Doctoral theses at NTNU, 2023:53

Femke B. Gelderblom

Evaluating Performance Metrics for Deep Neural Network-based Speech Enhancement Systems

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

NTNU
Norwegian University of
Science and Technology

Femke B. Gelderblom

Evaluating Performance Metrics for Deep Neural Network-based Speech Enhancement Systems

Thesis for the Degree of Philosophiae Doctor

Trondheim, March 2023

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

NTNU

Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Electronic Systems

ISBN 978-82-326-5791-9 (printed ver.)
ISBN 978-82-326-6863-2 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:53

Printed by NTNU Grafisk senter

Abstract

A recurring challenge for speech enhancement (SE) systems, is that removing/reducing noise and reverberance does not necessarily increase the intelligibility or the quality of the speech for human listeners.

Deep neural networks (DNNs) are promising models for speech enhancement systems due to their highly adaptive non-linear nature. While these models can be trained with standard deep learning (DL) techniques to perform a wide variety of tasks, their real-life performance is dependent on the predictive power of the evaluation tools that guide the development process of speech enhancement systems.

This thesis focuses on evaluating the reliability of popular objective performance metrics of DNN-based speech enhancement systems. For this purpose, a variety of single channel and multichannel SE systems were developed and subjectively evaluated with listening tests.

None of the tested metrics proved to be reliable indicators for subjective changes in performance. This lack of reliable indicators critically impedes progress within the field of speech enhancement systems for human listeners.

List of papers

I. F. B. Gelderblom, T. V. Tronstad and E. M. Viggen, 'Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement,' in *INTERSPEECH*, Stockholm, Sweden: ISCA, Aug. 2017, pp. 1968–1972

II. F. B. Gelderblom, T. V. Tronstad and E. M. Viggen, 'Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 583–594, Mar. 2019

III. F. B. Gelderblom, Y. Liu, J. Kvam and T. A. Myrvoll, 'Synthetic Data For DNN-Based DOA Estimation of Indoor Speech,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada: IEEE, Jun. 2021, pp. 4390–4394

IV. F. B. Gelderblom and T. A. Myrvoll, 'Deep Complex Convolutional Recurrent Network for Multi-Channel Speech Enhancement and Dereverberation,' in *IEEE International Workshop on Machine Learning for Signal Processing*, Gold Coast, Australia: IEEE, Oct. 2021, pp. 1–6

V. F. B. Gelderblom, T. V. Tronstad, T. Svendsen and T. A. Myrvoll, 'On the Predictive Power of Objective Intelligibility Metrics for the Subjective Performance of Deep Complex Convolutional Recurrent Speech Enhancement Networks,' [Submitted]

To Jacob and Tobias,
for reminding me of the need for speech enhancement,
during every single call with their grandparents

Preface

Approximately ten years ago, during my job interview for a researcher position at the Acoustics group at SINTEF, I was asked whether I considered it an option to take a PhD. This took place approximately two minutes after I had been shown the corporate slides showing ambitious targets for a high percentage of employees with a doctorate degree. "Of course," I said "*if* the right project comes along, that is."

I did not expect that to really happen. And soon after, I was also way too busy being what Norwegians call 'a potato': a label that is apparently not meant as an insult, but a reference to versatility(?). As a research scientist at SINTEF's Acoustics group I worked with aircraft noise calculations, algorithms for automatic audiometry, community noise annoyance surveys, military noise calculation software development, a hearing assistive app prototype, and then 3D-audio. I focused more and more on programming and became increasingly interested in machine learning.

And then, the right project *did* come along. It started with last-minute midnight proposal writing, an invaluable Dilbert cartoon, and funding that disappeared literally two weeks after I had been admitted to the PhD programme. The project was disrupted twice due to the birth of my two wonderful boys. And then the Covid-19 pandemic came upon us with its lockdowns, quarantine regulations, shortened child-care opening times and travel restrictions. However, this very same pandemic also turned my PhD topic into something that friends, family and the media frequently discussed. Suddenly 'everyone' was asking for better microphones and complaining about participants that forget to unmute.

My 20 % position at SINTEF provided me with a constant stream of projects that I also wanted to work on, and far too often I felt that I never seemed to get a full week of work done anywhere, due to all of life's other distractions. However, I am also thoroughly convinced that the two main sources behind most of these distractions (who I love more than I'll ever be able to explain to them), also kept me from getting completely stuck on unimportant details during the project.

Sanity itself, however, I owe to my husband Wouter, for being with me through it all. This of course includes the nightly stomach-flue duty shifts and other not-so glorious moments of family life, but more importantly, the moments that tip the balance clearly to where it needs to be. Whether it is through forcing me away from my desk for short lunch-walks while sharing a home office or dragging the

whole family on camping trips into magical places like Femund: he always seems to know what I really need, especially when I'm too busy to see it myself.

I'm also thankful to the many other people that have helped me get here. In a very literal sense, this PhD would never have happened without Odd Kr. Pettersen, whose 'cat herding' skills are unsurpassed: at times, even *I* thought this whole thing was my idea. I want to thank my main supervisor, Tor Andre Myrvoll, who was always available to discuss statistics, signal processing and Unix-systems, and my second supervisor, Torbjørn Svendsen, for his critical questions and valuable feedback on all written work.

I'm also highly appreciative for the input from the other co-authors for the papers included in this thesis: Tron V. Tronstad (who gets bonus points for sitting through countless listening tests), Izzie Yi Liu, and Johannes Kvam. I want to thank Nancy Eik-Nes, who is better with a red pen than anyone else I know. And of course, I am thankful to all participants of the different listening tests. I know it is boring to listen to sentences like "Benjamin har 3 fine kasser"[1] over and over again. Really, *I know*.

Likewise, I'm grateful to my parents, who've brainwashed me from an early age into thinking that science is fun. And then there are my friends here in Norway and abroad, who I need to thank for being 'there', wherever that is, physically or digitally. A special mention goes to those who kept me here in Norway when we suddenly found ourselves completely isolated from all our friends, family and our safety net in the Netherlands. Inga, Karoline and Kjersti: you made all the difference.

Lastly, I want to dedicate this thesis to my deceased grandmother, 'oma Anneke', whose intellect and kindness have always been an inspiration. Suffering from severe hearing loss, she was the first to introduce me to the need for speech enhancement, even though it took a long time before I could understand the truth in her words:

> "It's not that I can't hear well enough anymore: it is just so noisy everywhere".

[1] "Benjamin has 3 nice boxes"

Contents

Figures

Tables

Chapter 1

Introduction

"You're on mute."

When the Covid-19 pandemic hit, and countries one by one went into lockdown, this phrase rapidly claimed its place in the English language (and many others). Almost overnight, the expression was recorded 10 times more frequently in business phone transcripts [6], and in 2021, British adults working from home voted it to be the most annoying overused expression from the pandemic [7]. But the phrase does not directly mention the virus or the concept of working from home. Instead, more than anything else, it is about the failure of speech enhancement (SE) systems.

In physical meetings, participants do not need to be muted. They can breathe, shift in their chairs, drink coffee from their coffee cups and take notes, without annoying or interrupting anyone. And while privacy is naturally more of an issue when meetings take place from (literally) anywhere, the sound of the TV playing in the background (to keep the kids quiet) really should not bother those on the other side of the call any more than it bothers the ones who are actually in the same room as the source.

Speech enhancement is about improving the intelligibility and quality of speech in a recorded signal. The online meeting/conference situation (including the extra challenging hybrid meetings where some participants are in the same room, whilst others are calling in), is just one of its telecommunication applications. Other applications are, for example, in telephony (mobile phones, call centers, etc.) and in radio communication (with cockpit to flight tower communication being an example of a situation with extreme noise levels). Speech enhancement is also important for hearing assistive devices and as a front-end for automatic speech recognition (ASR).

Speech enhancement is something that the human brain is great at, but that the field of speech processing has struggled with since the invention of the telephone, well over a hundred years ago. Especially, to improve the the intelligibility and quality of a recorded speech signal, so that a human listener consistently understands more of, *and* indicates preference over, the original noisy record, has been shown to be quite the challenge, again and again [8–11].

To be able to say anything about the performance of a newly developed speech enhancement system, the SE system needs to be evaluated. To evaluate speech enhancement systems, there are both 'subjective' and 'objective' evaluation methods. 'Subjective methods' provide a direct measure of human response with tests that involve human subjects. In contrast, 'objective' methods attempt to predict this human response by means of an algorithm, generally implemented in easy-to-use scripts. Subjective testing is time-consuming and expensive, but can provide a true measure of the SE system's performance. Objective evaluation is fast and does not require manual labour, which means that it can be repeated frequently during the SE system development process. As such, objective evaluation methods guide the development of SE systems, while subjective evaluation is used to evaluate final system performance. Therefore, the ability of objective metrics to correctly estimate changes in subjective SE system performance is crucial to the SE system development process and therefore also the final SE system performance.

This thesis focuses both on the development of modern speech enhancement systems and the predictive power of commonly used objective evaluation tools.

The rest of this introduction presents a short review of relevant background on the speech enhancement systems themselves and on the methods used for evaluating these systems.

1.1 Speech Enhancement

A speech enhancement system can either be single channel or multichannel, depending on the number of speech signal channels (which equals the number of microphones used to record the speech) it receives as input.

Single channel speech enhancement systems are more widely applicable, as only one microphone is required. However, microphone elements have become more affordable, and many devices (like smartphones, laptops and webcams) now come with microphone arrays: multiple microphones placed (slightly) apart. This allows for multichannel speech enhancement, where the small differences between the different recordings of the same speech signal can be utilised. As such, the single and multichannel speech enhancement branches are fundamentally different, even if the task for these systems is the same.

Traditional single channel speech enhancement

In the late seventies and early eighties, the growing popularity of 'speech communication systems' (including telephony, radio broadcasting and public-address systems) already provided a wide range of applications for speech enhancement systems. During this time, these systems relied on single microphone input, and for many of the applications, bandwidth compression was required.

The bandwidth compression systems were based on models of clean speech signals, and therefore significantly reduced the speech quality and intelligibility when the input was not clean, but instead degraded by additive noise (and reverberance).

This lead to considerable focus on the topic of single channel speech enhancement, which resulted in several methods that are still in use today, and/or from which many current methods are derived. Well-known examples of these 'traditional' speech enhancement methods are spectral subtraction [12], Wiener filtering [13, 14], and minimum mean-square error (MMSE) estimation [15]. These methods can reduce the amount of stationary background noise, but generally introduce residual noise (musical noise) and distortion, and are unsuitable for non-stationary noise types.

A case for artificial neural networks

Part of the difficulty of speech enhancement lies in the fact that speech is a complex signal that varies widely depending on what is being said, who is saying it, and in what conditions. At the same time, the noise signal also varies, and may be very different from the speech, or very close to it. This limits what 'physics' or 'model/rule-based' enhancement approaches (such as the traditional methods mentioned above) can achieve.

The human brain, on the other hand, is excellent at understanding speech in noise, and it has been shown that this is a skill that children learn over time: young adults in their early twenties can tolerate several dB more noise in speech than twelve-year-old children [16]. The ability to understand speech in noise is 'learned' by exposure to many different degraded speech signals, from which the noise coping strategies are then indirectly inferred. This inspires the strategy to 'train' speech enhancement systems in a similar manner: not by dictating rules, or outlining physics-based models, but by exposing the system to many different degraded signals while giving it the task to estimate the the underlying clean speech. This type of approach is called 'supervised (machine) learning'.

The speech enhancement problem can therefore be defined as a supervised machine learning problem, where the goal is to find the highly complex and non-linear mapping (the regression relationship) between the degraded and clean speech. While supervised learning methods are very general and can also be applied to simple forms of regression (such as linear and logarithmic regression), the complexity of the mapping between degraded and clean speech motivates the use of non-linear complex functions called 'artificial neural networks' (ANNs), or 'neural networks' (NNs) for short.

ANNs are made up of a network of nodes, where each node receives input from other nodes, and then sends its output to other nodes. Each of these connections is associated with adjustable weights. This, combined with a non-linearity at each node, ensures that the ANN becomes a very general function, which can behave in just about any way possible, where the exact behaviour depends on the values of all its weights.

Recently there has been a great deal of focus on ANNs in both scientific and non-scientific channels, as ANNs are currently the technological core of the majority of systems that solve common problems in the fields of 'artificial intelligence' (AI).

The general nature of ANNs makes them relevant for widely different applications, and well-known uses include image classification, face recognition, automatic speech recognition, chatbots, etc. etc. However, ANNs are not new, and ANNs were already used for single channel speech enhancement back in the eighties and nineties [17–19]. Early ANNs had fewer nodes than modern ANNs, due to hardware restrictions and a lack of the appropriate methodology required for the 'training' (the optimizing of the weights) of these more complex neural networks. Therefore, the early ANNs could not hope to estimate the complicated relationship between noisy speech and clean speech. So for decades, the focus was on other supervised learning techniques, like Gaussian mixture models (GMMs) [20], support vector machines (SVMs) [21], and non-negative matrix factorization (NMF) [22]

At the same time, neural network technology continued to improve. In 2012, the victory of the convolutional neural network called 'AlexNet' in the image recognition contest ImageNet initiated a cascade of deep-learning (DL) based technology development. Here 'deep' refers to the fact that the nodes of the neural network are organized in 'layers', where each layer (a collection of nodes) acts as a function nested in the subsequent layer. Deep learning is therefore a particular type of machine learning in which deep (multilayered) artificial neural networks, called 'deep neural networks' (DNNs), are trained for particular tasks.

In 2013, Wang *et al.* showed that combining DNNs with SVMs could outperform SVMs at the speech enhancement task [23]. This system attempted to estimate the ideal binary mask (IBM): a spectral mask with values that are either equal to one (the frequency band will not be blocked) or zero (the frequency band will be blocked). In the same year, Lu *et al.* published a paper where a deep autoencoder was used for speech enhancement, introducing the first deep architecture that attempted to find the direct mapping (without the use of a mask) between noisy and clean speech [24]. This was also the strategy of Xu *et al.* in 2016, where a DNN was trained to map the low power spectrum of noisy speech to clean speech [25].

These early deep learning based systems had promising results and motivated a new wave of interest in the topic. Recently, Microsoft has been organizing Deep Noise Suppression (DNS) challenges at the two major conferences for the speech processing community: INTERSPEECH and ICASSP [26–28]). While earlier mentioned systems generally were designed with both automatic speech recognition (ASR) and human listeners in mind, the DNS challenges specifically encourage research into improving *subjective* quality for human listeners. Among the contributions to these challenges, DNN-based approaches are the norm, and the same two 'branches' within deep learning-based speech enhancement (namely masking-based approaches vs direct mapping) are still equally relevant today [29, 30].

Multichannel speech enhancement

While the DNS challenges have only put focus on speech enhancement for human listeners in the past few years, regularly repeated CHiME (Computational Hearing in Multisource Environments) speech separation and recognition challenges have

put focus on noise robust automatic speech recognition for over a decade [31–36]. From the very beginning, datasets for these CHiME challenges have been multichannel, to allow for spatial filtering, in other words, the ability to separate sources (such as noise and speech) based on the fact that they are at different locations.

Model based methods for spatial filtering are collectively called 'beamforming' models. Beamforming is about boosting (any kind of) desired signal that is arriving from a specific direction, while attenuating the signals coming from other directions. How well it works depends on the microphone array configuration, the reverberation in the room, and how much the sources are separated in space.

Beamforming strategies can give improved speech intelligibility, even for the very small arrays on hearing aids (see for example [37, 38]). Beamforming has also proven to be valuable in ASR applications, especially when combined with DNN-based speech enhancement [39, 40]. While beamforming on its own can give impressive results, it only does so when the reverberation is limited, the direction of the speaker is known (and the noise comes from somewhere else), and/or when the noise signal itself is also known (which is especially difficult for transient noises).

As such, it is a natural next step to consider combining beamforming and deep neural networks for better supervised multichannel speech enhancement systems for human listeners. This step requires multichannel training datasets, which can be acquired by augmenting single channel datasets with simulated room impulse responses (RIRs), one for each microphone element in the array.

Both the CHiME and DNS Challenges have provided simulated and real data test sets. These allow for comparison of performance of systems on these sets. While there is a general trend that indicates that systems that perform better on simulated data also perform better on real data, there are clear performance gaps and outliers to the trend [33]. This motivates looking into the generation of more realistic multichannel data for the training of DNN-based speech enhancement systems.

1.2 Evaluation of Speech Enhancement Systems

To ensure that an SE system indeed enhances a signal, performance needs to be measured. Performance indicators also allow for comparison of different systems and can guide the development of new algorithms.

The signal-to-noise ratio (SNR) is possibly the performance metric with the longest history. But it is quite easy to raise the SNR, while simultaneously introducing distortion and degrading the speech for human listeners. Therefore it is important to consider the effect on human perception in evaluation methods. Developing such metrics is a field of research on its own, but the results are of crucial importance for the field of speech enhancement.

First of all, it is important to separate the concepts of intelligibility and quality. Intelligibility refers to the amount of spoken information the listener has actually

understood. Quality, on the other hand, is an opinion-based indicator: how annoying is the noise, how would you rank this clip with respect to this other clip? Both intelligibility and quality increase in lower noise conditions, but this does not mean that SE systems tend to achieve the same. Instead, already in 1979, Lim *et al.* observe that many SE algorithms can improve quality, but that almost all of these systems reduce intelligibility and that those that do not, tend to degrade the quality [8].

For both quality and intelligibility, it is the human perception of the signal that determines the performance. Both measures can therefore be measured subjectively with listening tests. However, listening tests are expensive, time consuming, and do not show where the performance difference comes from. As such, they are poor tools for guiding the development of speech enhancement systems, especially when relying on the highly iterative process of machine learning.

Therefore, objective measures of both quality and intelligibility have been developed. Popular examples are PESQ (perceptual evaluation of speech quality), POLQA (perceptual objective listening quality analysis), STOI (short-time objective intelligibility), ESTOI (extended STOI), HASPI (hearing-aid speech perception index), CSII (coherence speech intelligibility index) and NCM (normalized covariance metric). These metrics are algorithms that attempt to predict the response of human listeners with respect to the quality or intelligibility of speech. While practical and popular tools for system development, their usability stands or falls with their predictive power. Given the complex nature of human hearing and audio perception, it comes as no surprise that objective measures of speech intelligibility and quality often struggle to estimate human response [41–44].

Therefore it is important to validate the use of popular objective performance metrics for DNN-based speech enhancement, by comparing predicted performance with subjective results.

1.3 Topics of this Thesis

In summary, this thesis investigates different topics motivated by the presented background. For this purpose, multiple single channel and multichannel speech enhancement systems were implemented and evaluated both objectively and subjectively. Specific background for the topics studied, is presented in the papers themselves.

For this thesis, the following speech enhancement systems were evaluated:

1. **single channel**: A fully connected feed forward network that estimates the log magnitude spectra of clean speech [Paper I][Paper II]
2. **single channel**: Same as 1, but with an added global variance normalization postprocessing step [Paper I]
3. **single channel**: Same as 1, but where the target is to reduce the noise, instead of removing it [Paper II]

4. **single channel**: A convolutional recurrent network that estimates a complex ratio mask to be applied to the noisy input combined with a dereverberation block [Paper IV][Paper V]
5. **Multichannel**: The system used in 4 combined with a beamformer, where the direction of the beam was estimated from noisy input [Paper IV][Paper V]
6. **Multichannel**: Same as 5, but with oracle direction input for the beamformer [Paper IV][Paper V]

All of these speech enhancement systems were expected to enhance the speech, with respect to intelligibility and/or quality according to the predictions by objective metrics. Therefore, the investigations specifically focused on:

- Do any of these systems improve subjective intelligibility? [Paper I][Paper II][Paper V]
- Do either the first and/or third system improve subjective quality? [Paper II]
- Is STOI a reliable indicator for the change in speech recognition thresholds for speech processed by any of these systems? [Paper I][Paper II][Paper V]
- Is POLQA a reliable indicator for the change in mean opinion scores for speech processed by the first and/or third system? [Paper I][Paper II][Paper V]
- Are NCM, CSII, ESTOI and HASPI reliable indicators for the change in speech recognition thresholds for systems four to six? [Paper V]

To train the multichannel systems, multichannel data had to be acquired. Here, the effect of using different simulation methods (aiming to reduce the gap between real and simulated data) was investigated through the application of direction of arrival estimation (DOA):

- Is the final performance of a DOA estimation system affected by the room impulse response simulation method used to generate training data? [Paper III]

1.4 Mathematical Notation

Throughout Part I of this thesis, bold font is used to indicate any kind of multidimensional tensor (vectors, matrices and multidimensional arrays), while signals, functions, and scalars are in regular font. For signals, uppercase indicates that the signal is defined in the frequency domain, whereas lowercase is used for time domain signals. For tensors, the choice of uppercase vs lowercase is based on convention, and does not convey any information. For signals, square brackets are used for discrete arguments, while parentheses indicate continuous arguments.

1.5 Thesis Structure

This thesis is a compilation thesis and is divided into two parts. The first part presents the theory, relevant literature, methodology, and results in a combined manner. The second part, the papers and articles of this thesis, presents these topics as self-contained studies.

Part I contains seven more chapters in addition to this introduction. In the first three of these, the different building blocks of a speech enhancement pipeline are discussed. Chapter 2 goes into the details of degraded speech data, showing what needs to be 'removed' in order for speech to be enhanced. Chapter 3, explains the general deep learning techniques needed to understand how state-of-the-art speech enhancement systems work. Then, in Chapter 4, the question of how enhancement systems can be evaluated is discussed: How can performance be measured?

Building upon the tools outlined in these earlier chapters, Chapter 5 presents the single channel speech enhancement networks evaluated for this thesis. The multichannel speech enhancement systems are covered in Chapter 6. Here the theories behind microphone arrays, direction of arrival estimation, and beamforming are also discussed. Results of all tested systems are presented in Chapter 7.

Finally, in Chapter 8, general conclusions of the combined body of work are drawn, with additional focus on opportunities for further work.

Chapter 2

Degraded Speech Data

Recording 'clean' speech is practically impossible and there are two main reasons for this. Firstly, there is the issue of noise: Other sources of sound are almost always present. Secondly, there is the presence of reverberation: The speech signal itself is usually reflected into different directions by nearby objects (including walls, floors and ceilings), causing several non-exact copies of the same speech signal to arrive at the microphone at different moments in time.

The speech recorded in an anechoic chamber is the 'cleanest'. Anechoic rooms are built with absorbing surfaces to stop reflections of sound, and they are generally isolated to prevent sound from outside of the room to enter. There is no such thing as a perfectly sound-proofed anechoic room, and all microphones have at least some self-noise. Moreover, the vast majority of speech recordings are not obtained in highly specialized rooms with high-end equipment. Instead, meeting participants record their speech with whatever device they have available at the time, and from just about anywhere: their (home) office, a meeting room, the bus stop or a car. Most speech recordings therefore contain noisy reverberant speech that is more difficult to understand and less comfortable to listen to than clean speech. Not only that, the recordings are usually also worse than the real-life non-recorded sound would have been; that is, if the listener's ears had been located where the microphone was.

This effect is not because microphones are so much worse at picking up a signal than human ears. Instead it has to do with how the signal is presented. The human brain normally spatially filters everything it hears, automatically adapting the experience to even the most minor movements of the head. It does this by interpreting the tiny but crucial differences of the two signals received at the two ears. When in a (conference) call, the signal is usually presented as either a single channel (one ear) signal, or a stereo signal, with two equal signals at both ears. This prevents the brain from being able to spatially filter the speech, thus the subjective experience is noisier and more reverberant. Speech enhancement systems attempt to improve the intelligibility and quality of speech, and the most obvious route to this is by removing the noise and reverberance.

This chapter starts at the very beginning of the SE processing pipeline. First the degraded (to be enhanced) speech signal is defined and different presentations of this signal are presented. These different presentations are relevant, both to be able to visualize important concepts, and to be able to present the speech to trainable SE systems. Additionally, relevant methods for the simulation of degraded speech are explained.

2.1 Signal Representation

2.1.1 Time domain

Audible sound — including speech, the interfering noise and reverberance — consists of acoustical waves propagated by vibrations in air. Microphone elements pick up these sound waves and convert them into electronic signals. Electronic signals can either be analogue or digital.

The noisy reverberant speech signal recorded by an analogue microphone element can be called $x(t)$, with t for time. Then its digital equivalent, sampled at a fixed sampling frequency is $x[n]$, with n indicating the time step. This $x[n]$ depends on the additive noise signal at the microphone($v[n]$), and the clean speech signal produced at the source ($s[n]$), as follows:

$$x[n] = h[n] * s[n] + v[n], \tag{2.1}$$

where $h[n]$, is the complete transfer function from the speech source to the recording unit (including the impulse response (IR), other possible transmission artefacts, and the microphone response). For signals recorded indoors, it is common to refer to the impulse response as the room impulse response (RIR). The noise signal may come from one or more sources, and each of these will have their own reverberance, but all of these signal components are here collected in the definition of $n[n]$.

Both noise and reverberance degrade the intelligibility and quality of speech [45, 46]. Therefore, from a mathematical point of view, the goal of speech enhancement, is to recover the signal $s[n]$.

The amount of noise varies with time, and can be anything from barely audible, to completely dominating over the speech signal. Another important concept is therefore the signal-to-noise ratio (SNR). When only additive noise is taken into account, the SNR is:

$$\text{SNR}_{\text{additive noise}} = 10 \log_{10} \frac{\sum_{n=1}^{N} s[n]^2}{\sum_{n=1}^{N} v[n]^2}, \tag{2.2}$$

where N is the total number of samples over which the SNR is obtained. When including distortions and reverberation, the SNR becomes:

$$\text{SNR} = 10 \log_{10} \frac{\sum_{n=1}^{N} s[n]^2}{\sum_{n=1}^{N} q[n]^2}, \qquad \text{where} \tag{2.3}$$

$$q[n] = h[n] * s[n] + v[n] - s[n] = x[n] - s[n]. \tag{2.4}$$

Care needs to taken when comparing SNRs for speech. Given a noise source that is constant through time and an equal speech signal, the SNR will depend on the duration of the recording. Longer recordings will contain more segments where only noise is present, and these recordings will have a lower SNR, even if the level of the speech and noise sources are unchanged. However, the intelligibility and quality will not change accordingly, as the amount of noise present during the speech has not been altered.

Figure 2.1 shows example plots of clean speech ($s[n]$), reverberant speech ($h[n] * s[n]$), the noise signal ($v[n]$), and noisy reverberant speech ($x[n]$). Here the SNR of the noisy reverberant speech equals 5 dB.

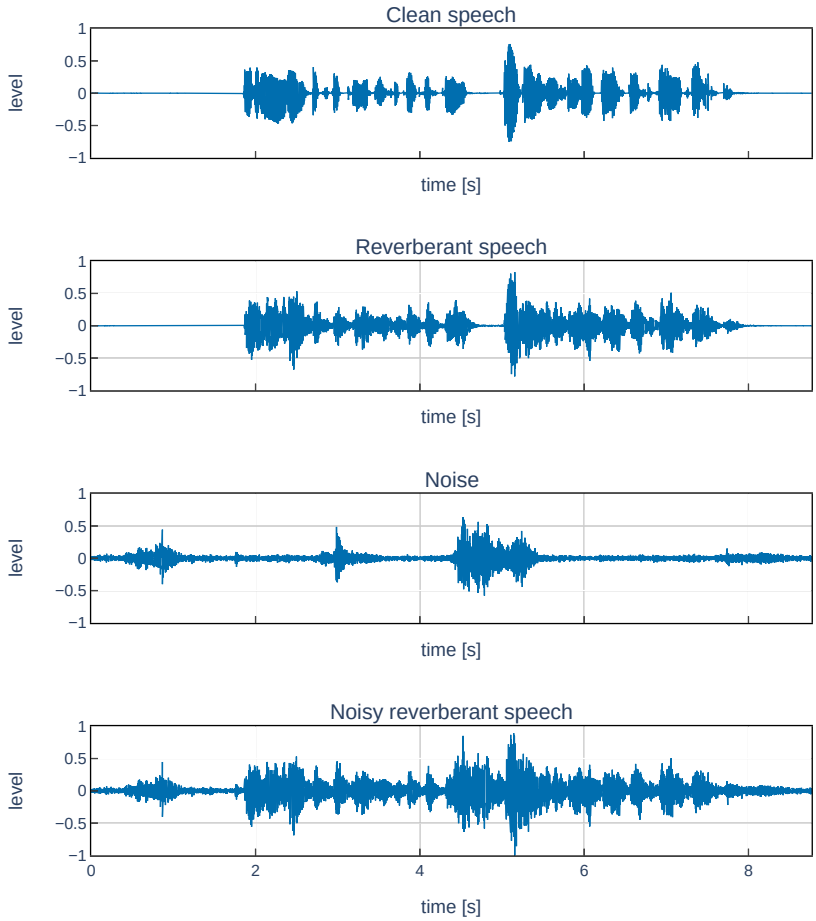


Figure 2.1: Time domain signals of the clean speech, reverberant speech, noise and noisy reverberant signals

All these signals are time domain signals. It is possible to send these signals to a loudspeaker to listen to them, but from a visualization perspective, the time

domain signal is not very informative. The plots show very little information about the type of noise in the recording, and the presence of reverberance is not/barely noticeable at all.

This is because all the different frequency components of audio signals are present in the time domain signal, but they are hard to discern as they are all added on top of each other at each time step.

2.1.2 Frequency domain

The frequency domain gives a more informative representation of the signal. The discrete Fourier transform (DFT), or its inverse, can be calculated with the fast Fourier transform (FFT). However, the speech signals also vary over time and it is critical that this information is not lost during the transformation.

Therefore, speech signals are often transformed with the short-time Fourier transform (STFT) algorithm. Here, the DFT is obtained (using the FFT) for overlapping windows of the signal. The windows are obtained by applying a sliding Hann (or similar) window to overlapping segments of the signal. The idea is that the windows are chosen short enough, so that the signal can be assumed to be stationary for the duration of the window.

The STFT can be used on all signals to obtain the frequency domain signals $X[m,k]$, $H[m,k]$, $S[m,k]$, and $V[m,k]$, where m refers to the STFT frame index, and k refers to the frequency bin.

$$X[m,k] = H[m,k]S[m,k] + V[m,k], \tag{2.5}$$

The STFT coefficients are complex, and therefore their real ($\Re$) and imaginary ($\Im$) parts can be presented separately, or as a combination of the magnitude and phase.

