itself, data augmentation could be explored for selected speaker groups, further improving the diversity and robustness of the dataset.

Additional speaker information, such as gender, what languages are used in the child's home, how long they have lived in Norway, target word length, or differences between stress-timed and syllable-timed languages, have not been investigated in this work. Further analysis exploring the effect of these factors could provide a broader understanding of children's language learning development and automatic prediction difficulties. Doing a similar analysis of the Swedish and Finnish data might also be insightful in understanding the similarities and differences in the data groups between languages and how they affect the performance of the APA model.

Furthermore, it would be natural to complete the multitask training with Combined Short from the Scribe project as the base model. In addition, using the new v2 300m bokmaal model from NbAiLab could further improve performance. It is also possible that using models that are trained on small subsets of all the intended users' first languages could improve APA performance. This would improve the model's ability to understand foreign sounds, but it would have to be robust enough to distinguish between multiple languages and only accept the Norwegian pronunciation as correct.

CHAPTER

# SIX

# CONCLUSION

The goal of this research was to gain a deeper understanding of the pronunciation difficulties of native and L2 child speakers of Norwegian, as well as the related APA system performance on this data. This thesis, therefore, conducted a comprehensive exploration of the factors influencing pronunciation difficulties among native and L2 child speakers of Norwegian and of the variables that affect prediction error in APA systems. In addition, the potential improvement of multitask ASR and APA system performance was tested through the adoption of a new base model fine-tuned on Norwegian data.

Addressing the first research question, this work recognized several fundamental elements that impact the pronunciation difficulties in child speakers. The analysis revealed that the developmental stage of the child, the speaker's first language background, highly affects pronunciation. The amount of noise in the recording, as well as the correctness of prosody, also affected human annotated scores. These elements manifest differently among native and L2 speakers, with L2 speakers facing additional challenges related to the interference of their first languages.

The second research question focused on what factors impact the prediction error of the APA system. The results of this suggest that prediction errors were increased by the biases in the data distribution and due to inadequate representation of specific linguistic groups in the dataset. The analysis highlighted that the APA system performed better on speakers whose first language shared phonetic sounds with Norwegian. As the model is only trained on Norwegian sounds, the amount of foreign sounds in speech samples might

lead to increased pronunciation errors. The APA model performs worse on recordings with a high amount of noise or that contain speech repetition, so one should strive to omit these types of speech recordings in the dataset or develop a more robust model that can handle these cases.

To answer the third and last research question, significant improvements were observed after training a new multitask ASR/APA model. The Norwegian fine-tuned Combined Short model was used as a base model for further training, utilizing the existing training setup from Aalto University. Three folds were completed, and the results from fold number two improved both WER, CER, ACC, and UAR from previous results. Specifically, WER decreased by 1.51 percentage points, CER improved by 1.09 percentage points, ACC increased by 4.36 percentage points, and UAR increased by 1.22 percentage points. This suggests that fine-tuning data that is linguistically related to the dataset substantially impacts model performance. These findings underscore the potential for further improvement in ASR and APA performance on this type of data.

As a result of all these findings, to further develop ASR and APA systems related to child L2 speech, it would be beneficial if future work included expansion of the dataset to more robustly represent the different age groups and first languages. It also ensures a good representation of utterances per target word, as these form the basis of the CALL gamification. These enhancements can enable more robust APA systems that will perform better for a wider range of child speakers.

In conclusion, while important advancements in understanding pronunciation difficulties and prediction errors for native and L2 child speakers of Norwegian have been made. The continued work in expanding datasets and further training of models is essential to improve the language learning process for kids in growing multilingual environments.

# REFERENCES


[1] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. *Low-resource Languages: A Review of Past Work and Future Challenges*. arXiv:2006.07264 [cs]. June 2020. DOI: 10.48550/arXiv.2006.07264. URL: http://arxiv.org/abs/2006.07264 (visited on 05/18/2024).

[2] *Teflon Project*. en-US. URL: https://teflon.aalto.fi/ (visited on 05/18/2024).

[3] Yaroslav Getman et al. "Developing an AI-Assisted Low-Resource Spoken Language Learning App for Children". In: *IEEE Access* 11 (2023), pp. 86025–86037. DOI: 10.1109/ACCESS.2023.3304274. URL: https://ieeexplore.ieee.org/document/10214276.