Magnitude and phase

The magnitude ($|X[m,k]|$) and phase ($\theta_{X[m,k]}$) of the noisy signal are:

$$|X[m,k]| = \sqrt{\Re\left(X[m,k]\right)^2 + \Im\left(X[m,k]\right)^2}, \qquad \text{and} \tag{2.6}$$

$$\theta_{X[m,k]} = \tan^{-1}\frac{\Im\left(X[m,k]\right)}{\Re\left(X[m,k]\right)}. \tag{2.7}$$

Figure 2.2 shows the magnitude spectra of the STFT of clean speech ($S[n,k]$). In these magnitude spectra, there is limited detail visible due to the large dynamic range of the magnitude coefficients. A more sensible presentation is therefore given in Figure 2.3, which shows the magnitude coefficients transformed with the log operator. The log-magnitude spectra show how the frequency content of signals varies over time.

However, to be able to reconstruct the time domain signal, the phase spectrum is also required. Here there are two options: to estimate the phase of the clean

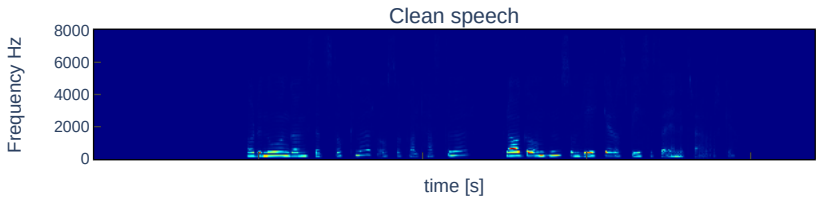


Figure 2.2: Magnitude spectrum of the STFT of the clean speech signal

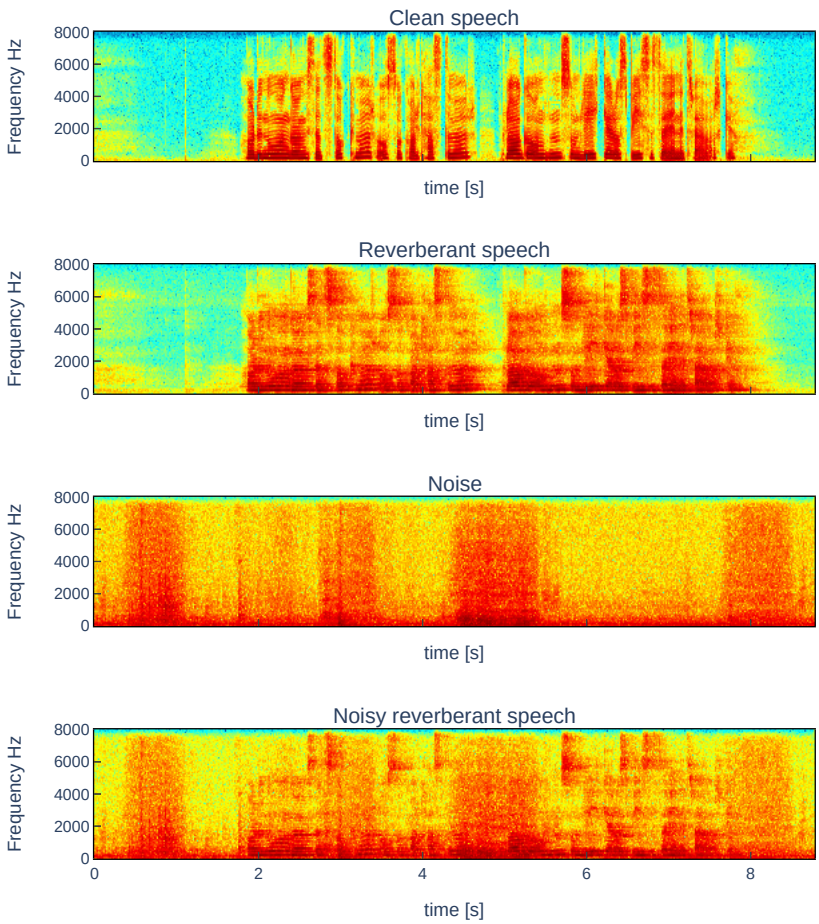


Figure 2.3: Log-magnitude spectra of the STFT of the clean speech, reverberant speech, noise and noisy reverberant signals

target speech, or to reuse the noisy phase directly, letting the enhancement system only work on the magnitude spectra.

Research from the early eighties advocated the 'unimportance' of estimating the phase. Wang *et al.*, for example, concluded that "an effort to more accurately estimate the phase from the noisy speech is unwarranted in the context of speech enhancement if the estimate is used to reconstruct a signal by combining it with an independently estimated magnitude" [47].

Therefore, based on the presumed lack of contribution by the phase, it has been common to enhance speech by combining the enhanced magnitude spectrum with the noisy (unaltered) phase. This was also the chosen method for the earlier papers in this thesis [Paper I][Paper II].

More recently, Paliwal *et al.* however presented clear evidence related to the importance of phase [48]. Their study showed that enhancement of the phase on its own (in other words, even without enhancing the magnitude) could already improve the subjective quality of the reconstructed signal. Additionally, they concluded that accurate phase spectrum estimates have the potential to significantly improve the performance of existing magnitude spectrum-based methods.

However, just because the retrieval of the clean phase is important, does not mean that it is easy. The 'shortcut' of reusing the noisy phase was additionally motivated from the machine learning perspective. Figure 2.4, shows the phase spectrum of the STFT coefficients of the clean speech signal. This spectrum contains very little structure — it appears rather random/noisy. This also means it will be difficult/impossible to learn anything from the phase spectra [49]. While there has been considerable effort put into clean phase retrieval, the task remains challenging [50].

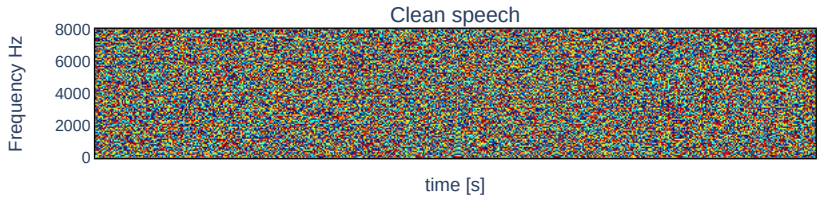


Figure 2.4: Phase spectrum of the STFT of the clean speech signal

Real and imaginary

Given that the phase actually is important, but too unstructured to be learned, other representations of the signal become more relevant. Both the phase and the magnitude spectra are obtained from the real and imaginary STFT coefficients. Therefore, improving the real and imaginary spectra will affect both phase and magnitude. Furthermore, Figures 2.5 and 2.6 show that these spectra do contain both temporal and spectral structure, making it possible to learn an informative mapping [49]. Therefore, the later work of this thesis utilised the real and imaginary spectra as input to the SE network [Paper IV][Paper V].

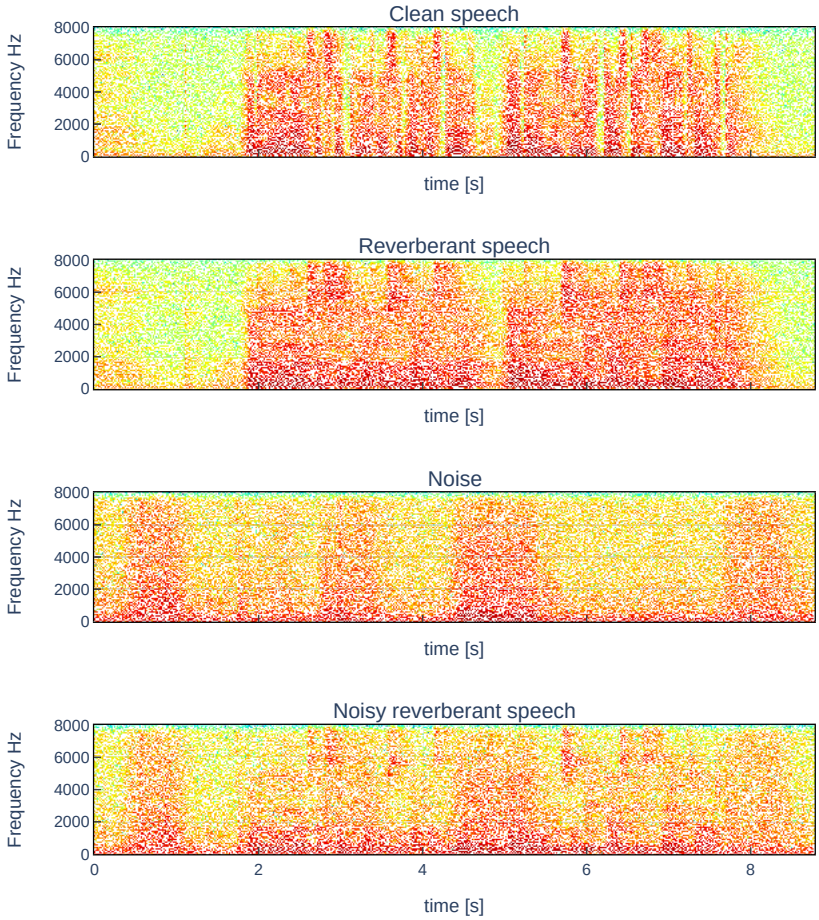


Figure 2.5: Log spectra of the real STFT coefficients of the clean, reverberant and noisy reverberant speech samples

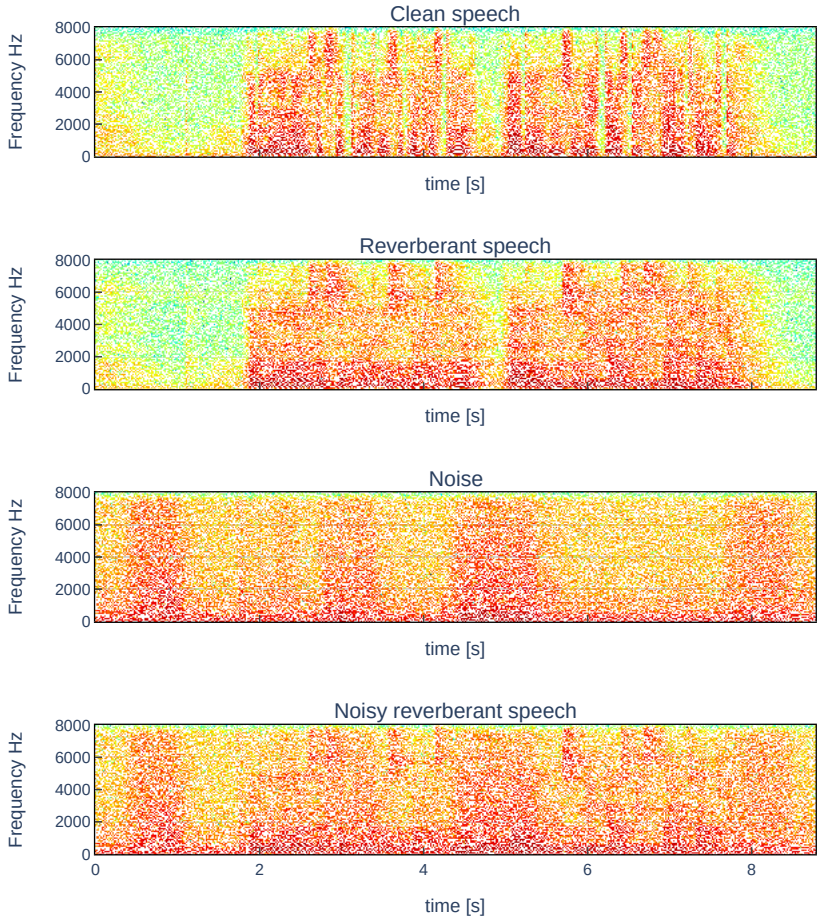


Figure 2.6: Log spectra of the imaginary STFT coefficients of the clean, reverberant and noisy reverberant speech samples

2.1.3 Multichannel

So far, this chapter has discussed the signals obtained by a single microphone. A microphone array is nothing more than a collection of multiple microphones (each located at a unique location). Therefore, the signal notation can be expanded to a multichannel problem using index i for each microphone element:

$$X[i,m,k] = H[i,m,k]S[m,k] + V[i,m,k], \qquad i = 1 \ldots N, \tag{2.8}$$

where N is the number of microphone elements in the array.

Here, the clean speech signal $S[m,k]$ does not depend on the microphone index (the speech source signal is the same, independent of where or how it is recorded). The RIR $H[i,m,k]$, however, is highly dependent on the relative positions of the speaker, the reflective surfaces, and the exact microphone element location. Also $V[i,m,k]$ depends on the microphone index, because here all different noise sources and their RIRs are combined into a single term. Adding the exact same noise to all channels would lead to a unrealistic situation where this noise can be used as a reference signal for direction of arrival estimation.

The microphone elements can be arranged in any possible configuration, but the multichannel SE systems from [Paper III][Paper IV][Paper V] are based on a circular microphone array. This particular array is shown in Figure 2.7. It is a prototype for a table top microphone, with a total of 37 microphone elements. The elements are configured in 4 concentric circles with diameters that increase from 4 cm to 16 cm in steps of 4 cm, plus a microphone placed in the center.

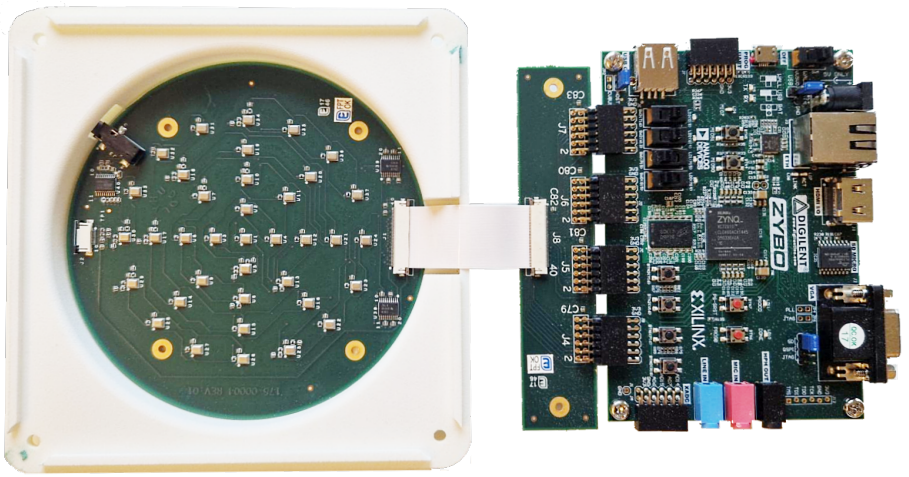


Figure 2.7: Photograph of the table top microphone array prototype. Here the backside of the array shows the 37 microphone elements.

While preliminary experiments were conducted with several different sets of microphone elements, all the published work on multichannel speech processing

is based on the circular configuration of nine of the microphone elements shown in Figure 2.8 [Paper V][Paper III][Paper IV].

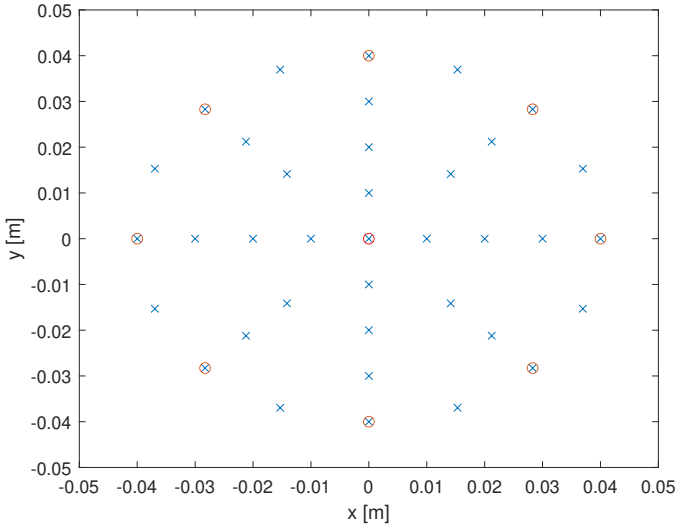


Figure 2.8: Configuration of the selected microphone array channels. Blue crosses indicate all available elements, red circles indicate the selected channels.

2.2 Clean Speech

As discussed in the introduction to this chapter, clean speech is hard to come by. It is possible to record speech in an anechoic chamber, but this is an expensive process that will limit the overall size of the database given time and cost restrictions. Therefore, for the purpose of trainable SE systems, it is common to rely on the next best thing: near-mouth microphone recordings in (relatively) quiet surroundings. Due to the short distance between the microphone and the signal source (the speaker's mouth), the direct signal should strongly dominate over all the reflections, and any background noise.

2.2.1 Speech databases

Many of the earlier speech databases used for speech enhancement were actually designed for automatic speech recognition (ASR). An example of such a database is TIMIT [51], which contains recordings of 630 speakers of eight major dialects/accents of American English, at a sampling frequency of 16 kHz. All speakers read ten phonetically rich sentences and the corpus includes time-aligned orthographic, phonetic and word transcriptions. Over the course of 2013 and 2014, a similar Norwegian database was collected: NB Tale [52]. NB Tale contains annotated 48

kHz recordings of 380 speakers, categorized by the main Norwegian dialects. The first papers included in this thesis relied on these databases [Paper I][Paper II].

The advantage of these databases is that they are controlled recordings, and all segments have passed through quality assurance. However, they are limited in size, and annotation data — while crucial for ASR — is not needed for speech enhancement.

Therefore, the speech data open-sourced for the 2021 INTERSPEECH Deep Noise Suppression Challenge [27] was used for the later papers in this thesis [Paper IV][Paper V]. The 'English read', the 'English emotional', and the 'Foreign language' speech subsets were included. Non-English languages included in this dataset are French, German, Italian, Mandarin, Russian and Spanish. Note that the Norwegian speech database was specifically not included in the training data, for easier comparison to other literature, and under the assumption that a system trained on large amounts of data from related languages like German and English should generalize well to Norwegian.

All subjective tests with listeners were based on 5-word Hagerman sentences speech material from Øygarden's speech-in-noise test [53]. These sentences were recorded in an audiometric room with a near-mouth microphone.

2.3 Additive Noise

People speaking in noisy environments often raise their voice, not only changing the SNR, but also the pitch of their voice. However, from a signal processing point of view, the noise signal is considered to be independent of the speech signal. This is shown in Equation 2.1 where the noise is simply additive.

For most applications of speech enhancement, there is little knowledge about the noise that may be present. There are countless sources of noise around us: it could be a dog barking in the background, a door being slammed shut, or a coffee machine grinding beans.

For the purpose of speech enhancement, it is therefore important to develop systems that generalize well to this wide variety of noise types. A general learning-based approach is therefore to expose trainable systems to all different sorts of noise, in the hope that it will eventually learn to find the speech signal, independent of the type of noise present.

2.3.1 Noise databases

The availability of audio databases containing suitable noise clips for training speech enhancement systems has increased a lot over the years. In 2015, Xu *et al.* presented their system for which they designed what was then considered a large training set, with many noise types [25]. Their noise database contained 104 different noise recordings of different noise types [54], and this set set was used for the earlier works of this thesis [Paper I][Paper II].

However, since then, Google has released a noise dataset called 'Audioset' with more than 2 million human labeled sound clips taken from YouTube videos [55]. In 2020, Microsoft provided a cleaned and more balanced subset of this database for their Deep Noise Suppression Challenges [26–28], containing 60 000 clips for 150 unique audio classes, plus an additional 10 000 clips from other sources. Here the noise types were specifically chosen based on their relevance for the VoIP (voice over internet protocol) application. This database is several orders of magnitudes larger than the one from Xu *et al.*. The later work of this thesis therefore relied on this much more extensive noise database [Paper IV][Paper V].

2.3.2 Recordings

For the purpose of evaluating multichannel speech enhancement systems, a set of noises was recorded with the microphone array prototype shown in Figure 2.7. The type of noises were chosen with the main application in mind: SE for the hybrid meeting setting. All noises are therefore typical 'meeting room' noises. Figure 2.9 shows random segments of the recorded noise types.

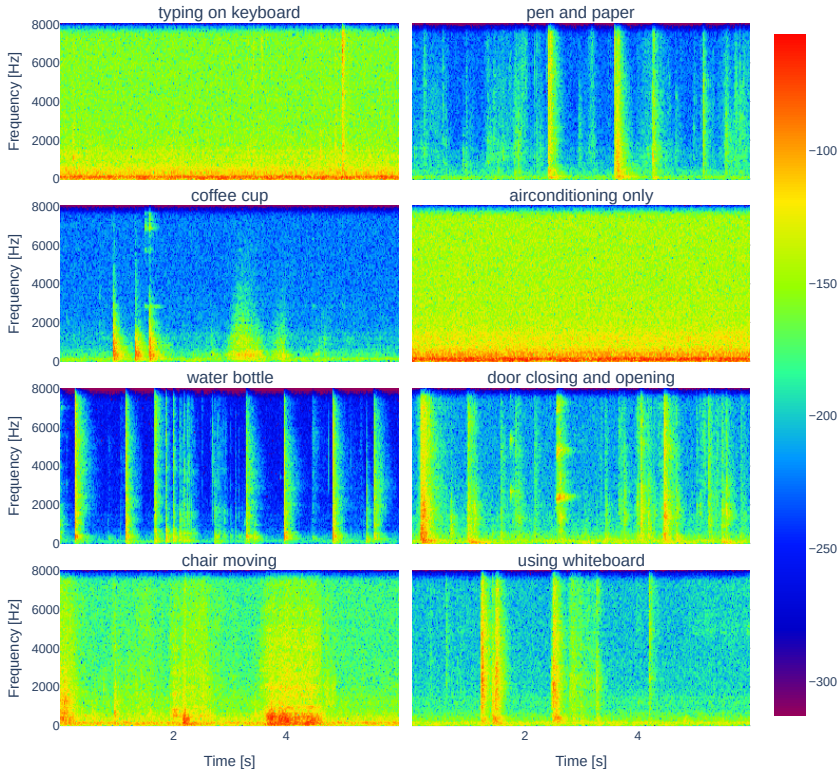


Figure 2.9: Random segments of the log-magnitude spectra of the recorded noise types. Only one channel shown.

The noises were recorded in a real meeting room at SINTEF, and contained therefore not only the chosen events, but also the general background noise in this room. This noise floor was dominated by the hum from the air-conditioning system, which is a stationary type of noise that only changes slowly over time. The other noises (like the clicking of a pen, or the movement of a chair) show more transient behaviour: they last only for a short time. This also means that SE systems should be able to deal with transient noises, where the long-term statistics of past segments have little relevance to the segment to be cleaned.

For the purpose of evaluating the single channel speech enhancement systems of [Paper I] and [Paper II], two noise clips were recorded: traffic noise and babble noise. These were recorded locally, at a crossroad in Trondheim, and in one of the university cafeterias of the NTNU, respectively.

2.4 Reverberance

Recorded speech signals are reverberant, because microphones also record all reflections of the direct speech signal signal that arrive at different moments in time. Reverberance is frequency dependent, as different frequencies are reflected and absorbed to different degrees.

The reverberance of the signal is captured by the impulse response h of Equation 2.1, and $H[m,k]$ of Equation 2.5. Each microphone element has its own RIR, which depends on the microphone's position relative to the speaker and all surfaces in the room.

The RIR can be separated into three main components: the direct path, the early reflections and the late reverberation tail. Figure 2.10 shows an example of a measured RIR.

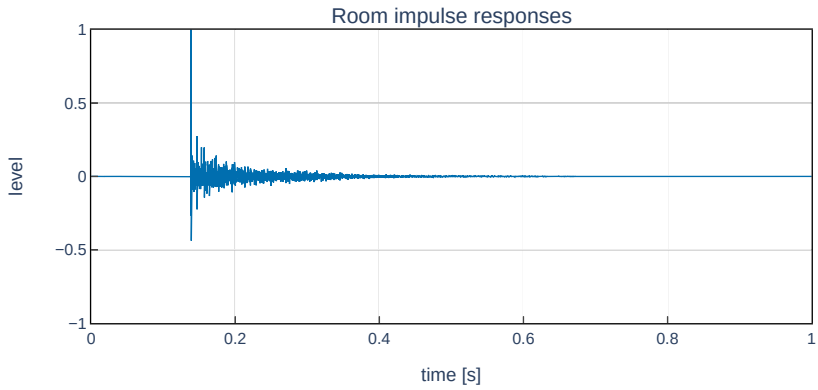


Figure 2.10: Example of a measured RIR

Often the direct path is the first and highest peak. However, this is not necessarily the case. For example, if the speaker is facing away from the microphone, the direct signal will be the damped signal coming out of the back of the head of the speaker. In this case, an early reflection from from a nearby wall or object in front of the speaker may be much louder.

The early reflections are relatively strong reflections of the signal that have 'bounced off' a limited number of surfaces before arriving at the microphone. The late reflections on the other hand have become 'diffuse'. They are caused by scattering effects due to small details on the surfaces in a room, and generate a sort of constant noise floor.

When anechoic clean speech is convolved with a measured RIR, the resulting signal should be exactly equal to a direct recording of the reverberant speech. However, recording RIRs takes time. The availability of RIR databases has increased in the recent years, but such databases are either for single channel microphones, or specific for the used microphone array.

Therefore, for the papers in this thesis, RIRs have been simulated and measured for the training and testing of the systems, respectively.

2.4.1 Simulations

The MCRoomSim package was used to simulate RIRs [56]. MCRoomSim operates in the frequency domain, where the phase of a source's directional response can also be simulated. MCRoomSim relies on the image source method (ISM) method to simulate the early reflections, and the diffuse rain algorithm for the diffuse reflections. Different combinations of these features are combined to generate different sets of RIRs.

Image Source Method

Allen *et al.* proposed the image source method (ISM) back in 1979 [57] and it has become the standard method for simulating RIRs for indoor environments. The idea behind this method is that sound hitting a wall, 'bounces' off this wall like light reflected by a mirror, with so-called 'specular' reflections.

With this in mind, the signal received at a microphone is obtained by adding the direct signal and all signals from 'image sources' located in 'mirror-rooms'. This concept is illustrated in Figure 2.11.

Figure 2.11 shows only one image source in a 2D situation with just one wall. In reality, each image will also have its own image sources, and there are four walls plus a ceiling and a floor. The number of image sources to include in the summation increases exponentially with the order of reflections.

Additionally, walls are not perfectly rigid: surfaces have finite impedance. This means that some of the incoming sound does not bounce off, but is instead absorbed or transmitted by the wall, where the degree to which this occurs is highly dependent on the angle of incidence. However, it is complicated to model the effects of surfaces with finite impedance [57]. Instead, the ISM assumes the

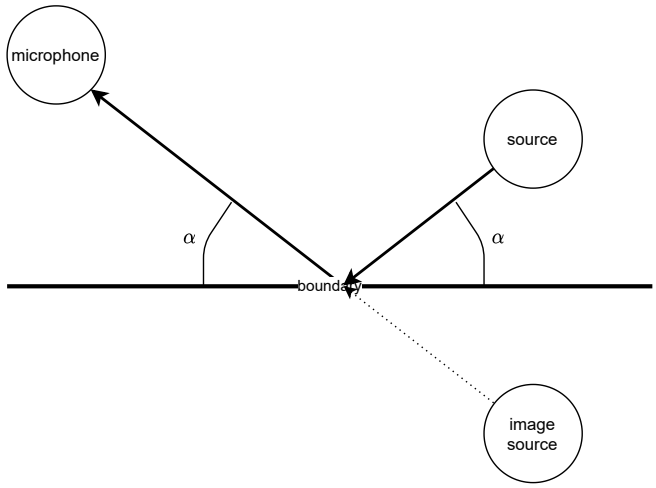


Figure 2.11: Illustration of the image source method

point image model even for nonrigid walls and defines the reflection coefficient for nonrigid walls to be finite and independent on angle.

Mathematically, the ISM defines the Fourier domain RIR (H), produced by a point source at position $X = (x, y, z)$, and received by a microphone element at position $X' = (x', y', z')$ as:

$$H(n, X, X') = \sum_{p=0}^{1} \sum_{r=-\infty}^{\infty} \beta \times \frac{\delta\left[n - \left(\left|R_p + R_r\right| / c\right)\right]}{4\pi \left|R_p + R_r\right|} \tag{2.9}$$

where $r = (u, v, w)$ and $p = (q, j, s)$ (both integer vector triplets), and

$$\beta = \beta_{x_1}^{|u-q|} \beta_{x_2}^{|u|} \beta_{y_1}^{|v-j|} \beta_{y_2}^{|v|} \beta_{z_1}^{|w-s|} \beta_{z_2}^{|w|} \tag{2.10}$$

$$R_p = (x - x' + 2qx', y - y' + 2jy', z - z' + 2sz') \tag{2.11}$$

$$R_r = 2(uL_x, vL_y, wL_z) \tag{2.12}$$

where L_x, L_y, and L_z are the dimensions of the room. The sums over the vector triplets p and r indicate three sums each; one for each of their components. These sums are therefore both over a three-dimensional lattice of points [57].

Diffuse Rain Algorithm

The diffuse sound field in a room is due to scattering effects from rough surfaces (see Figure 2.12).

The diffuse rain algorithm is a ray-tracing technique [56]. Like the ISM, ray tracing models the sound as travelling in 'rays', or beams of sound. These rays are (as with the ISM) partially reflected at different boundaries, and it is the energy of the rays that reach the location of the microphone that contribute to the RIR.

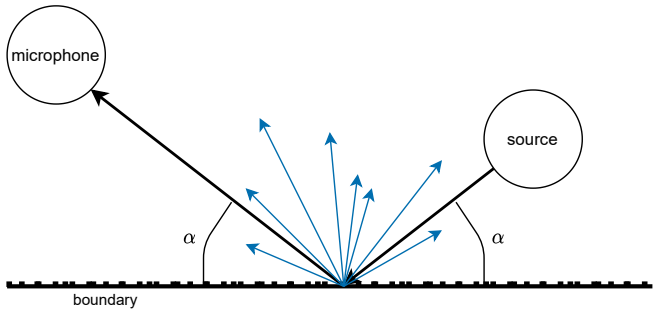


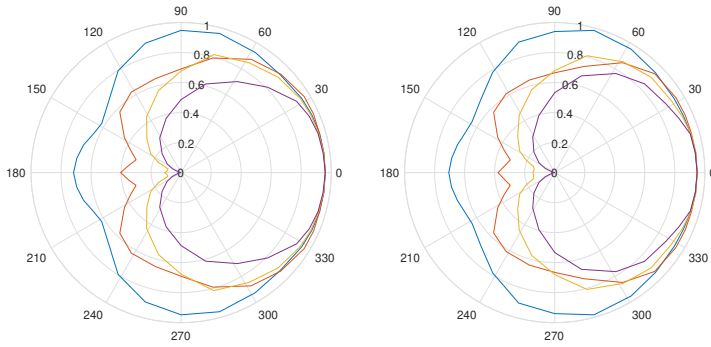
Figure 2.12: Scattering of sound on a rough surface

However, ray tracing is different in several ways. With ray tracing, the process is stochastic, and rays bounce off the reflective surfaces at random angles (instead of with specular reflections). Unlike with ISM (where only the exact rays that end up at the microphone are simulated), ray tracing is based on having a large number of rays, of which only a small proportion will reach the microphone and contribute to the simulated RIR. The accuracy of ray tracing increases with the number of simulated beams.

Ray tracing is used to simulate the scattering effect that occurs because surfaces are not perfectly smooth. The result is the diffuse field, which makes up all the late reflections in the RIR's tail.

Directional speakers

Figure 2.13 shows average directivity patterns for male and female speakers. A speaking person is not an omnidirectional source that emits an equal amount of sound in all directions. Instead, the speech signal is clearest when the speaker is



(a) Male speaker directivity **(b)** Female speaker directivity

Figure 2.13: Male and female directivity patterns. Used directivity patterns are defined in 3D and 256 frequency bands, but only zero elevation for a few selected frequencies (— 1 kHz, — 2 kHz, — 4 kHz, and — 8 kHz) ares shown here.

looking straight at the listener/receiver, and both attenuated and distorted when coming from behind. Note that the patterns depend on frequency.

When speakers are directional, it suddenly becomes relevant whether they are speaking towards the array or not. For an omnidirectional source, there will always be a strong direct path, but this is no longer the case for a speaker who is instead facing the opposite direction.

Simulated RIR sets

To ensure the task given to the learnable system is representative, the SE system input needs to represent the reverberant reality. This is true even if the goal is to just remove the noise, without any form of dereverberation. When exposed to reverberant input, trainable systems may also be able learn to extract useful information from the reverberance, and/or learn to remove it.

Additionally, it is important that the simulated reverberance then represents reality as closely as possible. Real recordings of speech are reverberant, due to both the specular reflections and scattering effects, and real speakers are directional sources.

To investigate which of these properties matter, four different *datasets* of RIRs were simulated, using four different simulation methods. First, 6000 (2000) virtual training (validation) rooms were modelled. Each room was randomly configured with parameters drawn from the uniform distributions specified in Table 2.1. Each room contained three speakers, and three noise sources, all placed randomly in the room within the restrictions given in the table.

Table 2.1: Details of the random virtual room configuration. Table adapted from [Paper III] (©2021 IEEE).

Item	Parameter	Min.	Max.
Room size	width	3 m	8 m
	length	3 m	10 m
	height	2.5 m	6 m
	RT60	0.2 s	1 s
	scattering coefficient	0	1
Array position	from walls	1 m	-
	from floor	0.6 m	0.9 m
Speaker positions (3x)	from walls	0.5 m	-
	from floor	1 m	1.8 m
	from array	0.5 m	-
	yaw (directive speakers only)	-180°	180°
Noise positions (3x)	from walls	0 m	-
	from floor	0 m	-
	from array	0.5 m	-

Positions were drawn such that they were evenly distributed in all directions. The average absorption of a room was determined from the random drawn RT60 time with Eyring's [58] algorithm with air absorption taken into account. Here RT60 stands for the time it takes for the sound to decay by 60 dB: the reverberation time. The RT60 depends on frequency, as different frequencies lose their energy at different rates.

Then for each source (three speakers plus three noise sources), four RIRs with the following methods were simulated:[1]

- **ISM-omni**: the basic RIR generated by ISM where sources are modelled as omnidirectional. No scattering and no diffuse field.
- **ISM-dir**: Like ISM-omni, but now sources are modelled as directive speakers, with either an average male or female directivity. No scattering and no diffuse field.
- **WithDiffuse-omni**: An advanced RIR with not just specular reflections, but also a diffuse field due to scattering, where sources are modelled as omnidirectional.
- **WithDiffuse-dir**: Like WithDiffuse-omni, but sources are again modelled as directive speakers.

2.4.2 Measurements

To be able to determine the effect of using different methods for generating training data, real data is needed to test the trained neural networks. For this purpose, RIRs were measured in the same meeting room as there where the background noise was recorded.

All measurements were obtained in the same room, which was one of SINTEF's meeting rooms. This particular room has dimensions 4.5 m x 3.8 m x 2.6 m, and $RT60_{1kHz}$ of 0.3 s. It is a rather typical medium-sized room, with two glass walls, two wallpapered hardboard walls, carpet on the floor, and sound-absorbing ceiling tiles. Centered in the room, there is a large oval table surrounded by eight chairs. The layout of the room where measurements were conducted is shown in Figure 2.14.

The exponential sine sweep (ESS) method proposed in [59] was used to measure the RIRs. A sine signal with exponentially varied frequency is played over a NTi TalkBox [60], placed at various locations in the room. This loudspeaker has head-size dimensions and is specifically designed for human speech measurement. The signal was recorded by a microphone array that was positioned approximately in the middle of the meeting room's table. The loudspeaker was aimed either towards the array, or rotated at a 90° angle. To obtain the room impulse response, the recorded signal was then convolved with the inverse of the original sine sweep.

The prototype microphone array (See Figure 2.7) interfaces through a LAN network communication port. As such, recordings had to be synced in time with the audio playback. For this purpose, a single earbud was mounted directly adjacent

[1] List of methods copied from [Paper III] (©2021 IEEE)

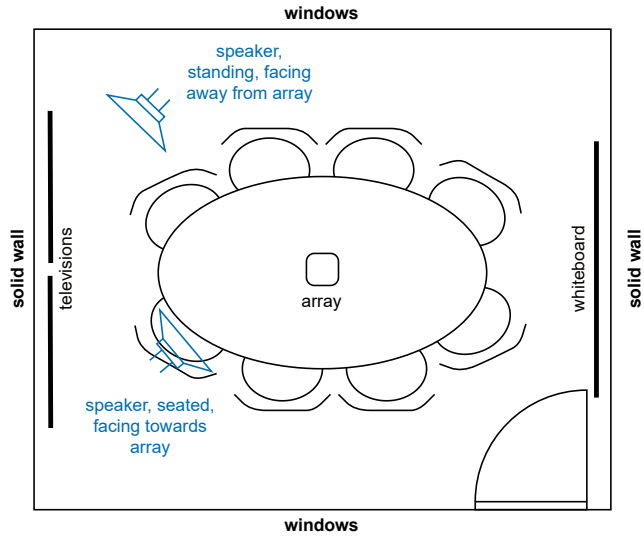


Figure 2.14: Room layout for RIR measurements

to the center microphone of the microphone array. A maximum length sequence (MLS) pulse was then played at a known delay before the sine sweep, allowing for time-synchronization of the final RIR.

A total of 57 RIRs were obtained for a speaker facing towards the array, and 107 with the speaker rotated at a 90° angle. The true DOAs were measured with an uncertainty of ± 1°, from positions approximately evenly spread out over the room, within the restraints of the furniture present. The overall aim was to collect RIRs at 'natural' speaker positions - both standing and seated, facing towards potential meeting participants, or towards the whiteboard or TV-screens. The distance to the microphone varied from 1 to 2 m.

From the RIR measurements, two test sets (to be convolved with speech) were obtained, dubbed 'Easy' and 'Challenging'. The first of these contained all RIRs where the speaker was facing towards the array, the second all RIRs where the speaker was facing away.

Chapter 3

Deep Learning

This chapter describes the general deep learning technology used in the speech enhancement systems based on deep neural networks (DNNs) and the direction of arrival (DOA) estimation system that were developed for this thesis. Supervised regression machine learning is presented, followed by a discussion of the types of neural nets and layers relevant to this work. The purpose of this chapter is to introduce the concepts needed to understand later chapters.

Complete systems often contain preprocessing and postprocessing steps in addition to the DNN core. In this chapter, the focus is only on the DNN model. To stress this, the following general variables will be used: $\mathbf{y}_0$ as the model input (meaning the input to the first layer), $\hat{\mathbf{y}}$ for the model output, and $\mathbf{y}$ for the model's target (the desired output). Note that while $\mathbf{y}$ is used for all these, they do not have to be similar in any way. For example, the input to a model can be the STFT coefficients of a speech signal, while its output may just be a scalar: the direction from which the speech is coming.

In the next section, the model is first treated as a black box that applies some weight dependent function to its input, to obtain an output. This, because the training framework does not change with the type of layers used in the model. Then, in Section 3.2, the different layer types used in the DNN-based systems included in this thesis are discussed. The details of the exact systems used in the work included in this thesis will be explained in Chapter 5 and 6.

3.1 Supervised Regression Learning

A computer program is said to *learn* if its performance of a task, improves with experience [61]. Any type of machine learning that tries to learn from paired input/target data is called 'supervised'. Regression is the term used for the general technique to estimate the relationship between two variables.

It is therefore quite natural to define the speech enhancement problem as a supervised regression learning problem. For this thesis, the main goal is to learn the relationship between pairs of clean and degraded speech. However, supervised

regression is equally suitable for the work on direction of arrival estimation, where the goal instead is to find the mapping between input speech and the azimuth of where the speaker is located.

3.1.1 A trainable system

For a DNN model to learn something, it needs to be 'trained'. This means that the model is exposed to input samples ($\mathbf{y}_0$), from which it estimates a target $\mathbf{y}$. The aim is to obtain information from each exposure and to use it to improve the model. The model is defined as the function $f(\mathbf{y}_0, \mathbf{w})$, where $\mathbf{w}$ are the tunable weights[1]. This gives:

$$\hat{\mathbf{y}} = f(\mathbf{y}_0, \mathbf{w}), \tag{3.1}$$

where $\hat{\mathbf{y}}$ is the model's estimate of $\mathbf{y}$.

Applying the model $f(\mathbf{y}_0, \mathbf{w})$ to the input is called the forward pass. At first the weights are initialized to random values[2], making the first estimate of $\hat{\mathbf{y}}$ also rather random. However, once the first estimate is obtained, the error between this estimate and the target can be obtained with a loss function. Based on the output of the loss function, the weights of the network are updated, with the aim to make it a better estimator for the samples to come. This part where the weights are updated is called the 'backward pass', and it only takes place during training.

Therefore, during training, the loss function plays a key role. Generally speaking, the loss function can be defined as $L(\mathbf{y}_0, \mathbf{y}, \mathbf{w})$, which depends on the noisy inputs, the clean targets and the weights of the model.

A very commonly used loss function is the mean squared error (MSE) function:

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^2 \tag{3.2}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(f(\mathbf{y}_o)_i - \mathbf{y}_i\right)^2, \tag{3.3}$$

where m is the total number of samples over which the loss function is calculated, and i the index of the sample pair. The MSE loss is very versatile, because its general underlying concepts are relevant for many applications. First of all it reduces to zero when the estimate and target are equal, and an error in one direction (i.e. $\hat{\mathbf{y}}_i > \mathbf{y}$)) is punished exactly like an equal deviation in the opposite direction ($\hat{\mathbf{y}}_i < \mathbf{y}$)). Furthermore, the squaring operation ensures that large deviations are punished relatively more than small ones, which encourages convergence and is sensible for many applications. The MSE loss was used for the work on earlier SE systems and DOA estimation in [Paper I][Paper II][Paper III], but there are also many other options.

[1] Here lowercase $\mathbf{w}$ indicates both the weights and biases, which will later separate into the variables $\mathbf{W}$ and $\mathbf{B}$

[2] This a simplification. There are for example, also initialization approaches that rely on (unsupervised) pre-training or heuristics.

The same loss function has a very different effect for different presentations of the output of the network and the target. Using the MSE loss with a time domain target and a time domain output will have a very different outcome than when using the same MSE loss with a spectral target and and network output that is a spectral mask of the noisy signal, for example.

All loss functions give a loss landscape, in the highly dimensional space of all the weights of the model. Somewhere, this (usually extremely) complicated landscape has minima, and this is where the estimates are closest to the targets. Learning/training means searching for a minimum: optimizing/minimizing the loss function

In the work of this thesis, Adam (adaptive moment estimation) [62] has been used for all final systems. This optimization method computes individual adaptive learning rates for the different parameters from estimates of the first moment (the mean) and second moment (the uncentered variance) of the gradients. While some sources have shown that Adam, despite its popularity, is not a one-size-fits-all solution for all problems [63], our experiments consistently showed it worked well for the systems developed for this thesis.

3.1.2 Validation

During training, a model is fitted to the training data. Validation is then used, to test how well the model works for unseen data. However, validation is not the same as testing the final performance of the model, which is described in detail in Section 4. Instead, validation is the phase where the model is checked for overfitting, and where different hyperparameters are tried, so as to select the model that performs best, for further testing.

During validation the backward pass is not applied, even if the loss function is calculated. In other words: the weights are not updated.

When relying on DNNs, which by their very nature can provide highly complicated mappings, the risk of overfitting is substantial. Overfitting means that the mapping is too specific to the training data, and does not generalize well to unseen data.

To check that the model is not overfitting, it is important to test it on unseen data. For the models of the earlier papers in this thesis, where the number of hours of speech was limited, it was assured that the validation set was a gender/dialect balanced subset of the whole training set, where speakers were never part of both sets. Also, different unseen noise types were selected for validation. For later models, the subsets suggested by the DNS database [27] were used.

3.2 Deep Neural Network Types

Two types of DNNs were used during the work for this thesis. For the earlier speech enhancement systems of [Paper I] and [Paper II] and the direction of arrival system of [Paper III], the networks were so-called fully connected feed forward

networks (also called fully connected neural networks, or multilayer perceptrons (MLPs)). Later work was based on a network with both convolutional and recurrent layers. In the following sections, the basics of these layers and the encoder-decoder architecture are explained. A far more extensive introduction can, for example, be found in [61].

3.2.1 Fully connected layers

Figure 3.1 shows a typical visualization of a deep neural net, with fully connected feed forward layers. The name feed forward comes from the fact that all information flows forward in this network (from left to right), and fully connected from the fact that all nodes are connected to all nodes in the layer before and after, as represented by the many arrows in Figure 3.1. These models are called networks, because they combine different functions, and they are called deep, because of the multiple layers.

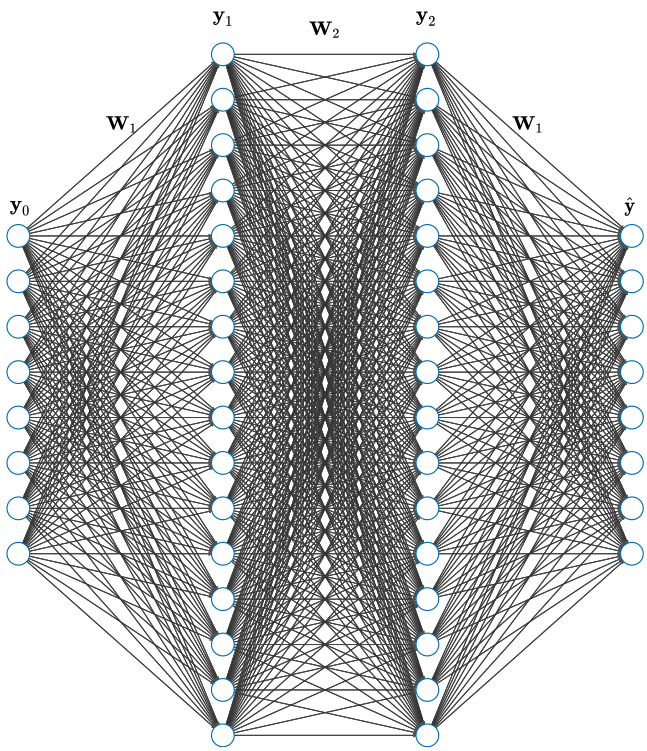


Figure 3.1: Deep neural net with fully connected layers. The circles represent the elements of each vector, the arrow the weights. Here the biases are not drawn, as is common practice.

The network (or model) is essentially a mapping function with learnable parameters. These learnable parameters are generally called weights, but this

concept can be divided into the weights $\mathbf{W}$ and biases $\mathbf{B}$. Both of these parameters are learned during optimization.

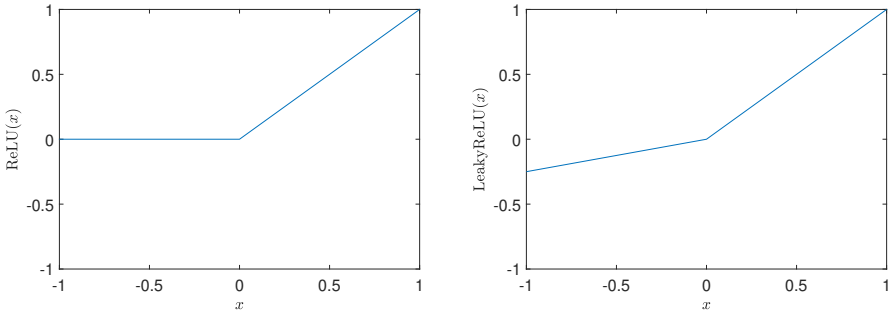
For the i^{th} layer, the output $\mathbf{y}_i$ is then:

$$\mathbf{y}_i = f^{(i)}(\mathbf{y}_{i-1}, \mathbf{W_i}, \mathbf{B_i}) = g^{(i)}\left(\mathbf{W_i}^{\mathrm{T}}\mathbf{y}_{i-1} + \mathbf{B}_i\right), \tag{3.4}$$

where $g^{(i)}$ is the non-linear function of the i^{th} layer, applied element-wise to its input $\mathbf{W_i}^{\mathrm{T}}\mathbf{y}_{i-1} + \mathbf{B}_i$, where $\mathbf{W_i}$ is the weight matrix of this layer, and $\mathbf{B}_i$ the corresponding bias vector.

Strictly speaking, the activation g at the different layers can also be linear, but a linear function of a linear function, would still be linear. Having a nonlinear activation function is essential to achieve the befamed nonlinear behaviour of DNNs.

The systems of this thesis mostly rely on the rectified linear unit (ReLU) or one of its close cousins (like LeakyReLu and PReLU) to introduce the non-linearity (see Figure 3.2). These functions are 'almost' linear (piecewise linear), which means the network will still have many of the properties that make linear models easy to optimize with gradient-descent-based algorithms. However, when there is one such 'slightly' non-linear component at each and every node of the hidden layers in the network, the resultant mapping functions can be extremely complicated.



(a) ReLU **(b)** PreLU or LeakyReLU

Figure 3.2: Different ReLU type activation functions. The difference between PreLU and LeakyReLU is that the slope for $x < 0$ is learned, while for LeakyReLU the slope is fixed to a specific value.

As shown in Figure 3.2, the ReLU function's output is either equal to its input, or zero, if the input is below zero:

$$\text{ReLU}(\mathbf{z}) = \max\{0, \mathbf{z}\}. \tag{3.5}$$

Each layer in a feed forward network is a function, with an input based on the layer before, and an output that is sent to the next layer. As such, the layers become a function of nested functions.

For the network of Figure 3.1 there are two hidden layers and one output layer, this gives:

$$\hat{\mathbf{y}} = f(\mathbf{y}_0) = f^{(3)}\left(f^{(2)}\left(f^{(1)}(\mathbf{y}_0, \mathbf{W_1}, \mathbf{B_1}), \mathbf{W_2}, \mathbf{B_2}\right), \mathbf{W_3}, \mathbf{B_3}\right), \tag{3.6}$$

where $f^{(1)}$ is the first hidden layer, $f^{(2)}$ the second hidden layer, and $f^{(3)}$ the output layer. The input layer merely represents the network's input (not a function), and is therefore equal to the input vector $\mathbf{y}_0$. The other layers have the previous layer as input, and all layers have their own respective weights $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ and biases $\mathbf{B}_1$, $\mathbf{B}_2$ and $\mathbf{B}_3$.

To ensure it is possible to get any possible output (including negative numbers), linear activation is often used at the output layer. Then for the network of Figure 3.1:

$$\begin{aligned}\hat{\mathbf{y}}_1 &= \max\left\{0, \mathbf{W_1}^{\mathrm{T}}\mathbf{y}_0 + \mathbf{B_1}\right\}\\ \hat{\mathbf{y}}_2 &= \max\left\{0, \mathbf{W_2}^{\mathrm{T}} \max\left\{0, \mathbf{W_1}^{\mathrm{T}}\mathbf{y}_0 + \mathbf{B_1}\right\} + \mathbf{B_2}\right\}\\ \hat{\mathbf{y}} &= \mathbf{W_3}^{\mathrm{T}} \max\left\{0, \mathbf{W_2}^{\mathrm{T}} \max\left\{0, \mathbf{W_1}^{\mathrm{T}}\mathbf{y}_0 + \mathbf{B_1}\right\} + \mathbf{B_2}\right\} + \mathbf{B_3},\end{aligned} \tag{3.7}$$

where $\hat{\mathbf{y}}_j$ is the output of a specific layer, with j as the index of that layer.

A key property of fully connected feed forward layers is that every output unit is based on every input. This means these models can learn extremely complicated mappings, but they are also prone to overfit. The issue of overfitting can be reduced with regularization techniques. However, a related problem is that there is 'a lot to learn' due to all the possible relations between all inputs and output — many of which may have little relevance.

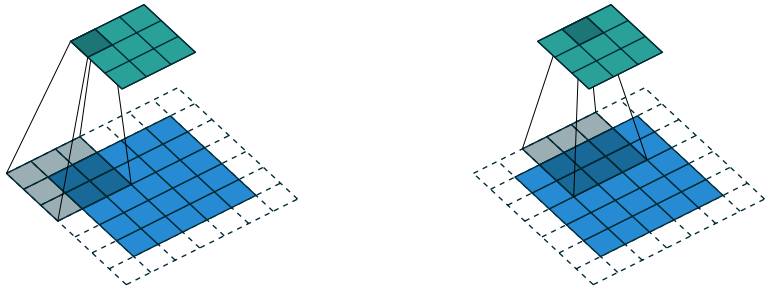
To feed noisy speech to a fully connected neural network, the vectors of the STFT frames can be stacked into a single long input vector. It is likely that there are certain frequencies that tend to be more noise dominated, but the values for this particular frequency band are now located separately in the input vector. This does not matter for the feed forward network that does not assume stronger relationships between values that are closer to one another in the input. However, it does mean that the feed forward layers always have to learn such relationships from 'scratch', which requires extra data.

3.2.2 Convolutional layers

Unlike fully connected layers, convolutional networks have layers with 'sparse interaction' — the outputs depend only on a subset of the inputs. Here the input is not stacked into a single vector, but instead kept as a 2D matrix. For speech enhancement, $\mathbf{y}_0$ is then generally some form of spectral input. For the convolutional layer to be the 'right' type of layer to use, it is important that $\mathbf{y}_0$ contains localised information, i.e. that there is meaning behind the fact that values are close to one another in the matrix.

As the name suggests, the convolution operation is key for convolutional layers. The input $\mathbf{y}_0$ is namely convolved with a kernel: the matrix containing the trainable weights $\mathbf{W}$. To result in a sparsly connected layer, the kernel must be smaller than the input.

The convolution process is shown in Figure 3.3. The kernel (in this case sized 3x3) slides over the 2D input $\mathbf{y}_0$, creating a single output value in each sliding step. Here it is also possible to increase the step size, or 'stride' above 1, which leads to subsampling.



(a) Step 1 **(b)** Step 2

Figure 3.3: Visualization of the convolution operation. Subfigures display how the kernel (grey, 3x3) slides over the input (blue, 5x5), here shown with stride 2. Dashed lines indicate zero padding. Images taken from [64], Vincent Dumoulin and Francesco Visin, distributed under the MIT Licence.

Figure 3.4 shows a typical visualization of an input layer and two convolutional layers. Following the same notation as in Section 3.2.1, the three stacked

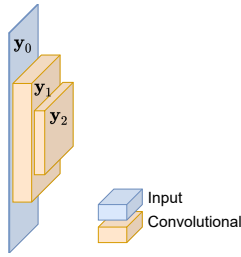


Figure 3.4: Convolutional neural net

convolutional layers of Figure 3.4 give:

$$f(\mathbf{A}_1) = f^{(2)}\left(f^{(1)}(\mathbf{y}_0, \mathbf{W_1}, \mathbf{B_1}), \mathbf{W_2}, \mathbf{B_2}\right), \tag{3.8}$$

where the output layer $f^{(3)}$ has been removed.

Most deep learning libraries do not contain convolution directly, but instead the related cross correlation function. This is the same as convolution, without 'flipping the kernel' (see [61] for a detailed explanation and derivation). As the weights are learned, their exact location in the kernel has no significance, meaning that in practice cross correlation and convolution give the same final performance.

The kernel ($\mathbf{W}$) could have the exact same number of dimensions as the input to the convolutional layer (as shown in Figure 3.3), but it is common to give it one extra dimension for the number of desired output channels. These 'channels' should not be confused with audio channels. Instead the number of channels indicates the number of 'feature maps' coming out of the convolutional layer. Intuitively, these kernels can be seen as a set of filters (one for each output channel), where each filter obtains different features from the input.

The cross correlation function is linear, so again a non-linear activation is needed. Like in Section 3.2.1, ReLU is applied element-wise to the output of the convolution for this example. Given a (spectral) 2D input $\mathbf{y}_0$, the output of the first convolutional layer ($\hat{\mathbf{y}}_1$) is then obtained as follows:

$$\hat{\mathbf{y}}_1 = \begin{bmatrix} \hat{\mathbf{y}}_{1,1} \\ \hat{\mathbf{y}}_{1,2} \\ \vdots \\ \hat{\mathbf{y}}_{1,N_2} \end{bmatrix}, \tag{3.9}$$

where N_2 is the number of desired output feature maps, and

$$\hat{\mathbf{y}}_{1,i} = \max\left\{0, \mathbf{B}_{1,i} + \mathbf{W}_{1,i} \star \mathbf{y}_0\right\}, \tag{3.10}$$

where $\star$ indicates cross correlation, $\mathbf{B}_1$ and $\mathbf{W}_1$ are the trainable bias and weight matrices of the first convolutional layer, and i is the index of the output feature map.

The output of the first layer ($\hat{\mathbf{y}}_1$) has one more dimension than the original input ($\mathbf{y}_0$) as there are now N_2 feature maps. The second convolutional layer (with output $\hat{\mathbf{y}}_2$) will therefore have to aggregate the feature maps. Subsequently:

$$\hat{\mathbf{y}}_2 = \begin{bmatrix} \hat{\mathbf{y}}_{2,1} \\ \hat{\mathbf{y}}_{2,2} \\ \vdots \\ \hat{\mathbf{y}}_{2,N_3} \end{bmatrix}, \tag{3.11}$$

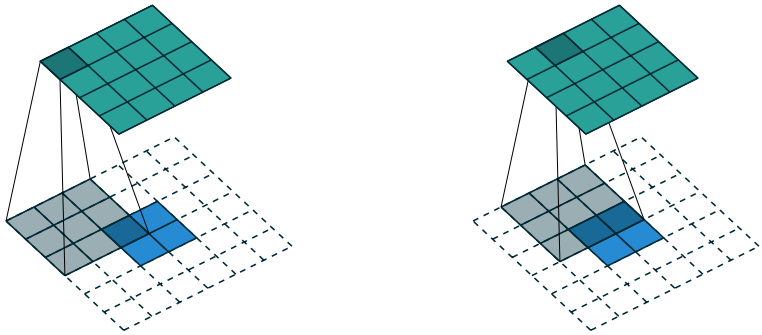
where N_3 is the number of output feature maps for the second layer, and

$$\hat{\mathbf{y}}_{2,i} = \max\left\{0, \mathbf{B}_{2_i} + \sum_{j=1}^{N_2} \mathbf{W}_{2,i,j} \star \hat{\mathbf{y}}_{1,j}\right\}. \tag{3.12}$$

The convolutional layers are still feed forward: There is no information going back from later layers to earlier layers. Due to the limited size of the kernel, the layers are, however, not fully connected. Another important difference is that for the fully connected networks, the number of nodes in the network defines the length of the vector obtained at that layer. Therefore, for a fully connected layer, the number of nodes in the output layer can always be set to be equal to the input vector length. This makes sense for a speech enhancement network where the output is either a cleaned version of the input, or a mask: it should have the same dimensions.

Now with the convolutional layers, the output dimensions depend on the number of filters in $\mathbf{W}$. While it is desirable to have many filters (to obtain many different feature maps), it also must be possible to go back to a tensor that represents the speech signal, or a mask of this signal.

Therefore, transposed convolutional layers can be applied to reverse the convolution operation. This is demonstrated in Figure 3.5.



(a) Step 1 **(b)** Step 2

Figure 3.5: Visualization of the transposed convolution operation. Subfigures display how the kernel (grey, 3x3) slides over the input (blue, 5x5), here shown with unit stride. Dashed lines indicate zero padding. Images taken from [64], Vincent Dumoulin and Francesco Visin, distributed under the MIT Licence.

3.2.3 Recurrent layers

Both fully connected and convolutional layers are feed forward: no output is ever presented to anything other than later layers. Another interpretation of this is that feed forward layers have no memory. This is where recurrent layers are crucially different.

For human listeners, context is very important for understanding speech. Even when just considering the speech signal itself (ignoring other context such as location, the time of the day, etc.), misheard words can be guessed, based on what comes before and after those words. On a smaller timescale, this process is (almost) completely subconscious: in noisy situations, humans are constantly guessing and filling in the blanks [65].

From that perspective, it makes sense to let SE model 'remember' what has been said (and for applications where delay is acceptable, to include what comes after). One option for this is to put past (and future frames) together with the current STFT frame in the input. Alternatively, it is possible to use recurrent layers that can *learn* what information should be remembered.

A particular popular recurrent layer is the long short-term memory (LSTM) layer [66]. Figure 3.6 shows an LSTM 'layer', which actually consists four interacting neural network layers. An LSTM has three inputs: the previous cell state $\mathbf{c}_{t-1}$, the previous hidden state $\mathbf{h}_{t-1}$, and the actual input vector $\mathbf{y}_{0,t}$, as used before, but now with a t subscript to keep track of the different time steps. The initial values of the cell and hidden states are zero.

The LSTM has only two outputs: the updated cell state and the $\mathbf{c}_t$, the updated hidden state $\mathbf{h}_t$. Both will be remembered for reuse in this same LSTM layer, while $\mathbf{h}_t$ will additionally be passed on to the next (LSTM) layer: acting as $\mathbf{y}_{0,t}$ for the next layer.