[4] Jacques Koreman. "Category Similarity in Multilingual Pronunciation Training". In: 2018, pp. 2578–2582. DOI: 10.21437/Interspeech.2018-1938. URL: https://www.isca-archive.org/interspeech_2018/koreman18_interspeech.html (visited on 05/23/2024).

[5] Silke Witt. "Automatic Error Detection in Pronunciation Training: Where we are and where we need to go". In: June 2012.

[6] Anne Marte Haug Olstad et al. "Collecting Linguistic Resources for Assessing Children's Pronunciation of Nordic Languages". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 3529–3537. URL: https://aclanthology.org/2024.lrec-main.313 (visited on 05/31/2024).

[7] Alexei Baevski et al. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. arXiv:2006.11477 [cs, eess]. Oct. 2020. DOI: 10.48550/arXiv.2006.11477. URL: http://arxiv.org/abs/2006.11477 (visited on 05/07/2024).

[8] Javier de la Rosa et al. *Boosting Norwegian Automatic Speech Recognition*. arXiv:2307.01672 [cs]. July 2023. DOI: 10.48550/arXiv.2307.01672. URL: http://arxiv.org/abs/2307.01672 (visited on 05/08/2024).

[9] *Wav2Vec 2.0 - a facebook Collection*. Jan. 2024. URL: https://huggingface.co/collections/facebook/wav2vec-20-651e865258e3dee2586c89f5 (visited on 05/26/2024).

[10] *scribe-project (SCRIBE)*. May 2023. URL: https://huggingface.co/scribe-project (visited on 05/26/2024).

[11] *NB-Wav2Vec - a NbAiLab Collection*. Mar. 2024. URL: https://huggingface.co/collections/NbAiLab/nb-wav2vec-651bd58f6a9194ff6b9b4fcd (visited on 05/26/2024).

[12] Julie Tvergrov. "Exploring phonetic annotations of TEFLON dataset as a preliminary step to end-to-end speech pronunciation assessment". Specialization Project. NTNU, 2023.

[13] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. eng. Voice/speech recognition technology. Upper Saddle River, NJ: Prentice Hall, 2001. ISBN: 978-0-13-022616-7.

[14] National Cancer Institute U. S. National Institutes of Health. *SEER Training Modules, Head & Neck Overview*. URL: https://training.seer.cancer.gov/head-neck/anatomy/overview.html (visited on 06/09/2024).

[15] *Parts of the Ear | NIDCD*. en. Mar. 2022. URL: https://www.nidcd.nih.gov/news/multimedia/medical-illustration-parts-ear (visited on 06/09/2024).

[16] M. Christian Brown and Joseph Santos-Sacchi. "Chapter 25 - Audition". In: *Fundamental Neuroscience (Fourth Edition)*. Ed. by Larry R. Squire et al. Fourth Edition. San Diego: Academic Press, 2013, pp. 553–576. ISBN: 978-0-12-385870-2. DOI: https://doi.org/10.1016/B978-0-12-385870-2.00025-1. URL: https://www.sciencedirect.com/science/article/pii/B9780123858702000251.

[17] Vassil Panayotov et al. "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. Apr. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964. URL: https://ieeexplore.ieee.org/document/7178964 (visited on 05/30/2024).

[18] Solveig Reppen Lunde. "Modeling the Interpretability of an End-to-End Automatic Speech Recognition System Adapted to Norwegian Speech". eng. Accepted: 2022-09-09T17:19:22Z. MA thesis. NTNU, 2022. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3016970 (visited on 05/07/2024).

[19] Yassine Kheir, Ahmed Ali, and Shammur Chowdhury. "Automatic Pronunciation Assessment - A Review". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8304–8324. DOI: 10.18653/v1/2023.findings-emnlp.557. URL: https://aclanthology.org/2023.findings-emnlp.557 (visited on 01/23/2024).

[20] *CC BY 4.0 Deed | Attribution 4.0 International | Creative Commons*. URL: https://creativecommons.org/licenses/by/4.0/ (visited on 05/30/2024).

[21] *NoW 10 Pronunciation - NTNU*. URL: https://www.ntnu.edu/now/10/pronunciation (visited on 05/23/2024).