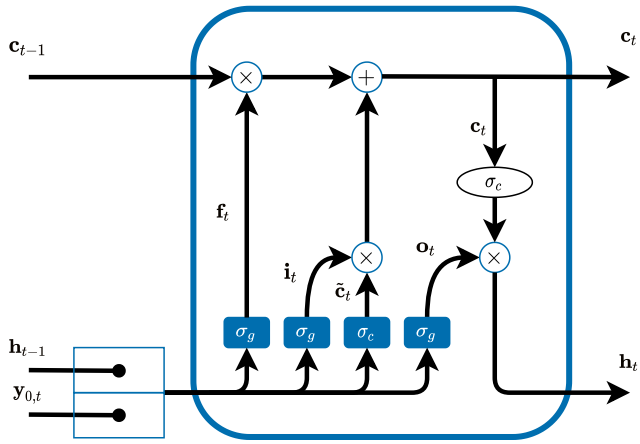
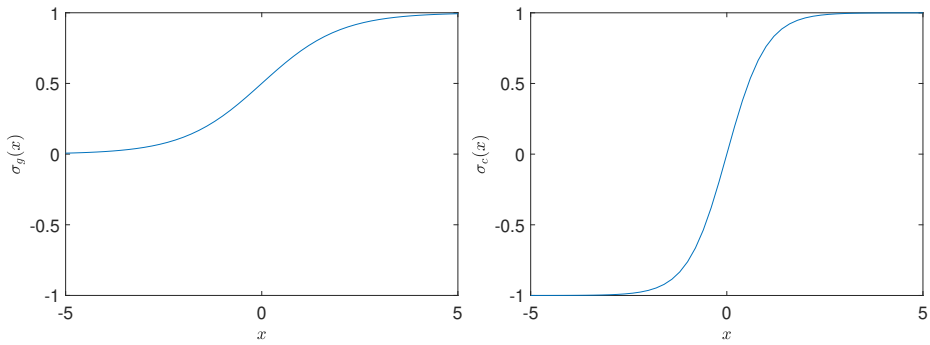


Figure 3.6: A single LSTM layer

An LSTM layer contains four neural layers named i) the forget gate (with output $\mathbf{f}_t$), ii) the input gate (with output $\mathbf{i}_t$), iii) the output gate (with output $\mathbf{o}_t$) and iv) the cell input (with output $\tilde{\mathbf{c}}_t$). Each of these layers has their own set of weights for the input $\mathbf{y}_{0,t}$ ($\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_o$ and $\mathbf{W}_c$), weights for the hidden state $\mathbf{h}_t$ ($\mathbf{U}_f$, $\mathbf{U}_i$, $\mathbf{U}_o$ and $\mathbf{U}_c$) and biases ($\mathbf{B}_f$, $\mathbf{B}_i$, $\mathbf{B}_c$ and $\mathbf{B}_o$.)

All four neural layers have a nonlinear activation, but it is not a ReLU type function. Instead, three of the layers act as gates, because they rely on the sigmoid activation (σ_g). The cell input layer with $\tilde{\mathbf{c}}$ output is the exception, and relies on the hyperbolic tangent function (σ_c). These activation functions are shown in Figure 3.7.



(a) sigmoid (σ_g) **(b)** tanh (σ_c)

Figure 3.7: The sigmoid (σ_g) and tanh (σ_c) activation functions

As shown in Figure 3.6, the updated cell state $\mathbf{c}_t$ is calculated as follows:

$$\mathbf{f}_t = \sigma_g(\mathbf{W}_f \mathbf{y}_{0,t} + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{B}_f) \tag{3.13}$$
$$\mathbf{i}_t = \sigma_g(\mathbf{W}_i \mathbf{y}_{0,t} + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{B}_i) \tag{3.14}$$
$$\tilde{\mathbf{c}}_t = \sigma_c(\mathbf{W}_c \mathbf{y}_{0,t} + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{B}_c) \tag{3.15}$$
$$\mathbf{c}_t = \mathbf{f}_t * \mathbf{c}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{c}}_t. \tag{3.16}$$

Given the updated $\mathbf{c}_t$, the actual LSTM output ($\mathbf{h}_t$, also called its hidden state), becomes:

$$\mathbf{o}_t = \sigma_g(\mathbf{W}_o \mathbf{y}_{0,t} + \mathbf{U}_o h_{t-1} + \mathbf{B}_o) \tag{3.17}$$
$$\mathbf{h}_t = \mathbf{o}_t * \sigma_c(\mathbf{c}_t) \tag{3.18}$$

This means that information in the network is no longer just moving forward. Instead, the output of an LSTM layer is reused as input to that same layer. This gives the LSTM layer the ability to remember useful information from earlier frames[1].

While it is possible to use LSTMs as the major building blocks for SE systems, the later papers included in this thesis relied on a combination of convolutional and LSTM layers, to benefit from their different strengths.

3.2.4 Convolutional recurrent encoder-decoder structure

For speech enhancement the output of the model is often either the enhanced input directly, or a mask to be applied to the input. In both cases, the input and output of the network have equal dimensionality. Encoder-decoder model architectures are suitable for this purpose.

In a model with an encoder-decoder architecture, the input is translated to a latent space by an encoder, while the decoder translates tensors from this latent space to the output. Other layers (those between the encoder and decoder) can act on the encoded input, before the decoder is applied, but not all encoder-decoder architectures have such layers. Any kind of layer can be used both in the encoder and the decoder. Examples of well known neural networks with an encoder-decoder architecture are SegNet [67] with only convolutional layers, and Google's recurrent (see Section 3.2.3) sequence to sequence learning neural network [68]. Fully connected layers are mostly used for autoencoders (i.e. [69]): a type of encoder-decoder architecture where the training targets are equal to the input of the model.

In a convolutional encoder-decoder structure, convolutional layers are used to first encode the signal into relevant feature spaces where the task of the model is easiest to achieve. This task can be to separate speech from noise — the main

[1] It is also possible to implement a so-called bi-directional LSTM layer, that doesn't just remember from the past, but also relies on input from future frames. Its implementation is conceptually similar to the LSTM discussed here, but outside of the scope of this work as none of the final SE networks included in this thesis had bi-directional LSTM layers

interest of this thesis — or something completely different (like reducing redundant information, or segmentation). Then, the encoding process is reversed with a set of *transposed* convolutional layers in the decoder.

Ronneberger *et al.* [70] proposed the 'U-Net' architecture for biomedical image segmentation: a convolutional encoder-decoder with so called 'skip connections'. Figure 3.8 shows the general layout of this kind of network.

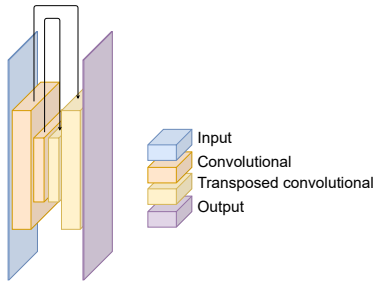


Figure 3.8: Encoder-decoder structure with convolutional layers, transposed convolutional layers and skip connections. Arrows indicate skip connections.

With skip connections, the output of the encoding steps is concatenated to the input of the respective decoder. This gives the decoder the input signal (transformed into a specific set of feature maps), in addition to the signal altered by later layers. This allows the network to make use of finer details that may have been lost on the way.

The original U-Net had max pooling layers for downsampling, but the convolutional encoder-decoder-based models used in this thesis instead use the stride of the convolution for downsampling. Additionally, LSTM layers are added between the encoder and decoder (while U-Net has no such layers). Further details of and argumentation for the exact chosen network architecture are presented in Section 5.2.

Chapter 4

Evaluating Enhancement

The goal of speech enhancement is to recover the speech signal $s[n]$ from the noisy signal $x[n]$. This turns speech enhancement systems into speech signal *estimators*. To be able to estimate $\hat{s}[n]$, a measure of how close the estimate is to its target, is needed. This in order to be able to say that the estimate is closer to it, than the noisy input was. Such a measure is relevant both to be able to train supervised SE systems, and to be able to evaluate their performance at a later stage.

It is common to separate the goal of speech enhancement into two dimensions: improving quality versus improving intelligibility. For both it is beneficial to reduce the noise and reverberance [45, 46]. However, quality and intelligibility are not the same thing.

The quality of a speech signal is determined by a person's opinion of that signal. Is it bad, or excellent? Preferred over another, or ranked as 'more annoying' or 'less natural'? The intelligibility of a speech signal is defined by how much a person understands of its content. How much can the listener repeat correctly?

An old-fashioned text-to-speech voice, reciting a phone number by merging separate voice recordings, is a typical example of a highly intelligible speech sample (all numbers are easily understood) that is low in quality; it is unnatural, robotic, and has a strange intonation. On the other end of the spectrum is a conversation in a foreign language. The signal can be as natural, clear and crisp as can be, but is quite possibly entirely unintelligible.

This shows that improving quality does not necessarily improve intelligibility, and vice versa. However, this somewhat counter-intuitive effect also occurs at a more subtle level. Already in the earliest papers included in this thesis, it was shown that reducing noise can lead to improved quality ratings of listeners, but quality came at the cost of intelligibility [Paper I][Paper II]. The fact that the price of increasing quality is often a reduction in intelligibility has been known for many decades [8].

Given that quality and intelligibility are independent factors, different methods are required to determine the quality and intelligibility of speech signals. For both, however, scores can be obtained in an objective manner, or a subjective manner. Here objective means that an algorithm provides a prediction of how the signal

will be perceived by human listeners. Subjective means that the scores are directly obtained from human listeners who have listened to different samples of speech.

Note that the aim of the measures of quality is to estimate how a signal is perceived (i.e. the user's opinion of this signal), and they are therefore often referred to as subjective measures.

In this thesis, the use of the terms objective and subjective are to be understood as follows:

- objective means based on an algorithm,
- subjective means based on listening tests.

Objective scoring is far more practical: quick, cheap and deterministic. Subjective testing is far more time-consuming, requires repeated measures to compensate for individual differences, and is more expensive. However, appropriate subjective testing provides the answer directly, instead of an estimate of this answer (which is what objective measures give us).

A main focus of the work for this thesis is on testing the validity and reliability of objective measures, and on comparing objective scores to subjective scores in three of the contributions [Paper I][Paper II][Paper V].

4.1 Speech Quality

The quality of a signal is a measure of opinion: Does the listener like the sound? There are many factors that play a part in this highly subjective measure: naturalness, clarity, pleasantness, brightness, etc. Quality evaluation often tries to capture all of these factors into a single score.

4.1.1 Subjective quality

There are many subjective methods for evaluating speech quality. These can be broadly categorized into two groups: relative preference (i.e. which out of the two clips do you prefer?) and absolute quality (i.e. rate the quality of this clip).

The mostly widely used methods require listeners to rate the quality of a speech signal on a five-point scale [71]. An average score, commonly referred to as the Mean Opinion Score (MOS) is obtained by averaging over multiple listeners. The ITU-T standard (P.835) [72] standardizes methodology to obtain MOS scores with focus on the speech signal alone, the background signal alone, and the complete signal. Table 4.1 shows the ratings and verbal scales of the English version of P.835.

The newer ITU-T standard (P.808 [73]) addresses how a P.835-like test can be performed with a crowd-sourcing approach, where the participants are connected via an online platform, and evaluate speech quality in their own environments, using their own devices. This has become a popular alternative to P.835, especially since Microsoft released an open source implementation of this standard, to be used in conjunction with their DNS-challenges [26–28]. P.808 implements several methods that attempt to compensate for the lack of having a controlled setting;

Table 4.1: English version of the ordinal scales used in ITU-T P.835 [72]. Table taken from [Paper II] (©2019 IEEE).

Rating	Speech	Noise	Overall quality
5	Not distorted	Not noticeable	Excellent
4	Slightly distorted	Slightly noticeable	Good
3	Somewhat distorted	Noticeable but not intrusive	Fair
2	Fairly distorted	Somewhat intrusive	Poor
1	Very distorted	Very intrusive	Bad

these methods include ensuring the listening environment works as expected, and including gold standard (samples with extremely high or alternatively extremely low quality, to check whether the participant responds as expected) and trap questions (questions that are not visually different, but ask the participant to select a specific answer).

However, this implementation was not yet available for the work included in this thesis. Instead, P.835, was implemented with Norwegian translations of the five point scales shown in Table 4.2, and listening tests were conducted in SINTEF's audio lab.

Table 4.2: Norwegian translation of the ordinal scales used in ITU-T P.835. Table taken from [Paper II] (©2019 IEEE).

Rating	Speech	Noise	Overall quality
5	Ikke forvrengt	Ikke hørbar	Veldig god
4	Litt forvrengt	Hørbar, men ikke påtrengende	God
3	Ganske forvrengt	Litt påtrengende	Middels
2	Betydelig forvrengt	Påtrengende	Dårlig
1	Voldsomt forvrengt	Veldig påtrengende	Veldig dårlig

The translation is based on the official English and French versions, and the Danish version presented in [74]. The Norwegian translation is not a literal translation from English. Instead of using 'slightly noticeable', 'noticeable but not intrusive' and 'somewhat intrusive' as rating 4, 3 and 2, the Norwegian version uses 'noticeable but not intrusive', 'somewhat intrusive' and 'intrusive'. The reason for this adjustment was that several of the participants of a pilot study provided feedback that it was difficult to distinguish between 'slightly noticeable' and 'noticeable but not intrusive'.

4.1.2 Objective quality

There are many objective measures that estimate the quality of a speech signal, but PESQ (Perceptual Evaluation of Speech Quality [75]) is the most popular. PESQ

is recommended by ITU (the same body that is behind the subjective evaluation recommendations P.835 [72] and P.808 [73]) for speech quality assessment in ITU P.862 [75] and ITU P.862.2 [76].

PESQ is an intrusive measure, where 'intrusive' means that it requires both a clean reference signal, and the noisy/distorted/processed signal to be tested. These two signals are then compared, and their 'closeness' determines the PESQ score. This closeness is based on several signal transforms that are perceptually motivated. The final PESQ score is based on a weighted sum of different components.

PESQ was originally standardized in ITU P.862 [75] for quality assesment of narrow-band telephone networks and speech codecs. With this original PESQ, the test and reference signals are first level aligned. Then they are filtered using an FFT to model the standard telephone handset. Next, the signals are also aligned in time and processed through an auditory transform. The results of this transform for the two signals are subtracted from one another, and two parameters of distortion are obtained. Finally, these parameters of distortion are aggregated in frequency and time and mapped to the subjective mean opinion scores (MOS) [77].

PESQ was then extended for wideband applications in ITU P.862.2 [76]. There were only two differences. Firstly, the filter that mimics the standard telephone handset was replaced with a filter that is more suitable for headphones, with a flat response above 100 Hz, and a gentle roll-off below 100 Hz. Separate filter coefficient are defined for 16 kHz (wideband) and 8 kHz (narrowband) input. Secondly, the output-to-MOS mapping function was replaced to include calibration for wideband listening test conditions.

There are therefore three variants of PESQ: the original PESQ for narrowband telephony [75], PESQ narrowband for 8 kHz input [76], and PESQ wideband for 16 kHz input [76].

Furthermore, ITU has updated P.862 with P.863: POLQA (Perceptual Objective Listening Quality Analysis) [78]. POLQA is calculated in a similar manner, but adds new capabilities for super-wideband (HD) and full-band voice signals, along with support for most recent voice coding and VoIP/VoLTE transmission technologies [78].

The early work for this thesis relied on POLQA, assuming this metric would take over PESQ in popularity, and because of licensing concerns. For the later work, both PESQ narrowband and wideband were used in order to accommodate easier comparison with work in the literature.

4.2 Speech Intelligibility

Speech intelligibility is a measure of how understandable a speech signal is. Intelligibility evaluation often tries to capture this concept into a single percentage indicating the ratio that is understood. Speech intelligibility can be measured by asking a listener to repeat what they have heard. Objective measures try to predict this response.

But the intelligibility of a signal is not only determined by how audible the different phonemes[1] of the speech are. Speech is embedded in a lot of context. Who is talking, the topic, the structure of sentences, the language used, etc. all aid the brain in filling in the parts that are not audible. Knowing that someone has a friend called Addison, will make a person (unconsciously) assume he/she is talking about Addison (instead of Madison), in noisy conditions where only the '-dison' part was actually audible. There are experiments that show how easy it is to trick the brain into 'hearing' something recognizable that is not being presented at all [79]. Here language familiarity is an important factor: the brain will only hear words it knows, to the degree that it simply cannot discern phonemes that it doesn't know, even in the absence of noise.

Additionally, to be able to repeat a segment of speech, the brain not only needs to hear the segment, but also needs to remember it. Remembering a longer segment (like a five-word sentence), is more difficult than remembering shorter segments (like a three-word sentence) in the same noise conditions. Context and memory effects also add up: it is nearly impossible to remember/repeat they lyrics of a song in a language you have only limited familiarity with.

This means that the exact speech intelligibility test setup highly influences the score/percentage obtained in subjective testing. From the SE system developer's point of view, this is not really a problem. The interesting part is being able to compare different processing pipelines: does the system increase intelligibility, or does it make it worse? It is also from this perspective, that this thesis aims to evaluate the predictive power of the objective measures.

4.2.1 Subjective intelligibility

As mentioned, language familiarity is very important for subjective intelligibility. To ensure access to a sufficient number of native speakers, a Norwegian speech-in-noise test was chosen: the five-word Hagerman sentence test proposed by Øygarden in [53], because the material of this test was designed to be suitable for repeated measurements on the same person.

Øygarden's test is suitable for repeated tests, because the five-word Hagerman sentences are build up with the form [Name]-[Verb]-[Numeral]-[Adjective]-[Noun], and there are ten alternatives for each of these word categories. As such, it is possible to generate 100 000 unique sentences; from these, a large number of phonemically balanced sets for testing. Therefore, the risk that a listener recognizes a sentence from an earlier test is negligible.

For this thesis, the internal SINTEF implementation[2] of Øygarden's test was adapted to the use-case. This implementation presents the listener with a table from which all the correct answers can be selected, see Figure 4.1.

[1] Phonemes are units of sound that can distinguish one word from another in a particular language

[2] Originally implemented by Tron V. Tronstad in 2009

Figure 4.1: GUI of the speech intelligibility test. Figure taken from [Paper I] (©2017 ISCA).

First, 500 noise and reverberation free 5-word sentences were obtained. For the earlier work, these sentences were corrupted with additive noise [Paper I][Paper II]. For the last study, the sentences were corrupted with measured RIRs and recorded noise samples [Paper V]. For all studies, each test sentence was corrupted with noise with SNRs ranging from -36 dB to 10 dB in steps of 2 dB. Sentences were then processed with the enhancement systems to be tested, to create multiple test sets, each containing 500 x 24 = 12 000 sentences.

Each listener first had to complete a training round of the speech-in-noise test, and then one round for each system to be tested. Systems were presented in random order to ensure the training effect would not affect the results averaged over all participants. Every round consisted of 20 sentences, each containing five words, giving 100 stimuli per round.

The SNR of each test sentence was dependent on the responses given by the listener, as they were calculated using an adaptive estimation procedure called the Ψ method [80]. This method attempts to always test at the SNR that provides the maximum amount of information required to estimate the psychometric function. The psychometric function shows each listener's intelligibility score (as a percentage) against SNR. After twenty sentences, the SNR at which 50 % of the test material was understood is obtained from the estimated psychometric function: the speech recognition threshold (SRT).

4.2.2 Objective intelligibility

As with objective quality evaluation, there are many objective intellgibillity metrics (OIMs) to estimate subjective intelligibility. Also here there is the distinction between intrusive measures (that require a clean reference signal in addition to the test signal), and non-intrusive measures (that need only the test signal). Intrusive measures generally give more accurate predictions than non-intrusive

measures [81], while the latter have a wider application range. However, for speech enhancement systems trained in a supervised manner, the clean reference is readily available (already required as a target), making intrusive measures the logical choice.

There are, however, many such intrusive measures. For this thesis, five relatively commonly used metrics were chosen for comparison to subjective results: the short-time objective intelligibility (STOI) [82, 83], the extended STOI (ESTOI) [84], the normalized covariance metric (NCM) [85, 86], the coherence speech intelligibility index (CSII) [87], and the hearing-aid speech perception index (HASPI) [88, 89]. A brief descriptions of these metrics is provided below.

All these methods are similar in the sense that first the reference and test signal are transformed, and then a measure of how close the two signals are is obtained. However, the intelligibility of speech does not only depend on the speech degradation, but also on the subjective testing conditions. Longer test sentences, for example, are more difficult to remember, than single words. Test sentences can also contain context from which the listener can guess missing parts. Therefore the speech recognition thresholds obtained for the exact same speech material can vary widely even for the same individual.

As such, it is common to map the measure of proximity to the average intelligibility percentages obtained for listeners during subjective testing. Some of the metrics come with this mapping (obtained for various testing conditions), while others are only proposed as a score that still has to be mapped.

For the speech enhancement application, only the change in intelligibility from the unenhanced to the enhanced signal is of relevance (not the absolute scores of the listeners, specific to the testing conditions). Therefore, for [Paper V], all metrics (including those that already came with a mapping function) were mapped to the subjective results of the noisy unprocessed baseline test condition during subjective testing. Then the test setup and the mapping were kept constant for evaluation of all other processing conditions (the different models to be tested).

For STOI and ESTOI, the mapping function proposed [82–84] was used:

$$\hat{I} = \frac{100}{1 + \exp\left(a\tilde{I} + b\right)}, \tag{4.1}$$

where $\hat{I}$ is the predicted intelligibility (in percentage correct), $\tilde{I}$ the predicted score from STOI or ESTOI, and a and b are the mapping coefficients to be determined with the non-linear least squares method.

This same mapping function was empirically found to also work well for NCM scores. For CSII, non-linear least squares were used to find the coefficients a_1, a_2, a_3, and a_4 of the mapping function proposed by the original authors of CSII [87]:

$$c = a_1 + a_2\mathrm{CSII}_{\mathrm{Low}} + a_3\mathrm{CSII}_{\mathrm{Mid}} + a_4\mathrm{CSII}_{\mathrm{High}}, \tag{4.2}$$

$$\hat{I} = \frac{100}{1 + \exp\left(-c\right)}, \tag{4.3}$$

where $CSII_{Low}$,$CSII_{Mid}$ and $CSII_{High}$ are separate CSII scores for all segments with low-level, medium-level, and high-level speech (See Section 4.2.2).

For HASPI (which has already been fit to subjective data), it was found that simply translating objective results along the SNR-axis gave the closest match.

The following sections discuss the chosen intellgibility metrics in greater detail.

NCM (normalized covariance metric)

The normalized covariance measure (NCM) was originally proposed in [85], and updated with signal dependent weights in [86].

For NCM, the test signal and the clean reference are first band-pass filtered into different frequency bands. Then the normalized covariance (the Pearson correlation coefficient) is calculated for each of the frequency bands. Using the normalized covariances, an apparent SNR for each frequency band is calculated. These SNRs are limited to be within a -15 dB to 15 dB range. A linear transformation subsequently turns the apparent SNRs into a single transmission index (TI) for each frequency band. Finally, the NCM index is obtained by taking a weighted average of the transmission indices of all frequency bands.

CSII (coherence speech intelligibility index)

The coherence speech intelligibiligy index (CSII) was originally proposed in [87] as an extension of the speech intelligibility index (SII) metric, which is standardized in ANSI S3.5-1997 [90].

With CSII, the clean reference signal is first separated into windowed segments. For each of these windows, the root-mean-squared (RMS) level is calculated, and used to determine whether this particular segment is a low-level segment (10 to 30 dB below the overall RMS), a mid-level segment (0 to 10 dB below the overall RMS), or a high-level segment (at or above the overall RMS). Then a separate CSII score ($CSII_{Low}$,$CSII_{Mid}$ and $CSII_{High}$, respectively) is calculated for for these three levels.

For the CSII calculations, first the reference and test signal are time-aligned. Then the magnitude-squared coherence function is obtained, from which a speech power spectrum and noise power spectrum can be estimated. These are then used to obtain a signal-to-distortion ratio (SDR) for each frequency band. From this SDR, and the estimated speech and noise power spectra, the CSII is calculated following the SII procedure, standardized in ANSI S3.5-1997 [90].

STOI (short-time objective intelligibility)

Short-time objective intelligibility (STOI) index was originally proposed in [82, 83].

With STOI, the signal first has to be resampled to a 10 kHz signal. Silent regions (those segments with at least 40 dB less energy than the maximum energy of the reference signal speech frames) are removed. Then a one-third octave band

analysis is performed for both signals, and for each band (each time-frequency (TF) unit) the norm is obtained. From these TF units, the short-time temporal envelope of both the reference and the test signal. After this step, the test signal's envelope is normalized and clipped, before being compared to the refence signal's envelope with the Pearson correlation coefficient. Finally, the STOI score is obtained by averaging the correlation coefficient results over all bands and frames.

ESTOI (extended STOI)

The extended short-time objective intelligibility (ESTOI) was originally proposed in [84] as an extension to STOI.

ESTOI is similar to STOI, but does not assume mutual independence between frequency bands and incorporates spectral correlation, to improve its performance on modulated noise sources.

HASPI (hearing-aid speech perception index)

HASPI was first introduced in [88], and later updated to better predict the intelligibility of reverberant speech (HASPI version 2) [89].

HASPI relies on a complex auditory model that includes biologically motivated steps, such as a middle-ear transfer function, an auditory filterbank, outer hair-cell dynamic-range compression, two tone suppression and adaptation of the inner hair-cell firing rate. The model can also account for hearing loss.

With HASPI version 1, both the reference signal and the test signal are passed through the auditory model, which outputs both a signal envelope and a signal temporal fine structure (TFS). The two signal envelopes are compared with cepstral correlation. The TFS parameters are compared by averaging the cross correlation coefficient calculated for each segment in each frequency band. Finally, the cepstral correlation and auditory coherence results are weighted and transformed with a logistic function to obtain the HASPI index score. Here the weights were found by mapping the HASPI index to subjective intelligibility datasets.

With HASPI version 2, two modifications were made. Firstly, an envelope modulation filterbank was used instead of the TFS calculation, and secondly, the parametric model was replaced with a neural network.

Chapter 5

Single Channel Speech Enhancement

The guiding concept behind machine learning is that the mapping between the input and output is learned, rather than designed. Thus, feature engineering — transforming model input with rule-based models — is less relied upon than before. The idea is that the complexity of DNNs and the abundance of training data allows the systems to learn the optimal mapping directly from the data. There are therefore researchers working with end-to-end SE systems [91–93], where the time domain systems are directly used as input and output.

However, the best systems from recent challenges still include STFT preprocessing [29, 30, 94, 95]. An important argument against end-to-end SE systems, is that, for example, the discrete Fourier transform step leads to more informative features at less computational cost than a network that has to find this mapping from the data. It is questionable whether one should try to learn something, when there are solutions available that do not require learning and cost less computational power.

Thus, in addition to the trainable DNN model, the speech enhancement systems used in the work for this thesis have preprocessing and postprocessing steps without any learnable parameters. The following sections describe the two main single channel systems used during this thesis.

5.1 Fully Connected Log Magnitude Estimator

Figure 5.1 shows an overview of the fully connected log magnitude estimator (FCLME) systems.

This model was inspired by the model proposed by Xu *et al.* [25], but pretraining with restricted Boltzmann machines was omitted, and LeakyReLU activation was used instead of sigmoid activation.

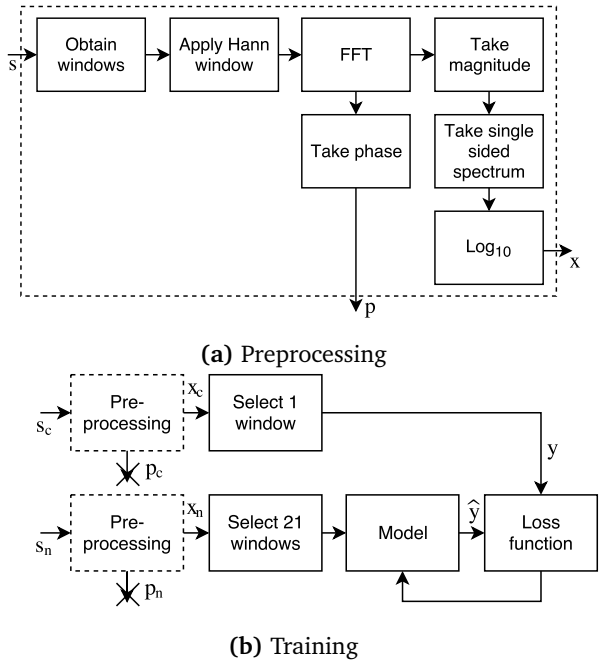


Figure 5.1: Training of the single channel FCLME system. Here s is a speech signal, x the $\log_{10}$ magnitude preprocessing output, and p the phase. The subscripts c and n indicate 'clean' and 'noisy', respectively. The model's target is y, while $\hat{y}$ is the model's estimate. Figures taken from [Paper I] (©2017 ISCA).

Three different variants of the FCLME were tested [Paper I][Paper II]:

1. **Model 1**: The system of Figure 5.1, trained to remove noise
2. **Model 2**: The same system as 1, but trained to *reduce* noise, instead of remove it. This means that the target speech samples had a 10 dB lower SNR than the noisy input
3. **Model 3**: The same system as 1, but with an extra postprocessing step as shown in Figure 5.2

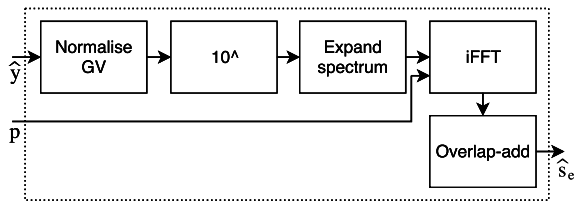


Figure 5.2: The postprocessing steps of FCLME, Model 3. The postprocessing pipeline for Models 1 and 2 are nearly identical, but for these models the 'Normalise GV' step is skipped. Here $\hat{y}$ is the model's estimate, p the noisy phase of the input speech, and $\hat{s}_e$ the estimated speech. Figure taken from [Paper I] (©2017 ISCA).

5.1.1 The data

The FCLME systems were all trained with noisy speech (no reverberance, additive noise only), downsampled to 8 kHz. The noise dataset contained 104 noises (first collected by Xu et al.) from either the Aurora database [96] or Guoning Hu's collection [54]. Clean speech was obtained from the Norwegian-language library NB Tale, often referred to as 'Språkbanken' [52]. SNRs ranged from -5 dB to 20 dB.

The speech data for validation was taken from the same speech database, but the validation and training sets were balanced with respect to gender and dialect, and no specific speakers or sentences occurred in more than one set. Unseen noise types for validation were taken from the Aurora [96] and NOISEX-92 [97] databases.

This gave 1984 hours of training data and 98 hours of validation data.

5.1.2 Training

The processing pipeline starts with a preprocessing/feature extraction block. This is where the raw time-domain samples are transformed into to the input to the neural network. For this system, the neural network input $\mathbf{y}_0$ equals $\log_{10}|X[n,k]|$, the noisy log magnitude spectrum. This was obtained through the STFT operation as explained in Section 2.1.2. The noisy phase was kept for later use during postprocessing, but did not pass through the network.

The neural net was trained to predict the clean log magnitude spectrum, one frame of the STFT at a time. For this purpose it was fed with 21 noisy frames: the current frame, ten historic frames, and ten future frames. Note that these frames were all 50 % overlapping. Inclusion of this kind of context allows the network to improve its estimate, based on the signal before and after the current timeframe — a concept that is familiar to human hearing where the brain 'fills in the gaps' when listening to noisy speech.

The MSE loss function was used to compare the noisy and estimated log magnitude spectra. The system was loosely based on the system proposed by Xu *et al.* [25]. However, the unsupervised pretraining stage with multiple restricted Boltzmann machines (RBMs) was skipped and LeakyReLU activation was used (instead of sigmoid), together with the Adam optimizer and 50 % dropout.

5.1.3 Usage

Figure 5.3 shows the SE system in enhancement mode. Once the network has estimated a log magnitude spectrum, postprocessing is required to obtain a time domain signal. During postprocessing, the preprocessing steps are reversed, and the noisy phase is used together with the inverse STFT to obtain a time domain signal. Postprocessing steps for Model 3 are shown in Figure 5.2. The Postprocessing steps for Model 1 and 2 are almost identical, but for these models the 'Normalise GV' step is skipped.

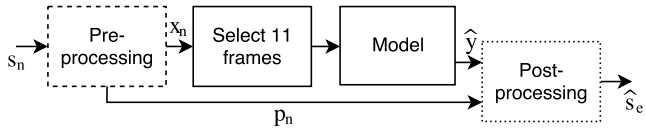


Figure 5.3: The single channel FCLME system during use. In this Figure, s_n is the noisy input speech, p_n the noisy phase, x_n the noisy $\log_{10}$ magnitude preprocessing output, $\hat{y}$ the model's esimate, and $\hat{s}_e$ the estimated enhanced speech. Figure taken from [Paper I] (©2017 ISCA).

5.1.4 The model layers

The DNN model had three fully connected layers, each layer with 2048 nodes, all with LeakyReLU activation. The output layer had linear activation.

5.1.5 Evaluation

Two test conditions were used to evaluate the different variants of the FCLME model. Both test conditions relied on the subjective speech material of Øygarden's Norwegian speech-in-noise test (see Section 2.2.1). The sentences were corrupted with additive noise at SNRs ranging from -36 dB to 10 dB in steps of 2 dB. The test conditions differed according to the type of locally recorded noise used: traffic noise versus babble noise (See Section 2.3.2).

Models 1 and 2 were evaluated for objective quality with POLQA and for subjective quality with a Norwegian translation of ITU-T P.835. Here both the traffic and babble noise conditions were tested with 23 listeners.

All three model variants were evaluated for objective intelligibility with STOI and for subjective intelligibility with Øygarden's Norwegian speech-in-noise test (see Section 4.2.1). For this purpose, two rounds of subjective intelligibility tests were conducted. During the first round, subjective performance of Model 1 and Model 3 was evaluated with 15 listeners for the traffic noise conditions. Then, during the second round, 12 listeners evaluated Model 1 (babble noise only), and Model 2 (both noise types).

All subjective intelligibility evaluation results were tested for significance with the Wilcoxon signed rank test. For the subjective quality evaluation, a cumulative link model (clm) from the ordinal package [98] in R [99] was used. The objective quality results were tested with a two sample t-test.

5.2 Deep Complex Convolutional Recurrent Mask Estimator

Later work included in this thesis was based on the Deep Complex Convolutional Recurrent Network (DCCRN) originally proposed by Hu *et al.* in [30].

For the work of this thesis, the DCCRN of [30] was adapted in only two ways: different training data was used (with the major difference being to rely on the RIRs simulated with directed speakers, as described in Section 2.4.1), and some minor hyperparameter optimization was conducted. To distinguish the original DCCRN by Hu *et al.* from the DCCRN system used here, the system will be referred to as the DCCRN-dir where the difference is relevant. Note that this distinction was not used in the paper where the system was proposed ([Paper IV]). DCCRN-dir is used both as a single channel baseline system, and as a building block that is incorporated into the multichannel systems described in Chapter 6.

The single channel DCCRN system was objectively evaluated in [Paper IV] and subjectively evaluated in [Paper V].

5.2.1 The data

The single channel DCCRN-dir was trained with noisy and reverberant input speech. The speech and noise datasets for training were obtained from the DNS Challenge dataset [27], while the RIRs were simulated using the six different methods described in Section 2.4.1. Pilot tests did not provide evidence that training with different RIR sets would give different final performance. Therefore, the ISM-dir dataset was chosen, as it led to the highest performance with respect to direction of arrival estimation in [Paper III]. The network was trained to remove noise, but the target speech was still reverberant.

All speech data was resampled to 16 kHz and cut into 4 s segments. With 441 hours of clean speech, 70 000 noise clips, and 24 000 unique RIRs, the number of unique random combinations of these was essentially limitless.

5.2.2 Training

Figure 5.4 shows the training process of the DCCRN. As with the fully connected feed forward system, the processing pipeline starts with a preprocessing/feature extraction block where the raw time domain samples are transformed to the input to the neural network.

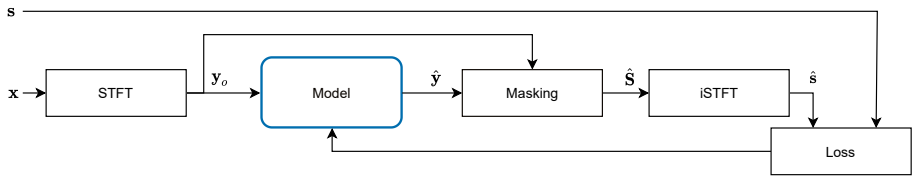


Figure 5.4: Training of the single channel DCCRN-based SE system. Here $\mathbf{s}$ is the reverberant speech signal, $\mathbf{x}$ the noisy reverberant speech signal, $\mathbf{y}_0$ the concatenation of the real and imaginary STFT coefficients of the noisy reverberant speech signal, $\hat{\mathbf{y}}$ the estimate of a mask for the model's input, $\hat{\mathbf{S}}$ the estimate of the STFT coefficients of reverberant speech, and $\hat{\mathbf{s}}$ the estimate of the reverberant speech signal.

For this system, however, $\mathbf{y}_0$ was the concatenation of $\mathbf{y}_{0,\Re}$ and $\mathbf{y}_{0,\Im}$, where

$$\mathbf{y}_{0,\Re} = \Re\left(\mathbf{X}[m,k]\right) \quad \text{and} \tag{5.1}$$

$$\mathbf{y}_{0,\Im} = \Im\left(\mathbf{X}[m,k]\right). \tag{5.2}$$

The coefficients of $\mathbf{X}[m,k]$ were obtained through the STFT operation as explained in Section 2.1.2. Only the current frame was presented to the DCCRN, as recurrent layers indirectly allow the network to use historic context (See Section 5.2.4).

The DCCRN was not trained to clean the noisy input, but instead to estimate a mask, to be applied to the noisy input. This means that the input to the network was also used for a postprocessing block that is applied during training.

The model's output $\hat{\mathbf{y}}$ was a complex mask, defined in the Fourier domain, and consisted of $\hat{\mathbf{y}}_{\Re}$ concatenated with $\hat{\mathbf{y}}_{\Im}$. Hu *et al.* found that the the masking approach in polar coordinates led to the best performance [30]. In accordance with those findings, the estimated STFT coefficients of the clean speech ($\hat{\mathbf{S}}$) were obtained as follows:

$$\hat{\mathbf{S}} = \left|\mathbf{y}_0\right| \cdot \left|\hat{\mathbf{y}}\right| \cdot e^{\theta_{\mathbf{y}_0} + \theta_{\hat{\mathbf{y}}}}, \tag{5.3}$$

where the magnitudes and phases of the model input and output were obtained as described in Section 2.1.2.

Another key feature of the DCCRN training process is that the loss was obtained in the time domain. Namely, the system was trained with the SI-SNR loss, $L_{\text{SI-SNR}}$ [100]:

$$\mathbf{s}_{\text{target}} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \cdot \mathbf{s}}{\|\mathbf{s}\|^2} \tag{5.4}$$

$$\mathbf{e}_{\text{noise}} = \hat{\mathbf{s}} - \mathbf{s} \tag{5.5}$$

$$L_{\text{SI-SNR}} = 10 \log_{10} \left(\frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2} \right), \tag{5.6}$$

where $\hat{\mathbf{s}}$ is the estimate of the clean speech $\mathbf{s}$, which is obtained from $\hat{\mathbf{S}}$ via the inverse STFT transform.

5.2.3 Usage

Usage of the network did not require any additional steps as a time-domain signal is already produced during the training process.

5.2.4 The model layers

Figure 5.5 shows the layers of the DCCRN. A set of six convolutional layers encoded the signal into learned feature spaces. They relied on PReLU activation and no max-pooling was added. The kernel depths of the encoder layers were: 32, 64,

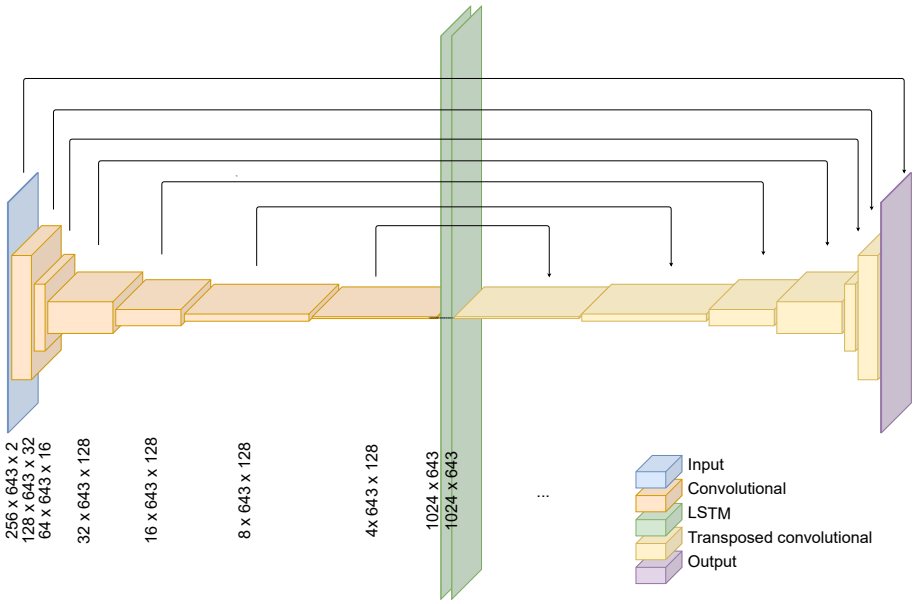


Figure 5.5: DCCRN model for the single channel SE system

128, 128, 256, and 256. The stride of the convolution was set to two. The figure shows how the dimensions of the input changed at each step.

Then, in the learned feature spaces most of the mask is, in essence, estimated in a form that is translated to the encoded feature space. The LSTM layers can retain information from earlier layers for this purpose. There were two LSTM layers, each with 256 nodes.

A set of subsequent transposed convolutional decoder layers returned the signal to the Fourier domain. Also here PReLU activation was used. Skip connections between corresponding encoder and decoder layers ensured that the decoder also had the encoder input available at each decoding steps. This allowed fine detail to be restored, meaning that the decoder layers also contributed directly to the final mask. The kernel depths of the decoder layers were the same as those of the encoder layers, but in opposite order, so that they reversed the encoder steps.

5.2.5 Evaluation

The single channel DCCRN-dir was evaluated for objective quality using both narrowband and wideband PESQ (see Section 4.1.2) to be able to compare it to the original DCCRN from [30], and another winning system of the Interspeech 2020 DNS Challenge [26]. Here the test set of the DNS Challenge was used, to allow for direct comparison.

PESQ wideband and STOI results were then obtained for the performance of DCCRN-dir on the multichannel datasets based on the 'Easy' and 'Challenging' RIR sets (see Section 2.4.2). The measured RIRs in each set were convolved

with random speech samples from 'NB Tale' [52] (see section 2.2.1). Then, noise recorded with the same array as used for RIR measurement was added to obtain SNRs of 0 dB, 5 dB or 10 dB. Here, a variation of typical meeting room noises was used (see Section 2.3.2). As DCCRN-dir is a single SE system, only the first channel of each multichannel test sample was processed, while all other channels were discarded.

Furthermore, DCCRN-dir was evaluated with five different objective intelligibility metrics (i.e.: NCM, CSII, STOI, ESTOI and HASPI) in order to be able to compare the predictions to subjective results. Therefore, in this case, the subjective speech material of Øygarden's Norwegian speech-in-noise test (see Section 2.2.1) was used. The sentences were first made reverberant with a random RIR from the 'Easy' and 'Challenging' RIR sets (see Section 2.4.2), and then corrupted with one type of recorded multichannel noise at SNRs ranging from -36 dB to 10 dB in steps of 2 dB. Here the chosen noise type was keyboard typing, with a climate control system audible in the background (See Section 2.3.2).

The subjective evaluation was conducted with the same noisy reverberant speech data, using Øygarden's Norwegian speech-in-noise test (see Section 4.2.1). This time 50 listeners were recruited, with an intentional inclusive recruitment process that also accepted non-native listeners and listeners with self-reported hearing loss, in addition to native speakers with normal hearing. These participants were then divided into three groups depending on their speech recognition threshold for the noisy baseline test.

Chapter 6

Multichannel Speech Enhancement

Humans can hear direction, mainly because they have two ears. These ears are present already before birth, but spatial hearing is a skill that needs to be learned: it develops slowly during childhood. Children first learn to follow movement, then to discriminate between left and right, and then finally to localize sources and to spatially filter speech from noise [101]. The fact that spatial hearing, and the ability to understand speech in noise, are so closely linked, strongly motivates the use of multiple microphones for speech enhancement.

The multitude of microphone elements in a microphone array, give multichannel recordings. Generally, the microphone elements are all located close together, and record the same sources. However, due to slightly different paths between sound source and element, the recorded signal will also differ slightly. The differences provide information that enable direction of arrival estimation (determining where the source signal is coming from), and beamforming (making the array 'listen' in a specific direction).

This chapter first delves into these two concepts, before they are combined with a DNN for the multichannel speech enhancement systems proposed in [Paper IV].

6.1 Direction of Arrival Estimation

Direction of arrival (DOA) estimation is, as the name suggests, about estimating the direction that a signal is coming from. This is not the exact same as sound/speech source localization (SSL), which also requires knowing how far away the source is located. From an SE perspective, this distance is often of less importance, while the direction can play a crucial role in beamforming (See Section 6.2).

The circular table top microphone array used for the work of this thesis, has a diameter of 8 cm. When the microphone is placed on a table in a meeting room,

speaker distances of more than a meter will be natural. As such, DOA estimation can be formulated as a far-field problem here.

If the source is in the far field, the sound waves will be travelling in parallel by the time they reach the array. Figure 6.1 shows such a wave hitting two microphone elements. The normal to the wavefront makes an angle θ with plane that connects the two elements. The signal received at the second microphone element is, in this case, a time delayed (but could also have been a time advanced) version of the signal reaching the reference sensor [102].

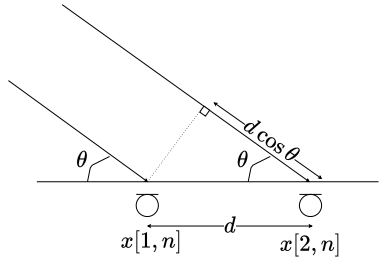


Figure 6.1: Direction of arrival problem in the far field. The wavefront needs more time to reach the second microphone element, as it is further away from the source.

In the figure, the time delay between the elements, the time-difference-of-arrival (TDOA) is equal to $d\cos\theta/c$, where c is the speed of sound. Once the TDOA is known, the direction of arrival is also known.

The figure illustrates this problem along one axis, but it can be expanded to all three axes of space. Given a circular array, where all elements lay in a 'flat' circle (have equal height, so to say), it is impossible to measure TDOAs along the z-axis. Therefore, the tabletop microphone in this thesis, can only be used to estimate the direction in 2D space. The DOA angle is here equal to the azimuth, and the angle of elevation is arbitrarily set to zero: all sources are projected onto the same plane of the microphone.

Finding the azimuth of a source is easy if that source is the only source, and if there is no reverberation (called 'free-field' conditions). In that case, the signals arriving at the different elements would merely be delayed and attenuated copies of each other. Then, it is possible to find the delay for two elements (the TDOA), by determining for which estimated delay $\hat{\tau}^{\mathrm{CC}}$, the cross correlation of the two recordings is maximum.

$$r^{\mathrm{CC}}(p) = E\left[x_1(t)\,x_2(t+\tau)\right], \tag{6.1}$$

$$\hat{\tau}^{\mathrm{CC}} = \arg\max_{\tau} r^{\mathrm{CC}}(\tau). \tag{6.2}$$

This still works if the recorded signal contains noise uncorrelated to the noise at the other sensor (like self-noise of the microphone element), or if the noise is really low and barely contributes to the output of the cross correlation.

However, in more realistic conditions, the reverberance and presence of correlated noise (coming from other directions) complicates the problem.

Generalized Cross Correlation

One of the most popular approaches to TDOA estimation, is the generalized cross correlation (GCC) algorithm. It is very similar to the direct cross correlation method already described, but the signals are first filtered, with frequency dependent filter weights $\vartheta(f)$. The cross correlation coefficient r then becomes:

$$r^{\text{GCC}}(\tau) = \int_{-\infty}^{+\infty} \vartheta(f) E\left[X_1(f) X_2^*(f)\right] e^{j2\pi f \tau}\, df, \tag{6.3}$$

and the estimated delay $\hat{\tau}^{\text{GCC}}$:

$$\hat{\tau}^{\text{GCC}} = \arg\max_{\tau} r^{\text{GCC}}(\tau). \tag{6.4}$$

Here the filtering function $\vartheta(f)$ can be chosen to best fit the application.

Generalized Cross Correlation with Phase Transform

One popular filtering function is to discard the amplitude, leaving only the phase as a contributor to the cross correlation coefficient.

$$\vartheta(f) = \frac{1}{\left|E\left[X_1(f) X_2^*(f)\right]\right|} \tag{6.5}$$

In this case, the cross correlation coefficient becomes

$$r^{\text{GCC-PhaT}}(\tau) = \int_{-\infty}^{+\infty} e^{j2\pi f\left(\tau - \hat{\tau}^{\text{GCC}}\right)}\, df, \tag{6.6}$$

which is infinitely high when $\hat{\tau}^{\text{GCC}} = \tau$ and zero everywhere else.

This is called the phase transform (PhaT) method, or GCC-PhaT. It is a computationally efficient method that performs reasonably well in moderately noisy environments [102]. It is suboptimal in non-reverberant conditions, but it relatively robust to reverberation.

With GCC-PhaT, a delay for each pair of microphones is obtained. For most arrays, there is a redundancy in microphone signals, giving a redundancy in TDOA estimates. These can be combined with the root-mean-squared-error (RMSE) minimization technique, to improve the accuracy of the individual TDOA estimates.

Steered Response Power with Phase Transform

The Steered Response Power with Phase Transform (SRT-PhaT) [103] technique, on the other hand, combines the microphone signals instead of the TDOA estimates, in order to improve the accuracy of the final TDOA estimates.

It does this through beamforming, a concept that is further explained in Section 6.2. Specifically, it relies on the simplest form of beamforming: the delay-and-sum beamformer. With this kind of beamformer, the different signals are time-aligned and summed, so that the parts of the signal that come from the same direction add constructively, while uncorrelated noise, and signals from other directions do not.

SRP-PhaT repeats this beamforming action over all the locations in a fine coarse search grid. Its output (a collection of beamformed signals for each location on the grid) is called a 'steered response', referring to the fact that the beamformer is being steered over the region.

SRP-PhaT also, as the name suggests, includes the PhaT filtering explained in Section 6.1. This filter is specifically useful for TDOA estimation. The individual beamformer outputs of the steered response are not meant to be listened to, but instead, expected to give a high power output at the source location.

DiBiase already showed in his original thesis where he introduced the SRP-PhaT method, that the power response of a PhaT filtered delay-and-sum beamformer (P) for a pair of microphones, can also be written in terms of the GCC-PhaT coefficient ($r^{\text{GCC-PhaT}}$) of these microphones [103].

$$P = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} r^{\text{GCC-PhaT}}, \tag{6.7}$$

where m_1 and m_2 are the two elements of the microphone pair, M is the total number of microphones of the array and $r^{\text{GCC-PhaT}}$ is calculated for the τ corresponding to the delay expected for the grid position for which the power response P is being calculated.

SRP-PhaT is a very popular DOA method, that has been shown to be robust in reverberation. The biggest issue with SRP-PhaT is not its accuracy performance, but the fact that it is computationally heavy.

DNN-based Direction of Arrival

As part of the work for this thesis, a DNN-based DOA estimation system was proposed in [Paper III]. Here, the problem of DOA estimation was formulated as a supervised problem (See Figure 6.2).

The training data was obtained by first simulating RIRs (See Section 2.4.1). These RIRs were convolved with clean speech, to obtain reverberant speech. For each of the microphone pairs in the simulated microphone array, GCC vectors were obtained from their respective reverberant speech signals. The input to the trainable DOA system was then a flattened matrix of truncated GCC vectors, one for each microphone pair. Figure 6.3 shows examples of the final input features, when only the method of RIR simulation is changed.

Each input, was paired with a target to be estimated: the azimuth of the source as used during the simulation of RIRs. Two different regression formulations and two different classification formulations were used.

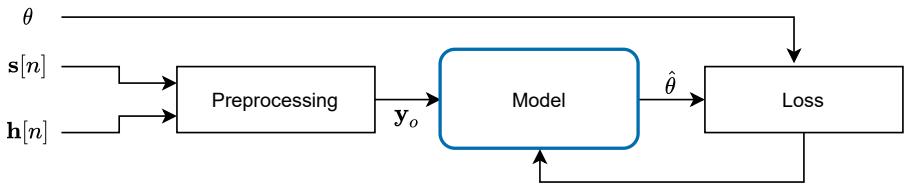


Figure 6.2: Overview of the DNN-based DOA estimation system. Here θ is the target azimuth, $\hat{\theta}$ the estimated azimuth, $\mathbf{s}[n]$ the clean speech signal, $\mathbf{h}[n]$ the RIR, and $\mathbf{y}_0$ the GCC vector model input.

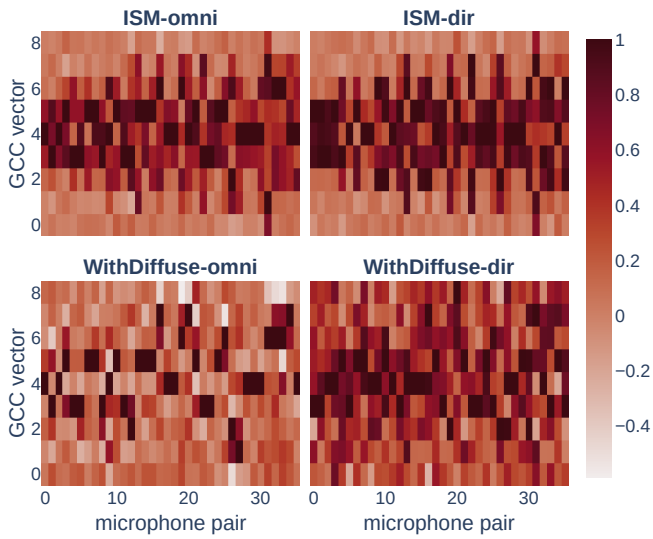


Figure 6.3: Examples of the final GCC vector input features. Here the elements of the GCC vector on the y-axis represent different values of τ, while the colour intensity indicates the value of $r^{\text{GCC-PhaT}}(\tau)$ for each microphone pair. Figure taken from [Paper III] (©2021 IEEE).

For regression, the formulations differed by the exact type of loss functions used. The two different loss functions were based on the general mean squared error (MSE) and mean absolute error (MAE) loss functions, respectively, but then defined such that the calculation always obtained the minimal error between two angles, which can either be clockwise or anticlockwise. Additionally, two classification formulations of the same problem were tested. Here the target angles were first binned into different classes (either 1° bins or 5 ° bins). These 'classification' models were trained with the categorial crossentropy loss.

A simple feed forward DNN with fully connected layers was used for this study, since the aim was not to find the best DOA estimation model. Instead, the goal was to find how different sets of RIRs (simulated using a different method) changed the final performance.

Therefore, all systems were tested with the two sets of reverberant speech. These sets were based on the 'Easy' and 'Challenging' RIR sets (see Section 2.4.2) and clean speech from 'NB Tale' [52] (see section 2.2.1).

Results of the performance evaluation of this DOA estimation system are presented in [Paper III]. Here it is the variance of the error from the true direction that indicates system performance (assuming zero bias error). Therefore the Brown-Forsythe statistical test [104] was applied. This test evaluates whether the variance of a pair of distributions is statistically significant without a strong assumption of normality.

6.2 Beamforming

The concept of beamforming has already been briefly discussed in Section 6.1, because of its role in the SRP-PhaT algorithm. While it can be used for direction of arrival estimation, it is better known for recovering a signal from noisy reverberant recordings.

The technique of beamforming multichannel signals is not specific to speech recordings. Instead, it has a longer history in the field of radio frequency processing. However beamforming speech recordings is quite different, as speech is a highly reverberant wideband signal, where the noise often has the same spectral characteristics as the desired signal, and the human ear is extremely sensitive over a wide dynamic range [102]. All of these factor have an effect on the process of beamforming multichannel speech signals.

The most intuitive beamforming algorithm is the delay-and-sum beamformer. If the source location is known (or has been estimated), the signals of the different microphones in the array are delayed by their respective TDOAs (see Section 6.1). Then the signals are simply added together, where constructive interference occurs for the signals from the right direction. The delay-and-sum beamformer is a so-called 'fixed' beamformer, where the beam pattern (how much each frequency is amplified/attenuated in what direction) does not depend on the recorded microphone signals.

For speech enhancement, the minimum variance distortionless response (MVDR), usually has higher performance, in the sense that there is less distortion of the estimated source signal. It is an adaptive beamformer, that adapts its beampattern to the signal it processes.

The 'distortionless' aspect is extremely important from a human speech perception perspective. The MVDR algorithm implements the criterion that the desired signal is not distorted. Bound by this constraint, the algorithm then attempts to minimize the output power of the beamformer, which is the same as maximizing the output SNR.

When using the MVDR beamformer, the enhanced output Fourier domain speech signal ($\hat{S}[k]$) is a weighted sum of the noisy input $X[k]$:

$$\hat{S}[k] = w_{\text{MVDR}}^{H} X[k], \tag{6.8}$$

where the MVDR beamformer itself is then defined as the following set of weights (See [105] for a complete derivation):

$$w_{\text{MVDR}}^{H} = \frac{a^{H} V[k]^{-1}}{a^{H} V[k]^{-1} a}, \tag{6.9}$$

where $V[k]$ is the noise signal, and a the steering vector, which is known when the TDOAs are known.

The disadvantage of the MVDR beamformer is its reliance on the noise signal, which is unknown. One solution to this, which works well for stationary noise, is to assume the noise signal is largely constant and can therefore be obtained from earlier speechless frames. This however, does not work for the typical meeting noise recorded during for this thesis, many of which are transient in nature (clinking of coffee cup, closing of door, etc.).

Another option is to estimate the noise signal using a DNN-based network. This is the approach taken, for example, by Heymann *et al.* in [39] and Erdogan *et al.* in [40]. Heymann *et al.* additionally obtain the steering vector by taking the principal component of the power spectral density matrix of the estimated speech [39], while Erdogan *et al.* rely on a formulation of the MVDR that does not explicitly contain the steering vector [40]. This means that for these approaches, the beamformer's behaviour is inherently dependent on the model's ability to estimate the noise (and speech) signal(s), and so is its 'distortionless' behaviour.

Closely related to the MVDR beamformer, is a beamformer called the minimum power distortionless response (MPDR). The weights of the MPDR beamformer are obtained as follows:

$$w_{\text{MPDR}}^{H} = \frac{a^{H} X[k]^{-1}}{a^{H} X[k]^{-1} a} \tag{6.10}$$

Here, the weights depend on the noisy reverberant input $X[k]$ (where the capitalization indicates the frequency domain) so that the steering vector is the only unknown in the equation. Given a correct steering vector, the MPDR weights are equal to the MVDR weights, and the performance of the two beamformers will

be equal [105]. The downside is that the MPDR beamformer performs worse for inaccurate steering vectors.

However, the MPDR also opens up the possibility of a multichannel approach that is different from the neural beamformers proposed in [39, 40]. With a direction guided MPDR, the speech enhancement model, DOA system, and beamformer are completely separate items of the SE pipeline, where the 'disortionless' behaviour of the beamformer depends only on the accuracy of the steering vector, not on the model's ability to separate speech and noise.

6.3 Multichannel DCCRN

Figure 6.4 shows a process diagram of the proposed multichannel speech enhancement system. Different variants of this network were tested in [Paper IV] and [Paper V].

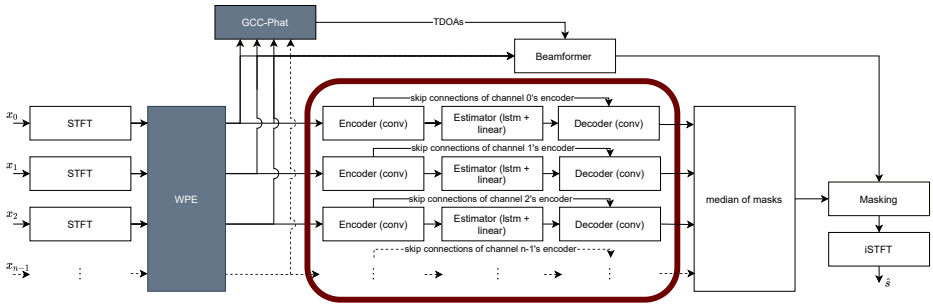


Figure 6.4: The multichannel DCCRN-based SE system. Here x is the multichannel noisy and reverberant input speech signal, and $\hat{s}$ the estimated clean speech. Figure taken from [Paper IV] (©2021 IEEE).

The multichannel noisy reverberant speech $\mathbf{x}$ is transferred into complex STFT coefficients, which are passed into a weighted prediction error (WPE) block to reduce the amount of reverberation. Here the reverberation tail is estimated and subtracted from the noisy reverberant speech using a maximum likelihood approach [106]. To prevent further delay in the processed signal, only one iteration is used. WPE dereverberation is included as a separate step, as, based on results from pilot studies, the DCCRN topology did not seem to be able to reduce reverberation: performance on reverberant signals got worse when clean instead of reverberant targets were used.

Following dereverberation, the output is passed on to the model, the MPDR beamformer (See Section 6.2), and (only when testing with estimated TDOAs) to a GCC-Phat block (See section 6.1).