[22] S. Davis and P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366. DOI: 10.1109/TASSP.1980.1163420.

[23] Arun Babu et al. *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. arXiv:2111.09296 [cs, eess]. Dec. 2021. DOI: 10.48550/arXiv.2111.09296. URL: http://arxiv.org/abs/2111.09296 (visited on 06/14/2024).

[24] Patrick von Platen. *Github patrickvonplaten/scientific_images*. Nov. 2021. URL: https://github.com/patrickvonplaten/scientific_images/blob/master/xls_r.png (visited on 05/08/2024).

[25] Alex Graves et al. "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks". In: *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*. Vol. 2006. Jan. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891.

[26] *CTCLoss — PyTorch 2.3 documentation*. URL: https://pytorch.org/docs/stable/generated/torch.nn.CTCLoss.html (visited on 05/30/2024).

[27] Daniel S. Park et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition". In: *Interspeech 2019*. arXiv:1904.08779 [cs, eess, stat]. Sept. 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680. URL: http://arxiv.org/abs/1904.08779 (visited on 05/30/2024).

[28] J. Garofolo et al. "TIMIT Acoustic-phonetic Continuous Speech Corpus". In: *Linguistic Data Consortium* (Nov. 1992).

[29] Zitha Sasindran et al. "HEVAL: A New Hybrid Evaluation Metric for Automatic Speech Recognition Tasks". In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Dec. 2023, pp. 1–7. DOI: 10.1109/ASRU57964.2023.10389717. URL: https://ieeexplore.ieee.org/document/10389717 (visited on 05/31/2024).

[30] Janine Rugayan, Torbjørn Svendsen, and Giampiero Salvi. "Semantically Meaningful Metrics for Norwegian ASR Systems". In: 2022, pp. 2283–2287. DOI: 10.21437/Interspeech.2022-817. URL: https://www.isca-archive.org/interspeech_2022/rugayan22_interspeech.html (visited on 05/31/2024).

[31] *Accuracy - a Hugging Face Space by evaluate-metric*. URL: https://huggingface.co/spaces/evaluate-metric/accuracy (visited on 05/08/2024).

[32] *Recall - a Hugging Face Space by evaluate-metric*. URL: https://huggingface.co/spaces/evaluate-metric/recall (visited on 05/08/2024).

[33] Yaroslav Getman et al. "wav2vec2-based Speech Rating System for Children with Speech Sound Disorder". In: 2022, pp. 3618–3622. DOI: 10.21437/Interspeech.2022-10103. URL: https://www.isca-archive.org/interspeech_2022/getman22_interspeech.html (visited on 05/31/2024).

[34] *MAE - a Hugging Face Space by evaluate-metric*. URL: https://huggingface.co/spaces/evaluate-metric/mae (visited on 05/08/2024).

[35] Yaroslav Getman et al. "Multi-task wav2vec2 Serving as a Pronunciation Training System for Children". en. In: *9th Workshop on Speech and Language Technology in Education (SLaTE)*. ISCA, Aug. 2023, pp. 36–40. DOI: 10.21437/SLaTE.2023-8. URL: https://www.isca-archive.org/slate_2023/getman23_slate.html (visited on 05/31/2024).

[36] *TeflonNorL2*. nb-NO. URL: https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-94/ (visited on 06/09/2024).

[37] Yaroslav Getman et al. "Github code related to Multi-task wav2vec2 Serving as a Pronunciation Training System for Children". In: *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*. 2023, pp. 36–40. DOI: 10.21437/SLaTE.2023-8. URL: https://github.com/aalto-speech/multitask-wav2vec2/tree/main.

[38] Per Erik Solberg et al. "Improving Generalization of Norwegian ASR with Limited Linguistic Resources". In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Ed. by Tanel Alumäe and Mark Fishel. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 508–517. URL: https://aclanthology.org/2023.nodalida-1.51 (visited on 05/31/2024).

[39] *huggingface/evaluate*. original-date: 2022-03-30T15:08:26Z. June 2024. URL: https://github.com/huggingface/evaluate (visited on 06/02/2024).

[40] Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, 2019. URL: https://www.python.org/.

[41] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007). Publisher: IEEE COMPUTER SOC, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[42] Michael Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (Apr. 2021), p. 3021. ISSN: 2475-9066. DOI: 10.21105/joss.03021. URL: https://joss.theoj.org/papers/10.21105/joss.03021 (visited on 06/02/2024).

[43] Per Erik Solberg and Pablo Ortiz. "The Norwegian Parliamentary Speech Corpus". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 1003–1008. URL: https://aclanthology.org/2022.lrec-1.106 (visited on 06/03/2024).

[44] Ingunn Amdal et al. "RUNDKAST: an Annotated Norwegian Broadcast News Speech Corpus." In: *Proceedings of the 6th International Language Resources and Evaluation (LREC 2008)*. Jan. 2008.

[45] Habibe Yanarkaya and Ayperi Sigirtmac. "THE CORRELATION BETWEEN DEVELOPMENT OF MOTHER TONGUE AND LEARNING SECOND LANGUAGE IN 48-72 MONTHS OLD CHILDREN". In: Jan. 2017.

[46] Teresa Cadierno et al. "Does younger mean better? Age of onset, learning rate and shortterm L2 proficiency in young Danish learners of English". en. In: *Vigo International Journal of Applied Linguistics* 17 (Jan. 2020). Number: 17, pp. 57–86. ISSN: 2660-504X. DOI: 10.35869/vial.v0i17.1465. URL: https://revistas.uvigo.es/index.php/vial/article/view/1465 (visited on 06/12/2024).

[47] Nicola Jones. "Who owns your voice? Scarlett Johansson OpenAI complaint raises questions". en. In: *Nature* (May 2024). Bandiera_abtest: a Cg_type: News Explainer Publisher: Nature Publishing Group Subject_term: Machine learning, Law, Computer science. DOI: 10.1038/d41586-024-01578-4. URL: https://www.nature.com/articles/d41586-024-01578-4 (visited on 06/13/2024).

[48] *Choban v. Meta | PDF | Legal Remedy | Class Action*. en. URL: https://www.scribd.com/document/670880272/Choban-v-meta (visited on 06/13/2024).