The model estimates a separate complex ratio mask for each channel, using one DCCRN per channel (See Section 5.2), where all DCCRNs have the same weights. All masks are subsequently merged into a single mask, by taking the median over each TF units for all masks.

This mask is then applied to the beamformer's output, and the signal is transformed to the time domain with the inverse STFT operation.

6.3.1 Variants

The following multichannel DCCRN-based SE systems were proposed and evaluated in [Paper IV] and/or [Paper V].

- **MPDR (estimated TDOAs) + DCCRN-dir**: Multichannel noisy and reverberant speech that has been passed through the complete multichannel system shown in 6.4. Here the DCCRN-dir model was trained separately, and TDOAs were estimated from the dereverberated output of the WPE blocks using GCC-Phat.
- **MPDR (oracle TDOAs) + DCCRN-dir**: Multichannel noisy and reverberant speech that has been passed through the complete multichannel system shown in 6.4. Here the DCCRN-dir model was trained separately, and oracle TDOAs were used.
- **Jointly trained system (estimated TDOAs)**: Multichannel noisy and reverberant speech that has been passed through the complete multichannel system shown in 6.4. Here the DCCRN-dir model was trained jointly with the MPDR, and TDOAs were estimated from the dereverberated output of the WPE blocks using GCC-Phat.
- **Jointly trained system (oracle TDOAs)**: Multichannel noisy and reverberant speech that has been passed through the complete multichannel system shown in 6.4. Here the DCCRN-dir model was trained jointly with the MPDR, and oracle TDOAs were used.

Note also that several other systems where explored, but not published, as they showed lower objective performance. These variants included i) using the enhanced signals as input to the beamformer, ii) providing the model with 2-channel input (one channel from the center array elements and a second from the beamformer output) as two separate feature spaces of the first convolutional layer of the DCCRN, and iii) a multichannel DCCRN that directly obtained all microphone channels as separate feature spaces of the first convolutional layer of the DCCRN. Furthermore, it was discovered that it is important to use the same mask for all channels, but that it did not matter significantly whether this mask was obtained through the median (as proposed) operator, the mean operator, or simply by obtaining it from the first channel.

6.3.2 The data

For the combined 'MPDR + Single channel DCCRN' systems (either with estimated or oracle TDOAs), the DCCRN was trained with the datasets described in Section 5.2.1.

For the jointly trained systems, the training data was changed only in the following manners:

- Noise was also made multichannel and reverberant, using omnidirectional RIRs simulated with the ISM-omni method (See Section 2.4.1).
- The target was set to clean speech (instead of reverberant speech, as done for the training of the single channel DCCRN)

6.3.3 Training

For the combined 'MPDR + Single channel DCCRN' systems (either with estimated or oracle TDOAs), the DCCRN was trained as described in Section 5.2.2. For the jointly trained systems, both the WPE and beamformer blocks were included in the training process. During training, the beamformer was supplied with oracle TDOAs.

As with the single channel DCCRN system, the pipeline of the multichannel systems starts with a preprocessing/feature STFT extraction block. The STFT coefficient of all nine microphone array channels are then passed on to the WPE block for intial dereverberation. The concatenation of real and imaginary STFT coefficients of the WPE output was then passed to a multitude of DCCRN networks with shared weights, which all estimated a mask for their respective channels. A single mask was then obtained by using the median operator on all masks. The final mask was applied to the output of the beamformer, which had beamformed the noisy input.

As with the single channel DCCRN, the loss is obtained in the time domain with the SI-SNR loss.

6.3.4 Usage

During usage, TDOAs can either be estimated using GCC-PhaT, or set to the oracle TDOAs.

The jointly trained networks do not require any other additional steps as the time-domain signal is produced in the same manner as during the training process. For the combined systems, the separately trained single channel DCCRN is set to produce the mask as output. Then operation is as described in Section 6.3.3.

6.3.5 Evaluation

The multichannel SE systems listed in 6.3.1 were evaluated in [Paper IV] and/or [Paper V]. Here [Paper IV] presents only objective results, while [Paper V] includes both objective and subjective results.

For comparison, the following baseline systems were defined:

- **Noisy**: Single channel noisy and reverberant speech.
- **MPDR (estimated TDOAs)**: The MPDR beamformer on its own, with or without WPE dereverberation (as indicated separately), where TDOAs were estimated using GCC-Phat on the noisy reverberant input.

- **MPDR (oracle TDOAs)**: The MPDR beamformer on its own, with or without WPE dereverberation (as indicated separately), where TDOAs were oracle TDOAs.
- **ConferencingSpeech 2021 baseline**: The baseline system for the INTERSPEECH 2021 ConferencingSpeech challenge 2021 [107].
- **GEV (oracle IBM mask) with BAN**: A multichannel neural beamformer system proposed by Heymann *et al.* [39].
- **DCCRN-dir**: The single channel DCCRN described in Section 5.2, based on the DCCRN proposed by Hu *et al.*, but trained with directive RIRs, with or without WPE dereverberation (as indicated separately).

Table 6.1 shows which SE/baseline variants were evaluated for this thesis, and in what manner. Here WPE stands for weighted prediction error, which is an dereverberation technique proposed in [106].

Table 6.1: verview of the differenct processing conditions and evaluation methods used. Objective results are reported in [Paper IV] and [Paper V], subjective only in [Paper V].

		Quality		Intelligibility	
SNR [dB]	**WPE**	**Obj.**	**Subj.**	**Obj.**	**Subj.**
No enhancement	No	✓	✗	✓	✓
	Yes	✓	✗	✓	✗
ConferencingSpeech	No	✓	✗	✓	✗
2021 baseline [107]	Yes	✓	✗	✓	✗
DCCRN-dir	No	✓	✗	✓	✗
	Yes	✓	✗	✓	✓
GEV (oracle IBM mask) with	No	✓	✗	✓	✗
BAN, by Heymann *et al.* [39]	Yes	✓	✗	✓	✗
MPDR (estimated TDOAs)	No	✗	✗	✗	✗
	Yes	✗	✗	✓	✓
MPDR (oracle TDOAs)	No	✗	✗	✗	✗
	Yes	✗	✗	✓	✓
MPDR (estimated TDOAs)	No	✗	✗	✗	✗
+ DCCRN-dir	Yes	✗	✗	✓	✓
MPDR (oracle TDOAs)	No	✓	✗	✓	✗
+ DCCRN-dir	Yes	✓	✗	✓	✓
Jointly trained system	No	✓	✗	✓	✗
(estimated TDOA)	Yes	✓	✗	✓	✗
Jointly trained system	No	✓	✗	✓	✗
(oracle TDOAs)	Yes	✓	✗	✓	✗

For objective evaluation, PESQ wideband and STOI results were then obtained to estimate the performance of the systems on the multichannel datasets based

on the 'Easy' and 'Challenging' RIR sets (see Section 2.4.2). The measured RIRs in each set were convolved with random speech samples from 'NB Tale' [52] (see section 2.2.1). Then, noise recorded with the same array as used for RIR measurement was added to obtain SNRs of 0 dB, 5 dB or 10 dB. Here, a variation of typical meeting room noises was used (see Section 2.3.2).

Furthermore, the systems that were evaluated subjectively, were also additionally evaluated objectively with five different objective intelligibility metrics (i.e.: NCM, CSII, STOI, ESTOI and HASPI) in order to be able to compare the predictions to subjective results. Therefore, the subjective speech material of Øygarden's Norwegian speech-in-noise test (see Section 2.2.1) was used. The sentences were first made reverberant with a random RIR from the 'Easy' and 'Challenging' RIR sets (see Section 2.4.2), and then corrupted with one type of recorded multichannel noise at SNRs ranging from -36 dB to 10 dB in steps of 2 dB. Here the chosen noise type was keyboard typing, with a climate control system audible in the background (see Section 2.3.2).

The subjective evaluation was then conducted with the same noisy reverberant speech data. For this purpose, Øygarden's Norwegian speech-in-noise test (see Section 4.2.1) was used. A total of 50 listeners were recruited, with an intentional inclusive recruitment process that also accepted non-native listeners and listeners with self-reported hearing loss, in addition to native speakers with normal hearing. These participants were then divided into three groups, depending on their speech recognition threshold for the noisy baseline test.

The jointly trained systems (where the DCCRN blocks were trained together with the beamformer and WPE blocks) had the highest objective performance in [Paper IV]. However, the difference from the combined systems where the DCCRN was trained separately, was minimal and not statistically significant at the lower SNRs relevant for speech intelligibility. Additionally, a small subjective pilot indicated the jointly trained system would do worse in the subjective test. Therefore, the combined system was chosen for further subjective evaluation.

Chapter 7

Results

7.1 Single Channel Speech Enhancement

7.1.1 Fully connected log magnitude estimator

Speech quality results

The objective and subjective quality performance of two variants of the fully connected log magnitude estimator (FCLME) network (see Section 5.1), were evaluated in [Paper II]. The variants studied are Model 1 and Model 2, where the latter is mostly equal to the first, but it attempts to reduce noise instead of remove it.

Figures 7.1 and 7.2 show the POLQA results of the two models and the noisy baseline condition for traffic noise and babble noise, respectively.

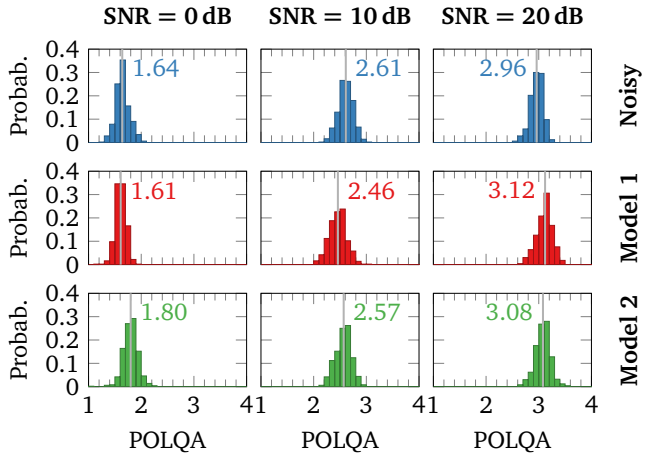


Figure 7.1: Objective quality results of the FCLME for traffic noise. The histograms show the relative probability of POLQA scores for traffic noise. The vertical gray lines and numbers represent median values. Figure taken from [Paper II] (©2019 IEEE).

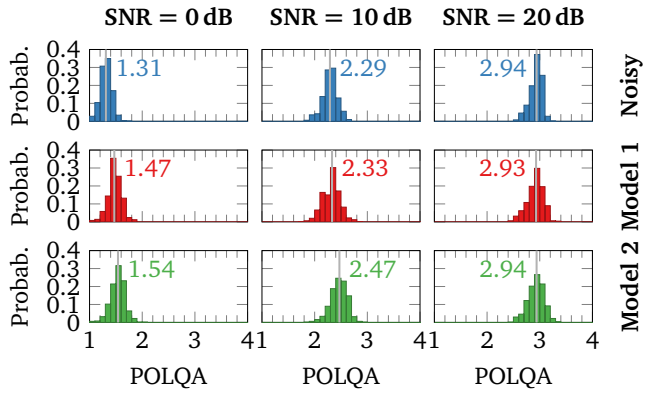


Figure 7.2: Objective quality results of the FCLME for babble noise. The histograms show the relative probability of POLQA scores for traffic noise. The vertical gray lines and numbers represent median values. Figure taken from [Paper II] (©2019 IEEE).

For traffic noise, POLQA predicts changes without clear direction. For babble noise, Model 2 always outperforms Model 1, which in its turn gets higher scores than the noisy baseline condition data (apart from at SNR = 20 dB, where performance is nearly equal for all processing conditions).

Figures 7.3 and 7.4 show the subjective results from the ITU-T P.835 evaluation on the same dataset.

Both Models 1 and 2 reduce noise, as evident from significantly higher scores for the Noise evaluation of ITU-T P.835 for all noise conditions except for babble noise at an SNR of 20 dB, where the change was insignificant for the two models.

However, both models also lower the quality of speech for all noise conditions, even if Model 2 does not distort the speech as much as Model 1. This is evident from the reduction in subjective scores obtained for the 'Speech' and the 'Overall' quality categories of the ITU-T P.835 evaluation.

Model 1 significantly degrades the overall quality of speech for all noise conditions, apart from at an SNR of 0 dB; here it does not affect the quality at all. Model 2 on the other hand improves the overall quality of speech for babble noise at an SNR of 0 dB, while leaving the overall quality unaffected for all other noise conditions.

There appears to be little correlation between the subjective scores of the overall quality as measured with ITU-T P.835 and the predictions by POLQA. In approximately half of the noise situations, POLQA's predictions are completely off: It predicts a relatively large improvement, while subjective testing shows a significant degradation, or it predicts no change, while subjective testing either shows significant degradation or improvement.

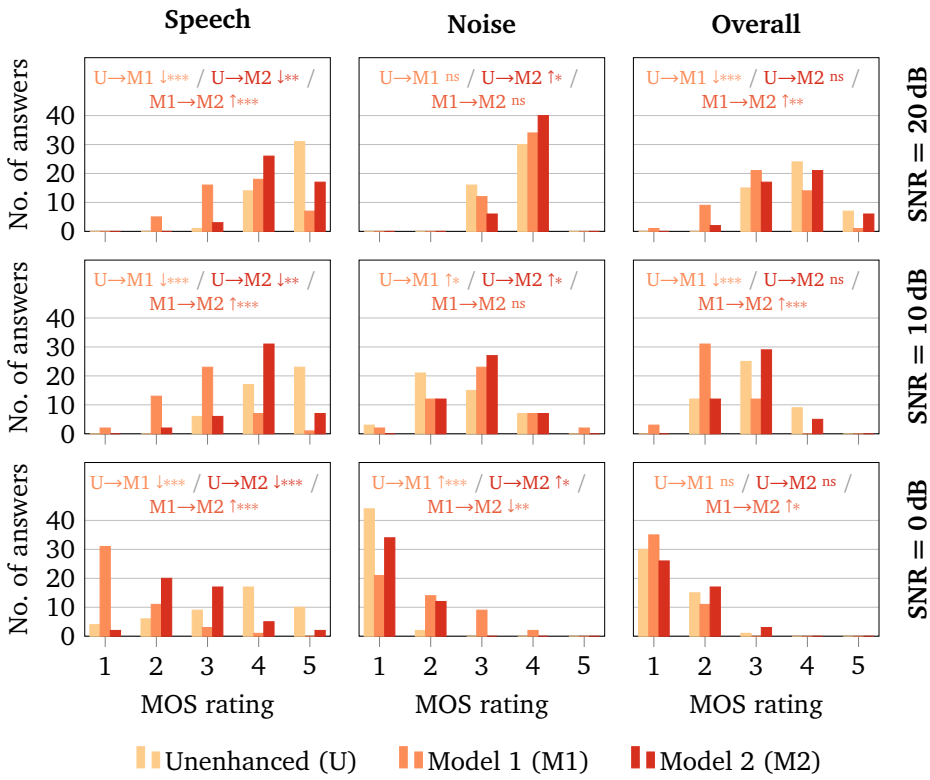


Figure 7.3: Subjective quality results of the FCLME for traffic noise. Histograms of ITU-T P.835 MOS ratings for the different SE models and SNRs. Three asterisks (***) indicate $p < .001$, two asterisks (**) indicate $p < .01$, one asterisk (*) indicates $p < .05$, and *ns* means 'not significant'. The arrows beside the asterisk(s) indicate the direction of change. Figure adapted from [Paper II] (©2019 IEEE).

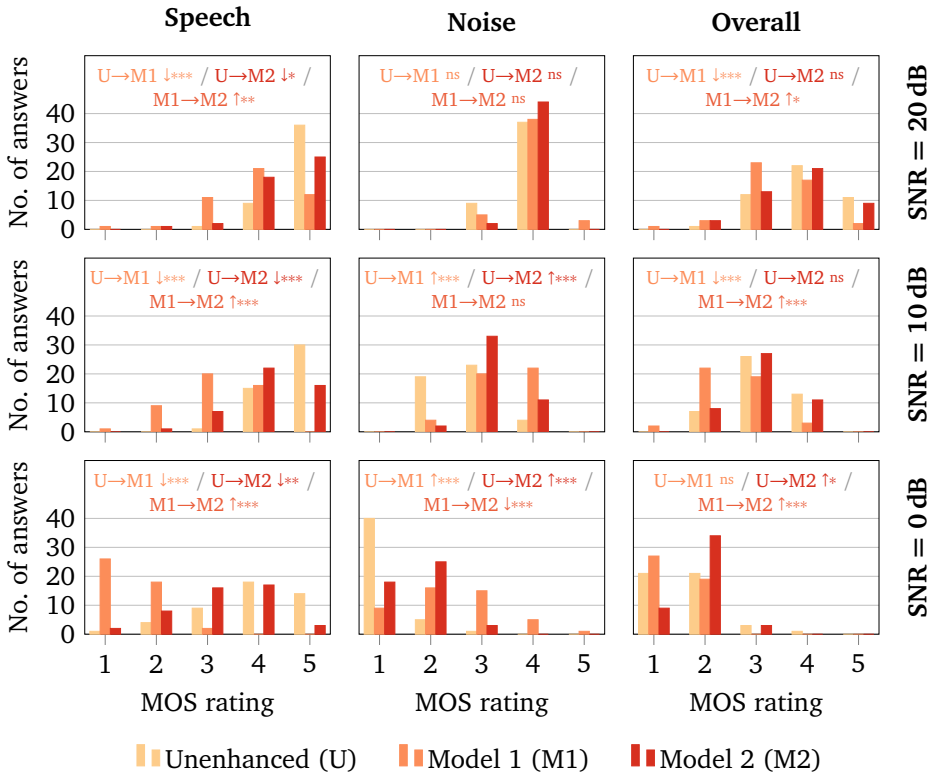


Figure 7.4: Subjective quality results of the FCLME for babble noise. Histograms of ITU-T P.835 MOS ratings for the different SE models and SNRs. Three asterisks (***) indicate $p < .001$, two asterisks (**) indicate $p < .01$, one asterisk (*) indicates $p < .05$, and *ns* means 'not significant'. The arrows beside the asterisk(s) indicate the direction of change. Figure adapted from [Paper II] (©2019 IEEE).

Speech intelligibility results

The objective and subjective intelligibility performance of all three variants of the fully connected log magnitude estimator (FCLME) network (see Section 5.1), were studied in [Paper I] and [Paper II].

Figure 7.5 shows the STOI scores obtained for the different variants. At an SNR of -6 dB, all models improve STOI scores by 8-12 %, with the exact improvement depending on the type of noise and, to a lesser degree, on the model variant. For really low SNRs (where intelligibility can be expected to be close to zero), the models that attempt to remove noise (Models 1 and 3) obtain scores below the noisy baseline condition. The model that attempts to reduce the noise (Model 2), instead obtains approximately equal STOI scores as the noisy reference at these SNRs. For the SNR range relevant to speech intelligibility, STOI predicts increased intelligibility for all model variants and both noise types.

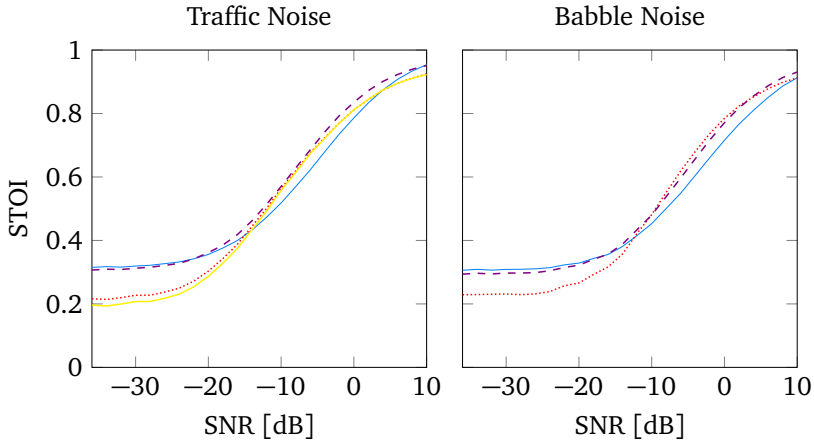


Figure 7.5: Objective intelligibility results of the FCLME for traffic noise and babble noise. Model 3 was only tested with traffic noise. Legend: Noisy, Model 1, Model 2 (like Model 1, but with reducing noise), and Model 3 (like Model 1, but with global variance normalization (GVN)).

Figure 7.6 shows the speech recognition threshold (SRT) scores obtained from the subjective evaluation of the same models under the same noise conditions. Here, all models raise the SRT, which means that intelligibility is actually reduced. The results shown in Figure 7.6 are obtained for two subgroups of participants. However, a Wilcoxon rank sum test shows no significant difference between the SRT results for the noisy reference for the two subgroups, which means that the groups performed equally well on the noisy condition.

The reduction in intelligibility of the processed speech is, on the other hand, statistically significant for all models ($p < 0.05$) according to the Wilcoxon signed ranks test. All models degrade the intelligibility of the speech signal.

The subjective results are therefore in the opposite direction as the predictions by STOI. Furthermore, STOI predicted no significant difference in the performance

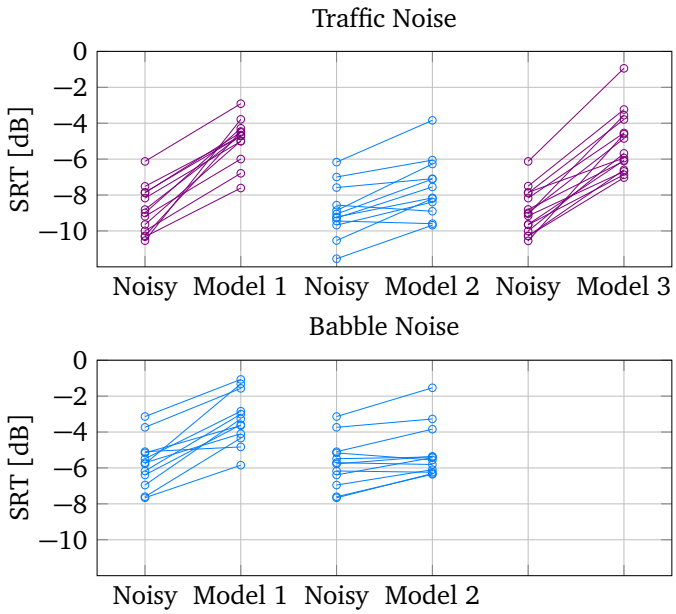


Figure 7.6: Subjective intelligibility results of the FCLME for traffic noise and babble noise. The connected lines represent results from each of the test subjects from noisy clips to the model indicated on the horizontal axis. Results were obtained from two separate studies with different listeners: — [Paper I] and — [Paper II]

of Model 1 and Model 2 in the SNR range relevant to speech intelligibility. Instead, Model 2 performs significantly better than Model 1 for both noise types. The estimated improvement of the SRT from M1 to M2 is 3.0 dB for traffic noise, and 1.9 dB for babble noise. STOI did predict correctly that adding GVN processing, would give no significant improvement (the performance of Model 3 and Model 1 is similar).

7.1.2 Complex convolutional recurrent mask estimator

Speech quality results

Performance with respect to speech quality of the single channel complex convolutional recurrent mask estimator (the DCCRN) was only evaluated objectively ([Paper IV]).

First, PESQ performance was compared to values of competitive systems from the literature (See Table 7.1). Here the difference between the DCCRN-E [30] and this thesis' DCCRN-dir is limited to using a different training set (with directional sources for RIR simulation), and general hyperparameter tuning.

PESQ performance on the non-reverberant DNS Challenge test set is similar to the performance of the reference systems. however, DCCRN-dir has superior

Table 7.1: Objective quality results of DCCRN-dir for the DNS Challenge 2020 dataset. Narrowband and wideband PESQ results, where the reverberant speech signal is used as reference.[1] Result without partial dereverberation, for unbiased comparison. Table taken from [Paper IV] (©2021 IEEE).

	PESQ nb		**PESQ wb**	
	No reverb	**Reverb**	**No reverb**	**Reverb**
Noisy	2.16	2.52	1.58	1.82
PoCoNet [94]	-	-	2.75	2.83[1]
DCCRN-E [30]	3.27	3.08	-	-
DCCRN-dir	3.28	3.44	2.76	2.94

performance on the reverberant set, when compared to the original DCCRN-E, and also possibly when compared to PoCoNet [94], depending on the standard deviation of their test scores (not published).

The same system was also tested on the two different datasets aquired with RIR measurements (See Section 2.4.2 and [Paper III]). Here 'Easy', indicates that the speaker is facing the array, while 'Challenging' means that the speaker is facing away at a 90° angle. PESQ results for these datasets are shown in Table 7.2.

Table 7.2: Objective quality results of DCCRN-dir for the 'Easy' and 'Challenging' datasets. Wideband PESQ results, where the clean speech signal is used as reference.

	WPE	**Easy**			**Challenging**		
SNR [dB]		**0**	**5**	**10**	**0**	**5**	**10**
No enhancement	No	1.25	1.33	1.39	1.22	1.29	1.35
	Yes	1.33	1.44	1.56	1.27	1.36	1.46
DCCRN-dir	No	1.46	1.49	1.51	1.41	1.44	1.46
	Yes	1.64	1.71	1.76	1.55	1.61	1.66

Independent of the test set used, PESQ predicts that all enhancement systems improve the quality of speech and that all enhancement systems also benefit from the WPE preprocessing step.

Intelligibility

Performance with respect to intelligibility of the single channel complex convolutional recurrent mask estimator (the DCCRN) was evaluated objectively (in [Paper IV] and [Paper V], and subjectively in [Paper IV]).

Table 7.3 shows the STOI performance on the two different datasets aquired with RIR measurements (See Section 2.4.2 and [Paper III]). STOI predicts that all enhancement systems improve intelligibility and also predicts that all enhancement systems benefit from the WPE dereverberation block.

Table 7.3: Objective intelligibility results of DCCRN-dir for the 'Easy' and 'Challenging' datasets. STOI score results, where the clean speech signal is used as reference.

	WPE	Easy			Challenging		
SNR [dB]		**0**	**5**	**10**	**0**	**5**	**10**
No enhancement	No	0.69	0.72	0.74	0.60	0.62	0.63
	Yes	0.72	0.76	0.78	0.18	0.66	0.68
DCCRN-dir	No	0.73	0.75	0.75	0.64	0.64	0.65
	Yes	0.77	0.78	0.79	0.68	0.69	0.70

Table 7.4 shows the predictions of change in speech recognition threshold (SRT) for five different measures of objective intelligibility, together with the SRT changes obtained for different groups of listeners through subjective testing.

Table 7.4: Objective intelligibility results of DCCRN-dir for the speech-in-noise test dataset. All statistically significant changes in SRT (predicted or measured) are marked with an asterisk (*).

	NCM	**CSII**	**STOI**	**ESTOI**	**HASPI**
Predicted	-7.4*	-4.5*	-4.0*	-3.1*	-3.8*
Low SRT group			2.5*		
Medium SRT group			0.8		
High SRT group			-0.2		

All measures predict statistically significant negative changes (improved intelligibility), while subjective evaluation either shows reduction in intelligibility, or no significant change.

7.2 Direction of Arrival Estimation

The direction of arrival (DOA) system described in Section 6.1 was proposed and evaluated in [Paper III]. Tables 7.5 and 7.6 show the mean absolute error (MAE) results for the different test sets, variants of the model, and when trained with different datasets based on the RIR simulation techniques described in Section 2.4.1.

For the 'Easy' test set, where the speaker was always facing the array, performance is high for all formulations and training sets.

The RIR simulation method matters more for the 'Challenging' dataset, where the speakers all looked away from the array at a 90° angle. For this dataset, both the ISM-dir and the WithDiffuse-dir based datasets, led to models that had significantly higher performance than the SRP-PhaT (see Section 6.1) baseline system.

The overall performance difference between ISM-dir and WithDiffuse-dir was insignificant, meaning that adding a diffuse field has no significant effect for the testing conditions.

Table 7.5: Mean absolute error for the 'Easy' test set, where speakers face directly towards the array. Table taken from [Paper III] (©2021 IEEE).

	Regression		**Classification**	
	$MSE_{\measuredangle}$	$MAE_{\measuredangle}$	1° bins	5° bins
SRP-Phat	1.5°			
ISM-omni	2.2°	2.1°	1.4°	1.3°
ISM-dir	3.0°	2.1°	1.5°	1.5°
WithDiffuse-omni	2.8°	1.1°	1.3°	1.4°
WithDiffuse-dir	3.8°	1.4°	1.1°	**0.9**°

Table 7.6: Mean absolute error for the 'Challenging' test set, where speakers face 90° away from the array. Table taken from [Paper III] (©2021 IEEE).

	Regression		**Classification**	
	$MSE_{\measuredangle}$	$MAE_{\measuredangle}$	1° bins	5° bins
SRP-Phat	16.5°			
ISM-omni	18.2°	18.2°	19.1°	18.8°
ISM-dir	12.7°	11.5°	8.9°	**8.1**°
WithDiffuse-omni	19.7°	19.6°	18.6°	17.9°
WithDiffuse-dir	13.0°	10.5°	9.9°	10.1°

The effect of changing the simulation method gives similar performance changes across the different model formulations, which provides evidence that the obtained results are indeed due to the different training sets, and not locked to the formulation of the problem, or caused by hyperparameter tuning.

For the 'Challenging' dataset, the model trained on RIRs simulated for directive sources achieved up to 51 % lower mean absolute error than the industry standard SRP-PhaT method, while the equivalent model trained with standard image source method RIRs from omni-directional sources performed worse than the SRP-PhaT baseline method.

7.3 Multichannel Speech Enhancement

Quality

Performance with respect to speech quality of the multichannel complex convolutional recurrent mask estimator (the DCCRN) was only evaluated objectively (in [Paper IV]). Table 7.7 shows the wideband PESQ results of the three variants of the multichannel DCCRN, together with state-of-the-art reference systems and the noisy unprocessed condition.

PESQ indicates that all speech enhancement systems statistically significantly (independent two-sample t-test, $p < 0.05$) benefit from the WPE preprocessing step. Furthermore, the independent two-sample t-test shows that all proposed

Table 7.7: Objective quality results of the multichannel DCCRN for the 'Easy' and 'Challenging' datasets. Wideband PESQ results, with clean signal used as reference. Best scores per SNR are shown in bold, where multiple highlighted values in the same columns indicate that the difference was not statistically significant. Table taken from [Paper III] (©2021 IEEE).