# APPENDICES

# A - SAMPA TO IPA PHONEME MAPPING

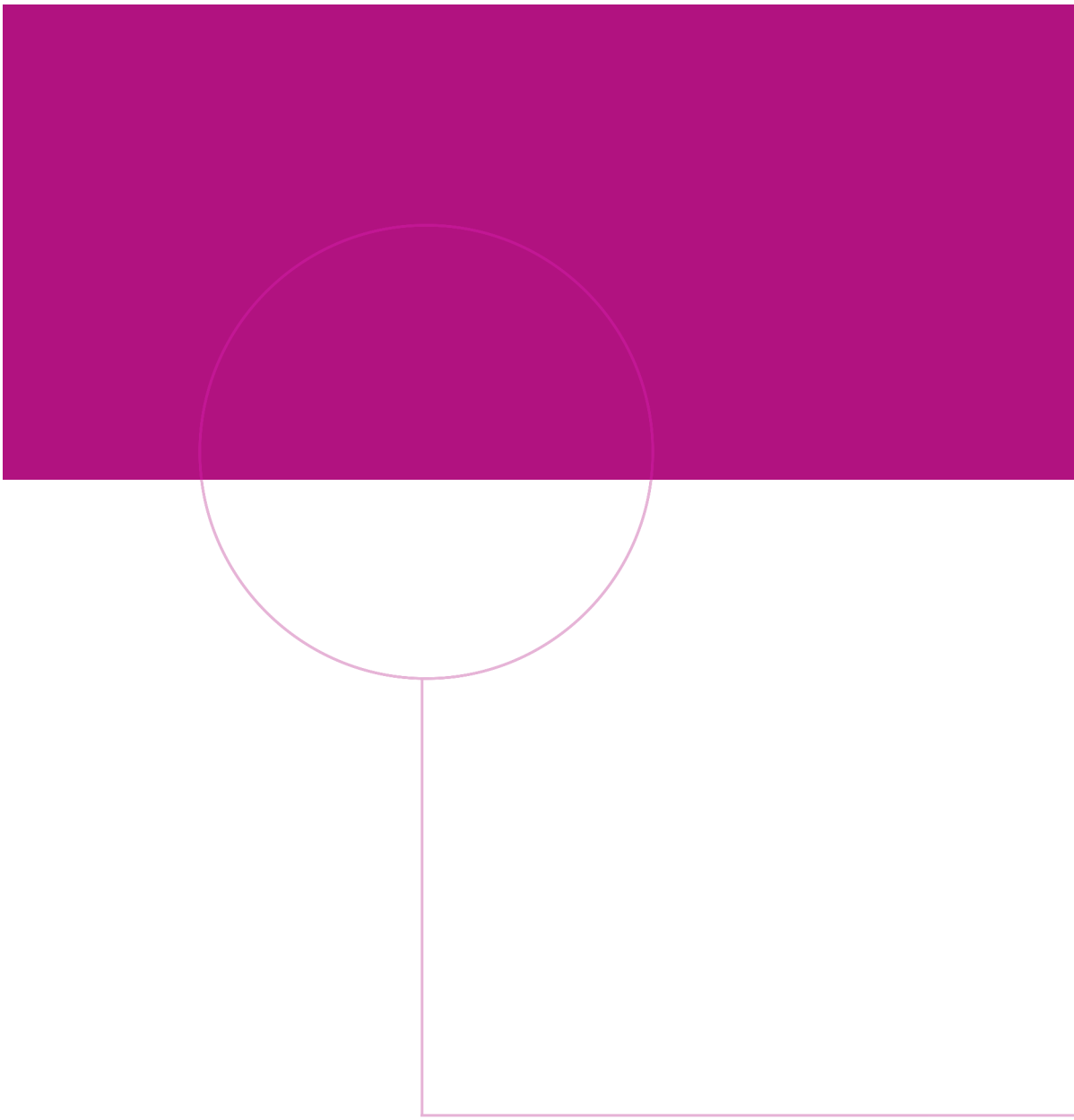
| Norwegian(SAMPA) | IPA Symbol |
|---|---|
| @ | /ə/ |
| A | /ɑ/ |
| A: | /ɑː/ |
| { | /æ/ |
| {: | /æː/ |
| b | /b/ |
| C | /ç/ |
| d | /d/ |
| E | /ɛ/ |
| e: | /eː/ |
| {I | /æɪ/ |
| f | /f/ |
| g | /g/ |
| h | /h/ |
| I | /ɪ/ |
| i: | /iː/ |
| j | /j/ |
| k | /k/ |
| l | /l/ |
| m | /m/ |
| N | /ŋ/ |
| n | /n/ |
| O | /ɔ/ |
| o: | /oː/ |
| 9 | /œ/ |
| 2: | /øː/ |
| 9Y | /œʏ/ |
| Eu0 | /æʉ/ |
| u0 | /ʉ/ |
| }: | /ʉː/ |
| p | /p/ |
| r | /r/ |
| n` | /ɳ/ |
| s` | /ʂ/ |
| t` | /ʈ/ |
| S | /ʃ/ |
| s | /s/ |
| t | /t/ |
| U | /ʊ/ |
| u: | /uː/ |
| v | /ʋ/ |
| n= | /nˌ/ |
| n`= | /ɳˌ/ |
| Y | /ʏ/ |
| y: | /yː/ |

**Figure .0.1:** Showing phoneme mapping from SAMPA to IPA symbols used in plots.

# B - OVERVIEW OF SPEAKER IDS DIVIDED INTO AGE GROUP AND SPEAKER TYPE

**Table .0.1:** Distribution of speaker age divided by native and non-native speaker IDs.

| **Childs age** | **Native** | **Non-Native** |
|---|---|---|
| 4 | d11 | |
| 5 | d08, d14, d16 | a07, a13 |
| 6 | d05, d06, d09, d12 | a02, a11 |
| 7 | d04, d17 | a01 |
| 8 | d01, d02, d03, d07, d15, d19 | a03, a04, a12 |
| 9 | d10, d13, d18 | a06, a08, a09 |
| 10 | | a05, a10, a14, a20, a34 |
| 11 | | a21, a22, a25, a27, a29, a30, a33 |
| 12 | | a15, a16, a17, a18, a19, a23, a26, a28, a31, a32 |



NTNU

Norwegian University of
Science and Technology