		Easy			Challenging		
SNR [dB]	**WPE**	**0**	**5**	**10**	**0**	**5**	**10**
No enhancement	No	1.25	1.33	1.39	1.22	1.29	1.35
	Yes	1.33	1.44	1.56	1.27	1.36	1.46
ConferencingSpeech	No	1.33	1.36	1.48	1.27	1.31	1.41
2021 baseline [107]	Yes	1.40	1.46	1.63	1.33	1.39	1.52
DCCRN-dir	No	1.46	1.49	1.51	1.41	1.44	1.46
	Yes	1.64	1.71	1.76	1.55	1.61	1.66
GEV (oracle IBM mask) with	No	1.48	1.59	1.60	1.41	1.46	1.52
BAN, by Heymann *et al.* [39]	Yes	1.58	1.75	1.80	1.49	1.58	1.67
MPDR (oracle TDOAs)	No	1.68	1.73	1.76	1.54	1.59	1.62
+ DCCRN-dir	Yes	**1.89**	**1.98**	**2.04**	**1.71**	1.79	1.85
Jointly trained system	No	1.68	1.86	1.88	1.61	1.73	1.78
(oracle TDOA)	Yes	**1.80**	**2.02**	**2.06**	**1.74**	**1.89**	**1.94**
Jointly trained system	No	1.60	1.80	1.85	1.50	1.62	1.69
(estimated TDOAs)	Yes	1.74	**1.95**	**2.04**	1.63	1.79	**1.88**

systems have statistically significant higher performance than the three reference systems ($p \ll 0.05$)

For the 'Easy' set (where the speaker is facing directly towards the array), there is no statistically significant difference in performance between integrating the MPDR in the training loop, or simply adding it as a preprocessing step to the single channel DCCRN.

For the 'Challenging' set (where the speaker looks away from the array at a 90° angle), there is a significant performance difference for the SNRs of 5 and 10 dB. Here the alternative to the proposed system (where the MPDR is added as a standalone preprocessing step before the pretrained DCCRN) performs statistically significantly worse ($p < 0.05$).

As expected, there is a statistically significant performance decrease from moving from *oracle* TDOAs to *estimated* TDOAs for lower SNRs. However, the system trained with estimated TDOAs still outperforms all baseline systems, indicating that beamforming is useful, even when it is difficult to estimate exact TDOAs.

Intelligibility

Performance with respect to speech intelligibility of the multichannel complex convolutional recurrent mask estimator (the DCCRN) was evaluated objectively in [Paper IV] and [Paper V], and subjectively in [Paper V].

Table 7.8 shows the STOI results of the three variants of the multichannel DCCRN, together with state-of-the-art reference systems and the noisy unprocessed condition, on the 'Easy' and 'Challenging' datasets.

Table 7.8: Objective quality results of the multichannel DCCRN for the 'Easy' and 'Challenging' datasets. STOI results, with clean signal used as reference. Best scores per SNR are shown in bold, where multiple highlighted values indicate that the difference was not statistically significant. Table taken from [Paper III] (©2021 IEEE).

		Easy			Challenging		
SNR [dB]	**WPE**	**0**	**5**	**10**	**0**	**5**	**10**
No enhancement	No	0.69	0.72	0.74	0.60	0.62	0.63
	Yes	0.72	0.76	0.78	0.18	0.66	0.68
ConferencingSpeech	No	0.68	0.72	0.73	0.59	0.61	0.62
2021 baseline [107]	Yes	0.71	0.75	0.77	0.63	0.66	0.67
DCCRN-dir	No	0.73	0.75	0.75	0.64	0.64	0.65
	Yes	0.77	0.78	0.79	0.68	0.69	0.70
GEV (oracle IBM mask) with	No	0.77	0.78	0.79	0.61	0.66	0.67
BAN, by Heymann *et al.* [39]	Yes	0.78	0.80	0.81	0.68	0.69	0.71
MPDR (oracle TDOAs)	No	**0.80**	**0.81**	**0.81**	0.71	0.72	0.73
+ Single channel DCCRN	Yes	**0.81**	**0.82**	**0.83**	**0.74**	0.74	0.75
Jointly trained system	No	**0.80**	**0.82**	**0.83**	**0.75**	**0.76**	**0.77**
(oracle TDOA)	Yes	**0.80**	**0.83**	**0.83**	**0.76**	**0.77**	**0.78**
Jointly trained system	No	0.78	**0.81**	**0.82**	0.72	0.73	0.75
(estimated TDOAs)	Yes	**0.79**	**0.82**	**0.83**	0.73	**0.75**	**0.77**

From a STOI performance perspective, integrating the MPDR in the training loop, or having it as a separate step, does not always give significantly different results. However, at SNRs 5 and 10 dB, and for the 'Challenging' dataset (where the speaker is not facing directly towards the array), the integrated MPDR obtains significantly (two-sample t-test, $p < 0.05$) higher STOI scores. For all SE systems, the WPE dereverberation step always leads to an improved STOI score, although the difference is not at all times statistically significant. The independent two-sample t-test shows that the three proposed systems have statistically significant higher STOI performance than the three reference systems ($p \ll 0.05$).

Figure 7.7 shows the predicted psychometric functions (intelligibility versus SNR), from five different objective intelligbility metrics. Here the scores from the metric have been converted to intelligibility in percentage (see Section 4.2.2).

Figure 7.7 shows the predicted intelligibility performance of two variants (with estimated or true TDOAs) of the multichannel DCCRN, where the MPDR was added as a separate block (and not integrated in the training loop). Additionally, results for several baseline systems are presented: the noisy condition, MPDR-only systems, and the single channel DCCRN-dir.

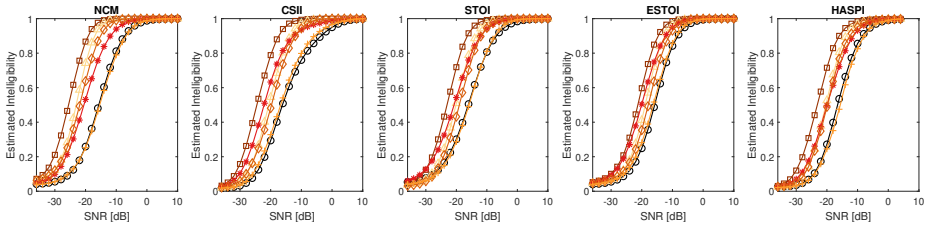


Figure 7.7: Objective Intelligibility (psychometric function) results of the multichannel DCCRN for the speech-in-noise test dataset. Conditions: noisy, Single channel DCCRN, MPDR (estimated TDOAs), MPDR (estimated TDOAs) + single channel DCCRN, MPDR (oracle TDOAs), MPDR (oracle TDOAs) + single channel DCCRN. Figure taken from [Paper V].

The noisy condition and the MPDR-only system with estimated TDOAs are predicted to have equal performance by all metrics. All other systems are predicted to achieve higher intelligibility than these baseline systems. This improvement is also predicted to be present, by all metrics, over the entire range of SNRs.

Figure 7.8 is a summary of Figure 7.7. It shows the expected change in speech recognition threshold (SRT) as predicted by each metric for each system, where the noisy unprocessed condition is taken as the baseline.

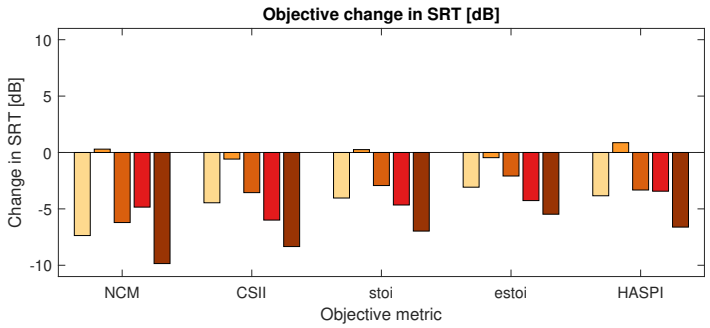


Figure 7.8: Objective Intelligibility (change in SRT) results of the multichannel DCCRN for the speech-in-noise test dataset. The following systems are compared to the single channel noisy condition: Single channel DCCRN, MPDR (estimated TDOAs) only, MPDR (estimated TDOAs) + single channel DCCRN, MPDR (oracle TDOAs) only, MPDR (oracle TDOAs) + single channel DCCRN. Negative numbers indicate improvement in speech intelligibility. Figure taken from [Paper V].

As before, the MPDR-only baseline system with estimated TDOAs is predicted to lead to similar intelligibility as the noisy baseline condition, as only ESTOI and HASPI predict significant changes ($p < 0.05$), and in those cases the predicted changes are still small.

All metrics predict significant decreases in SRT (meaning improved intelligibility) for all other systems with ($p \ll 0.01$).

Figure 7.9 shows the subjective results for all processing conditions, together with their respective objective predictions by the different metrics. These results were obtained from 16 respondents with SRTs below -15 dB on the single channel noisy baseline condition: the best hearing subjects.

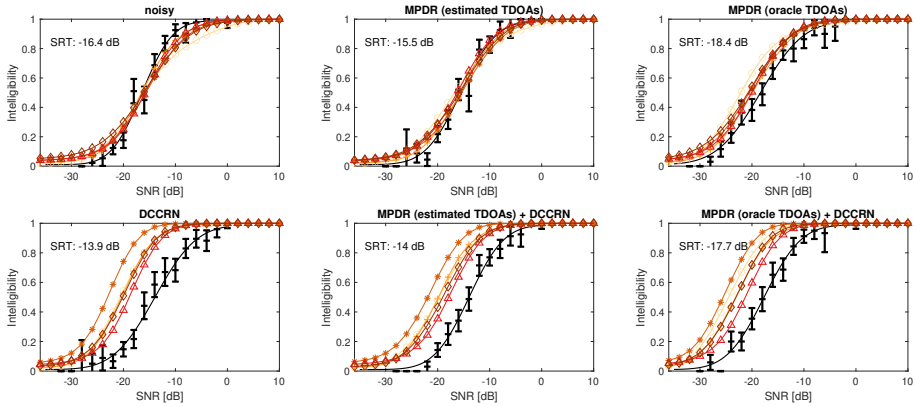


Figure 7.9: Subjective Intelligibility (psychometric function) results of the multichannel systems for the speech-in-noise test dataset. Psychometric functions obtained from normal hearing native speakers for the different processing pipelines. The subjective responses (error bars indicating confidence intervals) and their logistic fits are shown in black (——), together with the corresponding predictions from the objective metrics: CSII, HASPI, NCM, ESTOI, and STOI. Figure taken from [Paper V].

For the noisy condition, the objective scores have been mapped to the obtained subjective intelligibility, to calibrate the scores to the evaluation setup, which is held constant when testing the other processing conditions. All mappings slightly underestimate the slope of the psychometric function, but even at the extreme ends, the effect of this is minor. The same mapping also works reasonably well for the other baseline systems (Figure 7.9, top row), although there seems to be a slight systematic overestimation of intelligibility performance of the MPDR with oracle TDOAs.

However, for the DCCRN based systems, all metrics overestimate intelligibility across the entire SNR range relevant to intelligibility.

Only the MPDR systems supplied with oracle TDOAs obtain lower SRT scores (indicating improved intelligibility) than those obtained for the noisy condition. The MPDR-only system supplied with oracle TDOAs, also outperforms the multichannel DCCRN with oracle TDOAs. All other 'enhancement' systems instead degrade the subjectively measured intelligibility of speech.

This is also apparent from Figure 7.10, which shows the subjective results for three different subject groups (where subjects were divided by their SRT obtained for the noisy condition) and the six processing conditions.

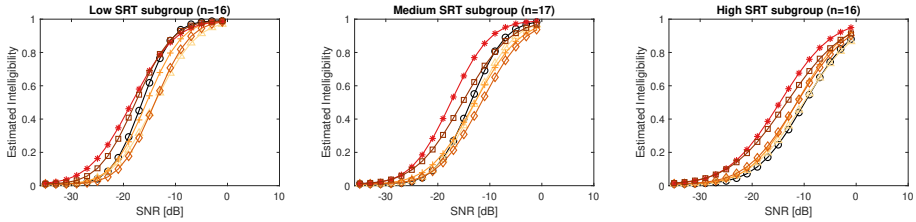


Figure 7.10: Subjective Intelligibility (psychometric function) results of the multichannel systems for the speech-in-noise test dataset, for the three subgroups of subjects. Subgroups: noisy, Single channel DCCRN, MPDR (estimated TDOAs), , MPDR (estimated TDOAs) + single channel DCCRN, MPDR (oracle TDOAs), MPDR (oracle TDOAs) + single channel DCCRN. Figure taken from [Paper V].

Figure 7.11 shows the change in SRT obtained for these subject groups and for the different systems. Again, only the MPDR systems supplied with oracle TDOAs improve the SRT in a statistically significant manner (as determined by the paired Wilcoxon rank sum test and marked with an asterisk). Other systems either degrade the speech or have no statistically significant effect.

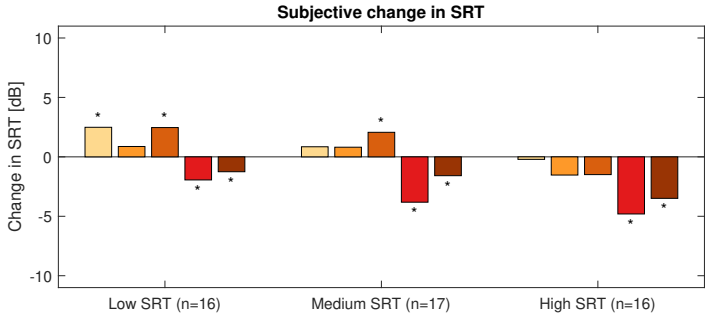


Figure 7.11: Subjective Intelligibility (change in SRT) results of the multichannel systems for the speech-in-noise test dataset, for the three subgroups of subjects. The following systems are compared to the single channel noisy condition: Single channel DCCRN, MPDR (estimated TDOAs) only, MPDR (estimated TDOAs) + single channel DCCRN, MPDR (oracle TDOAs) only, MPDR (oracle TDOAs) + single channel DCCRN. Positive numbers indicate a degradation in speech intelligibility. Statistically significant changes are marked with an *. Figure taken from [Paper V].

All objective metrics predicted increased intelligibility performance, but instead all DCCRN-based systems have no significant effect on the intelligibility, or even worse: They cause intelligibility to be reduced. Additionally, the metrics predicted significant increases in intelligibility for all SNRs, making the OIMs unreliable across the range.

Chapter 8

Conclusions and Further Work

This thesis aims to contribute to the search for speech enhancement systems that improve intelligibility (and quality) of speech signals for human listeners. Particular focus has been put on subjectively evaluating systems, to ensure findings are indeed representative for human subjects.

In 2017, at the start of this study, single channel DNN-based speech enhancement had begun to show potential for the purpose of automatic speech recognition (ASR). With speech recognition, a machine, instead of a human, 'listens' to a speech signal. The goal was to transfer the promising results from ASR to the domain of speech enhancement for human listeners.

The early studies of this thesis [Paper I][Paper II] immediately encountered the well-known, yet somewhat counter-intuitive challenge of subjective speech enhancement: Reducing/removing noise is not all that difficult, but doing so without making your signal less intelligible, is extremely challenging.

The first SE systems evaluated for this thesis were based on fully connected feed forward DNN models that directly estimated the clean spectrum. The effect of applying global variance normalization was evaluated with respect to subjective intelligibility, and no change in performance was found [Paper I].

A less 'aggressive' SE system that attempted to reduce the noise, instead of removing it all together, was then proposed. This system performed better than the systems that tried to remove the noise completely: It resulted in higher subjective quality ratings and improved subjective intelligibility scores. However, when compared to the unprocessed noisy condition, the system only barely improved subjective quality in one specific condition (leaving it insignificantly changed under all other tested situations) and actually slightly degraded speech intelligibility [Paper II]. So the concept of putting more focus on trying to avoid distortion was shown to have potential, but the actual speech enhancement capability of these fully connected feed forward networks was disappointing.

One of the major decisions during the work was then to move to multichannel input. This was motivated by the fact that human hearing is directional, an important ability that is lost in conference calls and most other applications of

speech enhancement. Furthermore, multichannel speech enhancement had (again) already lead to significant performance increases in the field of speech recognition.

To train the multichannel speech enhancement systems in a supervised manner, multichannel data were required. The availability of measured room impulse response data has recently increased, but the training process for a specific microphone array demands specific RIRs for an array's exact element layout. Here simulations provide a solution, and the the image source method is the standard go-to method for this purpose.

However, the image source method overly simplifies reverberance. Therefore, for the work of this thesis, alternative methods that include speaker directivity and diffuse reflections were studied. Especially including the directivity of speakers, lead to improved models for indoor direction of arrival estimation.

Following this work, several closely related multichannel speech enhancement networks were proposed in [Paper IV], which were all trained with input that relied on the directive RIRs. These multichannel networks were based on combining beamforming with a more complex deep neural convolutional recurrent network structure, and achieved much higher objective performance than the earlier fully connected networks. The multichannel system, where beamformer output was masked with a DNN, lead to reduced speech recognition threshold values (indicating improved intelligibility) in subjective testing, but only when supplied with oracle TDOAs. Comparison with a system that relied only on beamforming and did not include the DNN, showed that the neural network model actually degraded the signal, despite objective indicators predicting the exact opposite.

Therefore, the main contribution of this thesis, lies in showing that the tested objective measures for speech intelligibility and quality are not reliable tools to guide the development of DNN based speech enhancement systems, despite their popularity among researchers for exactly this purpose.

This thesis shows this by documenting that:

- STOI predicted improvements in intelligibility for two fully connected feed forward neural network based SE systems (which differed by including, or not including global variance normalization), but both of these system degraded subjective intelligibility [Paper I].
- STOI also predicted very *similar* improvements in intelligibility for two otherwise equal SE systems that were trained to either remove the noise or reduce it. However both systems degraded intelligibility and, additionally, to a different degree [Paper II].
- There was no correlation between POLQA's quality prediction for these noise reducing and noise removing SE systems and subjective measured quality [Paper II].
- NCM, CSII, STOI, ESTOI and HASPI all predicted large and significant improvements in intelligibility for the more advanced single channel DCCRN and the multichannel DCCRN-based systems proposed as part of this thesis, and still these systems degraded the actual subjective intelligibility [Paper V].

These findings are important, especially since relying solely on subjective testing is far too time-consuming to guide system development. The development of speech enhancement systems requires reliable objective estimation of performance, for both the training and selection of models. This thesis shows that the field is lacking tools for objective evaluation; this lack critically hampers progress.

It is also important to note that there are two sides to having unreliable objective metrics. First of all, such metrics may lure developers into thinking they have developed an enhancement system, while in practice the system only degrades the speech. But arguably equally problematic, they might stop the development of systems that may actually have had merit, by predicting (presumably just as unreliably) that these systems will only degrade speech, or have no effect at all.

For example, while the results of this thesis in the field of DOA estimation were promising, the focus on generating more realistic training data for the purpose of speech enhancement was largely abandoned when the objective measures did not predict similar effects for speech enhancement systems. This was done knowing full well that the objective measures could not be relied upon, but they were the only tools available.

8.1 Further Work

Combining the results of this thesis that were obtained in the fields of DOA estimation and SE, a clear opportunity for multichannel speech enhancement can be identified. Namely, with multichannel speech enhancement, the issue of unreliable objective performance measures can be avoided altogether. Instead of training models towards objective measures that do not represent reality, models can be trained for improved speaker localization. Such a procedure includes beamformers to maximally benefit from the understanding of the physics behind signals, while the power of deep learning is put there where traditional models fall short: locating a speaker in a highly noisy reverberant room. Most importantly, with direction of arrival estimation, there is a mathematically well defined error between target and estimate, which is in no way dependent on human hearing and perception.

The DOA models proposed in [Paper III], trained with directive RIRs achieved up to 51 % error reduction compared to a traditional baseline system, while the same models trained with RIRs from omnidirectional sources did worse than this baseline. This shows that the importance of the data generation step has been undervalued as a potential source of performance gain for speaker localization and therefore further work is warranted here. The proposed system of [Paper III] was intentionally kept simple, and focused only on the simulation of reverberance. Next steps include using state-of-the art DOA networks and adding noise. Improved speaker localization will lead to improved performance in all its applications, including speech enhancement for human listeners.

However, for single channel speech, speaker localization is not an option. Therefore, this thesis shows there is a dire need for further research on objective metrics for deep learning based speech enhancement. Specifically, it is important

to study why some signals, while being so much closer to their clean target signal, are still harder to understand than their noisy unprocessed counterparts. Where is the intelligibility lost? Which components of the speech signal are so important that they should never be touched/distorted, even if that means leaving more noise and reverberance in the enhanced signal? A deeper understanding of these fundamental questions should lead to better objective metrics that can predict whether a signal is enhanced or degraded. Without such metrics, the current development process of single channel speech enhancement systems for improved intelligibility, is essentially blind.

Bibliography

[1] F. B. Gelderblom, T. V. Tronstad and E. M. Viggen, 'Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement,' in *INTERSPEECH*, Stockholm, Sweden: ISCA, Aug. 2017, pp. 1968–1972.

[2] F. B. Gelderblom, T. V. Tronstad and E. M. Viggen, 'Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 583–594, Mar. 2019.

[3] F. B. Gelderblom, Y. Liu, J. Kvam and T. A. Myrvoll, 'Synthetic Data For DNN-Based DOA Estimation of Indoor Speech,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada: IEEE, Jun. 2021, pp. 4390–4394.

[4] F. B. Gelderblom and T. A. Myrvoll, 'Deep Complex Convolutional Recurrent Network for Multi-Channel Speech Enhancement and Dereverberation,' in *IEEE International Workshop on Machine Learning for Signal Processing*, Gold Coast, Australia: IEEE, Oct. 2021, pp. 1–6.

[5] F. B. Gelderblom, T. V. Tronstad, T. Svendsen and T. A. Myrvoll, 'On the Predictive Power of Objective Intelligibility Metrics for the Subjective Performance of Deep Complex Convolutional Recurrent Speech Enhancement Networks,' [Submitted].

[6] M. Bain and D. Yanofsky, *CEOs and executives can't remember to unmute either*, https://qz.com/1941207/ceos-and-executives-cant-remember-to-unmute-video-calls-either/, Dec. 2020.

[7] Pioneer, *You're On Mute*, https://www.pioneerproaudio.com/en/news/youre-on-mute-pioneer-pro-audio-sends-speaker-to-space, Sep. 2021.

[8] J. Lim and A. Oppenheim, 'Enhancement and bandwidth compression of noisy speech,' *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[9] Y. Hu and P. C. Loizou, 'A comparative intelligibility study of single-microphone noise reduction algorithms,' *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, Sep. 2007.

[10] P. C. Loizou and G. Kim, 'Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, Jan. 2011.

[11] D. Wang and J. Chen, 'Supervised Speech Separation Based on Deep Learning: An Overview,' *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[12] S. Boll, 'Suppression of acoustic noise in speech using spectral subtraction,' *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[13] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, 1964.

[14] R. McAulay and M. Malpass, 'Speech enhancement using a soft-decision noise suppression filter,' *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[15] Y. Ephraim and D. Malah, 'Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,' *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[16] M. S. Rashid, M. C. Leensen and W. A. Dreschler, 'Application of the Online Hearing Screening Test "Earcheck": Speech Intelligibility in Noise in Teenagers and Young Adults,' *Noise & Health*, vol. 18, no. 85, pp. 312–318, 2016.

[17] S. Tamura, 'An analysis of a noise reduction neural network,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, UK: IEEE, 1989, pp. 2001–2004.

[18] Fei Xie and D. Van Compernolle, 'A family of MLP based nonlinear spectral estimators for noise reduction,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Adelaide, Australia: IEEE, 1994, pp. II/53–II/56.

[19] E. A. Wan and A. T. Nelson, 'Networks for speech enhancement,' in *Handbook of Neural Networks for Speech Processing*, vol. 139, Boston, USA: Artech House, 1999, pp. 7–34.

[20] G. Kim, Y. Lu, Y. Hu and P. C. Loizou, 'An algorithm that improves speech intelligibility in noise for normal-hearing listeners,' *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, Sep. 2009.

[21] K. Han and D. Wang, 'A classification based approach to speech segregation,' *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, Nov. 2012.

[22] F. Weninger, J. L. Roux, J. R. Hershey and S. Watanabe, 'Discriminative NMF and its application to single-channel source separation,' in *INTERSPEECH*, Singapore, Singapore: ISCA, Sep. 2014, pp. 865–869.

[23] Yuxuan Wang and DeLiang Wang, 'Towards Scaling Up Classification-Based Speech Separation,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[24] X. Lu, Y. Tsao, S. Matsuda and C. Hori, 'Speech enhancement based on deep denoising autoencoder,' in *INTERSPEECH*, Lyon, France: ISCA, 2013, pp. 436–440.

[25] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, 'A Regression Approach to Speech Enhancement Based on Deep Neural Networks,' *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[26] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun, S. Srinivasan and J. Gehrke, 'The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework,' in *INTERSPEECH*, Shanghai, China: ISCA, 2020, pp. 2492–2496.

[27] C. K. A. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner and S. Srinivasan, 'Interspeech 2021 Deep Noise Suppression Challenge,' in *INTERSPEECH*, Brno, Czechia: ISCA, 2021, pp. 2796–2800.

[28] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner and S. Srinivasan, 'ICASSP 2021 Deep Noise Suppression Challenge,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada: IEEE, Jun. 2021, pp. 6623–6627.

[29] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng and X. Li, 'A Simultaneous Denoising and Dereverberation Framework with Target Decoupling,' in *INTERSPEECH*, Brno, Czechia: ISCA, Jun. 2021, pp. 2801–2805.

[30] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang and L. Xie, 'DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement,' in *INTERSPEECH*, Shanghai, China: ISCA, 2020, pp. 2472–2476.

[31] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta and M. Matassoni, 'The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada: IEEE, May 2013, pp. 126–130.

[32] J. Barker, E. Vincent, N. Ma, H. Christensen and P. Green, 'The PASCAL CHiME speech separation and recognition challenge,' *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, May 2013.

[33] J. Barker, R. Marxer, E. Vincent and S. Watanabe, 'The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines,' in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Scottsdale, USA: IEEE, Dec. 2015, pp. 504–511.

[34] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker and R. Marxer, 'An analysis of environment, microphone and data simulation mismatches in robust speech recognition,' *Computer Speech & Language*, vol. 46, pp. 535–557, Nov. 2017.

[35] J. Barker, S. Watanabe, E. Vincent and J. Trmal, 'The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines,' in *INTERSPEECH*, Hyderabad, India: ISCA, Sep. 2018, pp. 1561–1565.

[36] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka and N. Ryant, 'CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings,' in *INTERSPEECH*, Shanghai, China: ISCA, May 2020, pp. 1–7.

[37] J. G. Desloge, W. M. Rabinowitz and P. M. Zurek, 'Microphone-Array Hearing Aids with Binaural Output—Part I: Fixed-Processing Systems,' *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 529–542, 1997.

[38] J.-B. Maj, J. Wouters and M. Moonen, 'Noise Reduction Results of an Adaptive Filtering Technique for Dual-Microphone Behind-the-Ear Hearing Aids:' *Ear and Hearing*, vol. 25, no. 3, pp. 215–229, Jun. 2004.

[39] J. Heymann, L. Drude and R. Haeb-Umbach, 'Neural network based spectral mask estimation for acoustic beamforming,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China: IEEE, Mar. 2016, pp. 196–200.

[40] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel and J. L. Roux, 'Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks,' in *INTERSPEECH*, San Francisco, USA: ISCA, 2016, pp. 1981–1985.

[41] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang and D. Wang, 'An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type,' *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1660–1669, 2015.

[42] S. V. Kuyk, W. B. Kleijn and R. C. Hendriks, 'An Evaluation of Intrusive Instrumental Intelligibility Metrics,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, 2018.

[43] Y. Zhao, D. Wang, E. M. Johnson and E. W. Healy, 'A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions,' *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1627–1637, Sep. 2018.

[44] C. K. A. Reddy, V. Gopal and R. Cutler, 'DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada: IEEE, 2021, pp. 6493–6497.

[45] K. S. Helfer and L. A. Wilber, 'Hearing Loss, Aging, and Speech Perception in Reverberation and Noise,' *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 1, pp. 149–155, Mar. 1990.

[46] A. A. de Lima, S. L. Netto, L. W. P. Biscainho, F. P. Freeland, B. C. Bispo, R. A. de Jesus, R. Schafer, A. Said, B. Lee and T. Kalker, 'Quality Evaluation of Reverberation in Audioband Speech Signals,' in *E-Business and Telecommunications*, J. Filipe and M. S. Obaidat, Eds., vol. 48, Berlin, Germany: Springer, 2009, pp. 384–396.

[47] D. Wang and Jae Lim, 'The unimportance of phase in speech enhancement,' *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

[48] K. Paliwal, K. Wójcicki and B. Shannon, 'The importance of phase in speech enhancement,' *Speech Communication*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[49] D. S. Williamson, Y. Wang and D. Wang, 'Complex Ratio Masking for Monaural Speech Separation,' *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[50] T. Gerkmann, M. Krawczyk-Becker and J. Le Roux, 'Phase Processing for Single-Channel Speech Enhancement: History and recent advances,' *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[51] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, 1993.

[52] Nasjonalbiblioteket, *NB Tale - a basic acoustic phonetic speech database for Norwegian*, 2015.

[53] J. Øygarden, 'Norwegian speech audiometry,' Ph.D. dissertation, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, 2009.

[54] G. Hu, *100 Nonspeech Sounds*, http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[55] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal and M. Ritter, 'Audio Set: An ontology and human-labeled dataset for audio events,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, USA: IEEE, 2017, pp. 776–780.

[56] A. Wabnitz, N. Epain, C. Jin and A. van Schaik, 'Room acoustics simulation for multichannel microphone arrays,' in *International Symposium on Room Acoustics*, Melbourne, Australia: ICA, 2010, pp. 1–6.

[57] J. B. Allen and D. A. Berkley, 'Image method for efficiently simulating small-room acoustics,' *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[58] C. F. Eyring, 'Reverberation time in "Dead" rooms,' *The Journal of the Acoustical Society of America*, vol. 1, no. 2A, pp. 168–168, 1930.

[59] A. Farina, 'Simultaneous measurement of impulse response and distortion with a swept-sine technique,' in *Convention of the Audio Engineering Society*, Paris, France: AES, 2000, pp. 5093–5105.

[60] N. A. AG, *TalkBox*, https://www.nti-audio.com/en/products/talkbox, 2022.

[61] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.

[62] D. P. Kingma and J. Ba, 'Adam: A Method for Stochastic Optimization,' in *International Conference on Learning Representations*, San Diego, USA: ICLR, 2015, pp. 1–13.

[63] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro and B. Recht, 'The Marginal Value of Adaptive Gradient Methods in Machine Learning,' in *Conference on Neural Information Processing Systems*, vol. 30, Long Beach, USA: ACM, 2017, pp. 1–14.

[64] V. Dumoulin and F. Visin, *A guide to convolution arithmetic for deep learning*, 2016. arXiv: 1603.07285.

[65] D. N. Kalikow, K. N. Stevens and L. L. Elliott, 'Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability,' *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1337–1351, May 1977.

[66] J. Schmidhuber and F. Cummins, 'Learning to Forget: Continual Prediction with LSTM,' *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[67] V. Badrinarayanan, A. Kendall and R. Cipolla, 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[68] I. Sutskever, O. Vinyals and Q. V. Le, 'Sequence to Sequence Learning with Neural Networks,' in *Conference on Neural Information Processing Systems*, vol. 27, Montreal, Canada: ACM, 2014, pp. 1–9.

[69] G. E. Hinton and R. R. Salakhutdinov, 'Reducing the Dimensionality of Data with Neural Networks,' *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[70] O. Ronneberger, P. Fischer and T. Brox, 'U-Net: Convolutional Networks for Biomedical Image Segmentation,' in *Medical Image Computing and Computer-Assisted Intervention*, Cham, Switserland: Springer, May 2015, pp. 234–241.

[71] P. C. Loizou, 'Speech Quality Assessment,' in *Multimedia Analysis, Processing and Communications*, ser. Studies in Computational Intelligence, W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo and H. Wang, Eds., Berlin, Germany: Springer, 2011, pp. 623–654.

[72] ITU, *Recommendation P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, 2003.

[73] ITU, *Recommendation P.808: Subjective evaluation of speech quality with a crowdsourcing approach*, 2021.

[74] J. Ramsgaard and S. V. Legarth, 'Listening test on headset recordings applying the ITU-T P.835 with trained listeners – results from main systems under test,' Tech. Rep. SenseLab 006-14(2), 2014, p. 32.

[75] ITU, *Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.

[76] ITU, *Recommendation P.862.2: Revised Annex A - Reference implementations and conformance testing for ITU-T Recs P.862, P.862.1 and P.862.2*, 2007.

[77] A. Rix, J. Beerends, M. Hollier and A. Hekstra, 'Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Salt Lake City, USA: IEEE, 2001, pp. 749–752.

[78] ITU, *Recommendation P.863: Perceptual objective listening quality prediction*, 2018.

[79] D. Deutsch, *Musical Illusions and Phantom Words*. New York, USA: Oxford University Press, 2019.

[80] N. Prins and F. Kingdom, *Palamedes: Matlab routines for analyzing psychophysical data.* http://www.palamedestoolbox.org, 2009.

[81] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates and S. Scollie, 'Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools,' *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[82] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, 'A short-time objective intelligibility measure for time-frequency weighted noisy speech,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, TX, USA: IEEE, 2010, pp. 4214–4217.

[83] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, 'An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[84] J. Jensen and C. H. Taal, 'An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[85] I. Holube and B. Kollmeier, 'Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model,' *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

[86] J. Ma, Y. Hu and P. C. Loizou, 'Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,' *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[87] J. M. Kates and K. H. Arehart, 'Coherence and the speech intelligibility index,' *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[88] J. M. Kates and K. H. Arehart, 'The Hearing-Aid Speech Perception Index (HASPI),' *Speech Communication*, vol. 65, pp. 75–93, Nov. 2014.

[89] J. M. Kates, 'The Hearing-Aid Speech Perception Index (HASPI) Version 2,' *Speech Communication*, pp. 35–46, 2021.

[90] American National Standard, *ANSI/ASA S3.5-1997 (R2017): Methods For Calculation Of The Speech Intelligibility Index*, 1997.

[91] R. Giri, U. Isik and A. Krishnaswamy, 'Attention Wave-U-Net for Speech Enhancement,' in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA: IEEE, Oct. 2019, pp. 249–253.

[92] A. Pandey and D. Wang, 'TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton, United Kingdom: IEEE, May 2019, pp. 6875–6879.

[93] S.-W. Fu, Y. Tsao, X. Lu and H. Kawai, 'Raw waveform-based speech enhancement by fully convolutional networks,' in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Kuala Lumpur, Malaysia: IEEE, Dec. 2017, pp. 6–12.

[94] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani and A. Krishnaswamy, 'PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss,' in *INTERSPEECH*, Shanghai, China: ISCA, Oct. 2020, pp. 2487–2491.

[95] G. Zhang, L. Yu, C. Wang and J. Wei, 'Multi-Scale Temporal Frequency Convolutional Network With Axial Attention for Speech Enhancement,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Singapore, Singapore: IEEE, May 2022, pp. 9122–9126.

[96] D. Pearce, H.-G. Hirsch *et al.*, 'The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions.,' in *INTERSPEECH*, Beijing, China: ISCA, 2000, pp. 29–32.

[97] A. Varga and H. J. M. Steeneken, 'Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,' *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[98] R. H. B. Christensen. 'Ordinal–Regression models for ordinal data.' (2015).

[99] R Core Team, *R: A language and environment for statistical computing*, Manual, R Foundation for Statistical Computing, Vienna, Austria, 2017.

[100] Y. Luo and N. Mesgarani, 'TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, Canada: IEEE, Apr. 2018, pp. 696–700.

[101] R. E. S. Lovett, P. T. Kitterick, S. Huang and A. Q. Summerfield, 'The Developmental Trajectory of Spatial Listening Skills in Normal-Hearing Children,' *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 3, pp. 865–878, 2012.

[102] J. Benesty, J. Chen and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Ed. Berlin, Germany: Springer, 2008.

[103] J. H. DiBiase, 'A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays,' Ph.D. dissertation, Brown University, Providence, USA, 2000.

[104] M. B. Brown and Alan B. Forsythe, 'Robust tests for the equality of variances,' *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.

[105] H. L. Van Trees, *Optimim Array Processing*. Wiley, 2002.

[106] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and Biing-Hwang Juang, 'Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[107] W. Rao, Y. Fu, Y. Hu, X. Xu, Y. Jv, J. Han, Z. Jiang, L. Xie, Y. Wang, S. Watanabe, Z.-H. Tan, H. Bu, T. Yu and S. Shang, 'INTERSPEECH 2021 ConferencingSpeech Challenge: Towards Far-field Multi-Channel Speech Enhancement for Video Conferencing,' in *INTERSPEECH*, Brno, Czechia: ISCA, Apr. 2021, pp. 1–5.

Paper I

F. B. Gelderblom, T. V. Tronstad and E. M. Viggen, 'Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement,' in *INTERSPEECH*, Stockholm, Sweden: ISCA, Aug. 2017, pp. 1968–1972

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Subjective intelligibility of deep neural network-based speech enhancement

Femke B. Gelderblom[1], *Tron V. Tronstad*[1], *Erlend Magnus Viggen*[1]

[1]Acoustics Research Centre, SINTEF Digital, Trondheim, Norway

femke.gelderblom@sintef.no, tronvedul.tronstad@sintef.no, erlendmagnus.viggen@sintef.no

Abstract

Recent literature indicates increasing interest in deep neural networks for use in speech enhancement systems. Currently, these systems are mostly evaluated through objective measures of speech quality and/or intelligibility. Subjective intelligibility evaluations of these systems have so far not been reported. In this paper we report the results of a speech recognition test with 15 participants, where the participants were asked to pick out words in background noise before and after enhancement using a common deep neural network approach. We found that, although the objective measure STOI predicts that intelligibility should improve or at the very least stay the same, the speech recognition threshold, which is a measure of intelligibility, deteriorated by 4 dB. These results indicate that STOI is not a good predictor for the subjective intelligibility of deep neural network-based speech enhancement systems. We also found that the postprocessing technique of global variance normalisation does not significantly affect subjective intelligibility.

Index Terms: speech enhancement, deep neural network, subjective evaluation, speech intelligibility

1. Introduction

The field of speech enhancement (SE) aims to improve the quality and/or intelligibility of speech that has been degraded [1]. In the past few years, deep neural networks (DNNs) [2, 3] have emerged as a promising approach for SE, outperforming earlier approaches. SE has been proven useful as a preprocessing step for automatic speech recognition systems to decrease their word error rates [4, 5, 6], but the field also aims to make degraded speech easier to understand and/or more comfortable to listen to for humans [5, 7, 8].

The performance of each of these SE approaches with respect to intelligibility improvement is typically evaluated through objective measures. Especially popular measures are STOI [9], PESQ [10], or the word error rates of speech recognition systems. PESQ was originally designed as a measure for speech quality rather than intelligibility, but was then found to also correlate reasonably well with subjective intelligibility [11]. None of today's objective measures of intelligibility can perfectly predict intelligibility to humans, and their correlation depends on the type of speech degradation present [9, 12].

Thus, listening tests are necessary to quantify the benefit of DNN-based SE for human listeners. Listening tests for speech *quality* have previously been reported in the literature with positive results [5, 7, 8]. Quality is however highly subjective, since whether a signal sounds 'good' or 'poor' is based on listeners' preferences. Intelligibility tests are more objective in nature as these allow for quantitative scoring of how much information the listener actually understood. To our knowledge, and despite its popularity, no one has tested the predictive power of STOI for DNN-based SE against subjective listening tests.

In this work we report the results of a series of listening tests for *intelligibility*, where our test subjects attempted to comprehend speech in background noise, before and after DNN-based speech enhancement. Here, we evaluate whether STOI correctly predicts change in subjective intelligibility for a reasonably common DNN setup. Additionally, we analyse the effect of the 'global variance normalisation' postprocessing step (described in sec. 2.1.3) on intelligibility.

2. Methods

2.1. DNN system overview

The speech enhancement system is loosely based on the system Xu et al. proposed in [8], but omits pre-training with restricted Boltzmann machines as their results indicate that the effect of pre-training was negligible. The DNN was implemented using Keras 1.0.5 [13].

2.1.1. Speech and noise preparation

For training, clean speech was combined with noise to obtain noisy speech. The clean speech was obtained from the Norwegian-language library 'Språkbanken' [14], to ensure that the DNN trained on the same language as used during subjective evaluation. The setup of Språkbanken is similar to that of the more widely used TIMIT. The clean speech database was divided into a training set, a validation set, and a test set (not used for this article). Care was taken to ensure that each set was balanced with respect to gender and dialect, and that no specific speakers or sentences occurred in more than one set. The final training set consisted of 1932 sentences from 137 unique speakers, while the validation set contained 816 sentences from 48 speakers.

Periods of silence lasting longer than 75 ms were trimmed to 75 ms where their levels were 40 dB or more below the peak of the given sentence, to capture the average dynamic range of speech [11]. The 75 ms length was arbitrarily chosen as a compromise between minimising the number of quiet training samples, and maintaining a clear separation between words.

Noisy speech was obtained by combining the clean speech with the same 104 noises Xu et al. used in [8], all obtained from either the Aurora database [15] or Guoning Hu's collection [16]. Six different signal-to-noise ratios (SNRs) ranging from -5 dB to 20 dB, with SNRs applied at sentence level, were used for training. This range was chosen, despite the need for lower SNRs during speech intelligibility testing, as a DNN trained with a more suitable SNR range, but otherwise equal hyperparameters, actually performed worse in terms of STOI values at all SNRs.

The noisy speech, along with clean speech (with 'infinite SNR'), was used as input for the DNN. This lead to a total of 1984 hours of training data. Noisy speech for validation was obtained by combining the clean validation speech with the 15 unseen noises Xu et al. specified in [8], obtained from either the Aurora or NOISEX-92 databases [15, 17]. This resulted in 98 hours of validation data. Both the noisy and clean speech sig-

http://dx.doi.org/10.21437/Interspeech.2017-1041

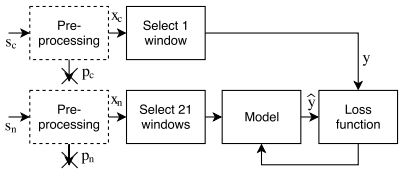


Figure 1: *Diagram of training procedure. The clean and noisy phases output by the preprocessing steps are discarded.*

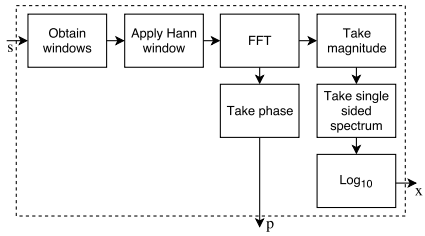


Figure 2: *Diagram of preprocessing steps*

nals were down-sampled to 8 kHz, as this was the lowest sampling rate of any of the original signals.

2.1.2. Training

Figure 1 shows a block diagram of the training procedure. The model learns in a supervised manner, with the standard mean squared error (MSE) loss function

$$\text{MSE} = \frac{1}{n}\sum_k \left(\hat{y}_k - y_k\right)^2, \tag{1}$$

where $\hat{y}_k$ and y_k represent the kth frequency bins of the enhanced and clean log-power spectral features, respectively. The features were obtained through the preprocessing steps shown in Figure 2. During preprocessing, the signal is first separated into windows that overlap by 50 %. The windows consist of 256 samples, and thus represent a timeframe of 32 ms at 8 kHz. The Hann window function is then applied to each window before the result is Fourier transformed. Redundant information above the Nyquist frequency is discarded from the resulting magnitude spectrum to obtain a single-sided output. Finally, log-power spectrum features are calculated for each window. After preprocessing, the input vector is obtained by stacking 21 sequential 50 % overlapping windows that contain the log-power spectral features. This provides the DNN with 160 ms historic and 160 ms future context. The phase of both clean and noisy speech is ignored during training. No normalisation of input or output was applied.

The DNN model is a multi-layer perceptron, a feedforward neural network with fully connected layers. It has three hidden layers, each with 2048 nodes and LeakyReLU activation functions. The model is trained with 50 % dropout on the hidden layers using the Adam optimiser with a learning rate of 10^{-5}. The activation function of the output layer is linear.

Training continued until the STOI value reached a maximum for the validation set at the 8th epoch. The model's state at this epoch was used for enhancement. We also trained a number of different models with different hyperparameters; the model described here was selected due to its better STOI performance.

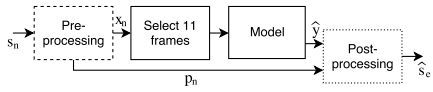


Figure 3: *Diagram of enhancement procedure*

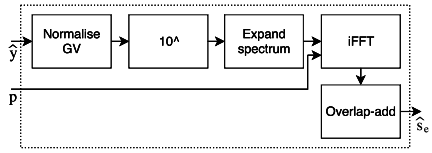


Figure 4: *Diagram of postprocessing steps*

2.1.3. Enhancement

After training, the model could be used to enhance noisy speech. Figure 3 shows the enhancement procedure, and Figure 4 shows the postprocessing steps.

Postprocessing mainly consists of reversing the steps that were taken during preprocessing, using the *noisy* phase for waveform reconstruction. The first step, global variance normalisation (GVN), is the exception to this reversal. This step aims to prevent over-smoothing by enforcing the variance of the enhanced speech to be equal to the variance of actual clean speech. During GVN, the DNN's output features are multiplied with a frequency bin independent factor calculated as

$$\beta = \sqrt{\frac{\text{var}_{m,k}[y_k(m)]}{\text{var}_{m,k}[\hat{y}_k(m)]}}, \tag{2}$$

where $\text{var}_{m,k}$ represents the variance over all values of m and k, with m indexing examples in the training set, and k indexing frequency bins. Furthermore, from the law of total variance we can calculate this variance as

$$\underset{m,k}{\text{var}}[a_k(m)] = \frac{1}{K}\sum_k \underset{m}{\text{var}}[a_k(m)] + \underset{k}{\text{var}}\left(\underset{m}{\text{var}}[a_k(m)]\right), \tag{3}$$

where K equals the total number of frequency bins and $a_k(m)$ represents either $y_k(m)$ or $\hat{y}_k(m)$. This specific method for the calculation of the global variance combines readily with Welford's online algorithm for variance computation, which is well suited to working with large data sets [18]. Two systems were tested for this work; one with, and one without the GVN step.

2.2. Objective evaluation

The short-term objective intelligibility (STOI) measure [9] was used to test the model's performance. The advantages of STOI include a documented strong correlation with subjective speech intelligibility [9] and the possibility to compare obtained results with earlier publications [8]. Additionally, unlike with some other popular objective measures like PESQ, use of STOI is not restricted by licencing.

Objective evaluation results were obtained both for the validation set and for the signals used during subjective testing.

2.3. Subjective evaluation

The subjective evaluation of intelligibility was performed using a speech recognition test. Figure 5 shows the user interface

Figure 5: *The GUI of the Norwegian-language subjective test*

implemented in MATLAB [19]. Random five-word sentences, all uttered by the same male speaker, were presented at different SNRs to determine the speech recognition threshold (SRT). All sentences were in Norwegian and structured the same way: [Name], [Verb], [Numeral], [Adjective], [Noun], with 10 options for each. The subjects' task was to pick out which word in each category was in the sentence they just heard. The speech material has been taken from Øygarden's hearing in noise test, which is based on Hagerman sentences [20].

To keep the subjective test to a manageable length, only one noise file was used: a road traffic recording from a crossroad in central Trondheim, a common type of background noise in cities. Each sentence was mixed with a random section of this noise file at the desired SNR. The SNR was calculated from the root-mean-square (RMS) value for the sentence without noise and the RMS value for the selected section of the noise signal. The background noise was kept constant at a comfortable level while the speech was varied to achieve the correct SNR. The speaker, utterances, and noise used in this test had not been included during DNN training nor during validation.

Each subject completed three tests. For each test case, all material was first down-sampled to 8 kHz. One test set was left otherwise untreated ('Noisy'), while for the other cases the speech was enhanced according to the method described in sec. 2.1.3 ('DNN with/without GVN'), where the GVN step was only included for one of these cases. The material of each test set was subsequently up-sampled to 44.1 kHz before being presented to the subject. All sentences were presented binaurally with Sennheiser HDA-200 headphones via an external sound card (Roland Edirol UA-101).

An adaptive procedure called the Ψ method [21] was used to determine the presentation levels during testing. The method uses the entropy of the posterior probability distribution in the determination of the next stimuli level. The Palamedes MATLAB toolbox [22] was used for the realisation of the Ψ method.

The test was not forced choice, but the test subjects were encouraged to guess whenever they thought they (partly) recognised a word. Both the guess and lapse rate were set to 0.01 in the method. The threshold and slope value were allowed to vary in the estimation of the psychometric function. The stimulation range of the SNRs was from -36 dB to 10 dB, in 2 dB steps.

15 persons, with ages from 39 to 65 (Mean $= 54.2$, SD $= 9.5$), participated. The only selection criteria observed was that all participants had to have Norwegian as their first language. All test subjects were given a training session before the three situations (Noisy, DNN with GVN, and DNN without GVN) were tested and the test sequence was randomised between each individual to reduce any further training effect that could occur during the session. The test subjects were also allowed to take a break during the test if they desired.

Table 1: *STOI results for the validation set. Results are averaged over the 15 unseen noise types and stated together with their sample standard deviation.*

SNR	Noisy	DNN without GVN	DNN with GVN
20	0.95 (0.01)	0.92 (0.01)	0.91 (0.01)
15	0.91 (0.02)	0.90 (0.01)	0.89 (0.01)
10	0.85 (0.03)	0.86 (0.02)	0.85 (0.02)
5	0.76 (0.04)	0.80 (0.02)	0.79 (0.02)
0	0.65 (0.04)	0.71 (0.03)	0.71 (0.03)
-5	0.55 (0.04)	0.61 (0.04)	0.60 (0.04)

3. Results

3.1. Objective evaluation

Table 1 shows the STOI results for the validation set. The GVN step shows no significant effect on the STOI results. DNN processing leads to improved scores as compared to the baseline for all SNRs under 10 dB. Looking at our unprocessed 'noisy' baseline, our STOI results at low SNRs are lower by 0.05 than what Xu et al. [8] found using the TIMIT speech library. As we use the same noise types, and we were able to reproduce their 'noisy' STOI scores using TIMIT, this discrepancy shows that STOI predicts different intelligibility for the two libraries under equal noise conditions.

Figure 6 shows a plot of the average STOI scores obtained for the files processed for subjective evaluation. As with the validation set results, the use of GVN did not significantly affect model performance. At higher SNRs, DNN processing performs worse than the noisy baseline. However, for low SNRs STOI scores suggest improvement even outside the training range. According to the objective evaluation, DNN processing ought to be beneficial for all SNRs in between -14 dB and 4 dB.

3.2. Subjective evaluation

Figure 7 shows the results from the subjective tests. Specifically, it shows the differences between the reference and the two DNN models, both for the SRT and the slope of the psychometric function at SRT. All test subjects performed worse on the SRT, while the slope values are more mixed.

To assess the normality of the data, we performed an Anderson-Darling test on all the differences. The SRT differences for the DNN without GVN failed the normality test. The non-normality is presumably a consequence of the small sample size. To cope with this, we performed a Wilcoxon signed rank test to compare the models with the reference. The tests

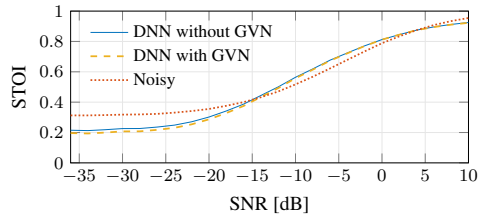


Figure 6: *STOI results for the subjective evaluation set*

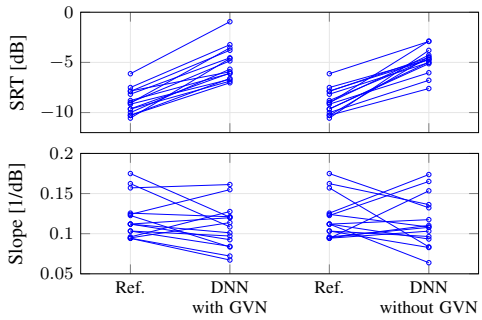


Figure 7: *Comparison between unenhanced reference data and DNN data. Upper: Speech recognition thresholds (SRT). Lower: Slope of the psychometric function at SRT.*

showed a significant difference ($W = 120, p = .0001$ for both) between the models and the reference; not surprisingly, since all the test subjects performed worse on the DNN models (see Figure 7). The differences in median SRT values were (using Hodges-Lehman estimators) 3.8 [3.2, 4.4] and 3.9 [3.2, 4.8] for DNN with GVN and without GVN, respectively. The numbers in brackets are the 95 % confidence intervals.

The slope of psychometric functions were compared using a two sample F-test. Neither DNN with GVN ($F_{14,14} = .91$, NS), nor DNN without GVN ($F_{14,14} = .69$, NS) showed any significant difference from the reference.

4. Discussion

The STOI results for unprocessed noisy validation files from the Norwegian database (Table 1) differ from those obtained for the TIMIT database by Xu et al. [8]. This complicates comparing model performance directly. However, the results are similar to those of Xu et al. in the sense that STOI improvement is arguably insignificant for SNRs of 10 dB and above. For lower SNRs, STOI predicts our system will achieve improvements of up to 6 percent on the subjective scale. This is less than Xu et al. achieved, but significant enough to predict that subjective SRTs ought to decrease, or at the very least, stay the same.

The DNN model was not trained at SNRs below -5 dB, but surprisingly, the STOI results shown in Figure 6 indicate that the model enhances noisy speech with SNRs up to 9 dB below its training range. This means that during subjective testing, 93.8 % of sentences presented to the listener had an SNR that fell in the functional range of the model (from -14 dB to 4 dB). All test subjects also achieved SRT values within this range. Nonetheless, the results from the subjective testing showed that the DNN models performed significantly worse (SRTs increased with approx. 4 dB) than the unprocessed sentences. Even from a conservative perspective where we could say that the changes the model attains in STOI are insignificant, the SRTs should not have increased this much. Thus, STOI significantly overestimates the speech intelligibility of our DNN-based speech enhancement system.

On the other hand, STOI correctly predicts that GVN has no significant effect on speech intelligibility. According to Xu et al. [8], PESQ results are, in contrast, significantly affected when GVN is used during postprocessing of a DNN-based speech enhancement system. This may indicate that GVN matters more to speech quality, but we did not investigate this further.

Our DNN model was selected because it obtained better STOI scores than similar networks trained for a larger range of SNRs or with different hyperparameters. Our results however indicate that STOI fails to predict the intelligibility of a DNN-based speech enhancement system. This directly undermines our model selection criterion. It is therefore possible that one of our other models would have lead to better subjective scores.

All test sentences were uttered by the same male speaker; it is likely that the DNN model will perform differently for different speakers. Similarly, the results are presumably affected by the choice of background noise. We expect that the traffic noise used here performs better than for example noise that consists mainly of human speech (babble), since the DNN models might try to enhance some of the speakers in the noise as well. Similarly, other types of noise may again be easier for the system to handle. A more comprehensive study of the suitability of STOI as an objective evaluation measure for DNN-based speech enhancement would need to include a variety of speakers and noises. Such a comprehensive study will be time-consuming and the material for the speech-in-noise tests will need to be carefully constructed for unbiased results.

The choice of sampling frequency (8 kHz) might also have affected the results. Increasing the sampling frequency to 16 kHz, or higher, would probably have improved the speech recognition for all the tests [23], but it is not clear if this would have changed the results of this study.

Another possible bias in this study is the effect of hearing loss. As the analysis of the subjective testing looked at the difference between a reference and the DNN models, we assumed that a hearing loss would not alter the results. Only one test subject had a hearing aid, but this was not used during the subjective test. Since the test subjects' ages were relatively high (mean $= 54.2$) it can be assumed that several of the test subjects were affected by presbycusis. Even if the intra-subject change in SRTs should be independent of hearing impairment, this may have affected results.

Our analysis is limited to speech intelligibility, and does not consider the effect of DNNs on speech quality. The relationship between these two parameters is not fully understood. For many communication systems, intelligibility may be approaching 100 %, while user satisfaction is still limited. Here, listening effort tests, where a speech intelligibility test is combined with another task, may provide a good compromise between providing objective results for the more quality related question of how comfortable or easy it is to listen to the enhanced speech.

5. Conclusion

We have tested a DNN-based speech enhancement system with listening tests to determine the subjective intelligibility of processed noisy speech. Our results show a significant degradation in intelligibility, even though STOI scores predicted otherwise. Therefore we advise against solely relying on STOI when designing DNN-based speech enhancement systems for human listeners. Our results further show that the postprocessing technique of global variance normalisation does not significantly affect subjective intelligibility.

6. Acknowledgements

We thank Tor Andre Myrvoll for his guidance in setting up the speech enhancement system and valuable insights during discussions. We also thank our volunteer test subjects.

7. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[3] M. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[4] Z.-Q. Wang and D. Wang, "A Joint Training Framework for Robust Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1–11, 2016.

[5] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, 2016.

[6] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *INTERSPEECH*, Singapore, 2014, pp. 616–620.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[8] ——, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[10] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, ITU-T Recommendation P.862, 2001.

[11] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, p. 3387, 2009.

[12] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7539284/

[13] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015, last accessed on 2017-06-01.

[14] Nasjonalbiblioteket, "NB Tale - a basic acoustic phonetic speech database for Norwegian," http://www.nb.no/sprakbanken/show?serial=sbr-31, 2015, last accessed on 2017-06-01.

[15] D. Pearce, H.-G. Hirsch, and others, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." in *Interspeech*, 2000, pp. 29–32.

[16] Guoning Hu, "100 Nonspeech Sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, last accessed on 2017-03-14.

[17] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[18] T. F. Chan, G. H. Golub, and R. J. LeVeque, "Algorithms for computing the sample variance: Analysis and recommendations," *The American Statistician*, vol. 37, no. 3, pp. 242–247, 1983. [Online]. Available: http://www.jstor.org/stable/2683386?origin=crossref

[19] The MathWorks, Inc., *MATLAB R2016a*. Massachusetts, United States: Natick, 2016.

[20] J. Øygarden, "Norwegian speech audiometry," Ph.D. dissertation, Norwegian University of Science and Technology (NTNU), Faculty of Art, Department of Language and Communication Studies, 2009.

[21] L. L. Kontsevich and C. W. Tyler, "Bayesian adaptive estimation of psychometric slope and threshold," *Vision research*, vol. 39, no. 16, pp. 2729–2737, 1999.

[22] N. Prins and F. Kingdom, "Palamedes: Matlab routines for analyzing psychophysical data." http://www.palamedestoolbox.org, 2009, last accessed on 2017-03-14.

[23] A. B. Silberer, "Importance of high frequency audibility on speech recognition with and without visual cues in listeners with normal hearing," Ph.D. dissertation, University of Iowa, 2014.

Paper II

F. B. Gelderblom, T. V. Tronstad and E. M. Viggen, 'Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement,' *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 583–594, Mar. 2019

Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement

Femke B. Gelderblom, Tron V. Tronstad, Erlend Magnus Viggen

***Abstract*—Speech enhancement systems aim to improve the quality and intelligibility of noisy speech. In this study, we compare two speech enhancement systems based on deep neural networks. The speech intelligibility and quality of both systems was evaluated subjectively, by a Speech Recognition Test based on Hagerman sentences and a translation of the ITU-T P.835 recommendation, respectively. Results were compared with the objective measures STOI and POLQA. Neither STOI nor POLQA reliably predicted subjective results. While STOI anticipated improvement, subjective results for both models showed degradation of speech intelligibility. POLQA results were overall hardly affected, while the subjective results showed significant changes in overall quality, both positive and negative, in many of the tests. One of the systems was trained to remove all noise; a strategy that is common in speech enhancement systems found in the literature. The other system was trained to only reduce the noise such that the signal-to-noise ratio increased with 10 dB. The latter system subjectively outperformed the system that attempted to remove noise completely. From this, we conclude that objective evaluation cannot replace subjective evaluation until a measure that reliably predicts intelligibility and quality for deep neural network based systems has been identified. Results further indicate that it may be beneficial to move away from more aggressive noise removal strategies towards noise reduction strategies that cause less speech distortion.**

***Index Terms*—speech enhancement, artificial neural networks, subjective evaluation, speech intelligibility, speech quality**

I. Introduction

THE field of speech enhancement (SE) deals with improving speech signals that have been degraded by noise [1]. Speech enhancement is commonly applied in automatic speech recognition (ASR) systems as a preprocessing step to improve these systems' accuracy in noisy environments [2], [3], [4]. Recently, research into this application has flourished, resulting in significant performance increases of ASR systems. This success has also lead to a renewed interest in the application of speech enhancement for human listeners, where the goal is to make the speech easier to understand (i.e., increase *speech intelligibility*) and/or more comfortable and less tiring to listen to (i.e., increase *speech quality*) [3], [5], [6]. The latter application is especially important within the fields of telecommunication and hearing assistive technology.

There exists a wide range of SE techniques. As in many other fields, techniques based on deep neural networks (DNNs) [7], [8] are currently receiving a lot of interest due to their potential to outperform earlier techniques. For ASR systems, performance is measured by a SE system's ability to decrease the word error rate. For human listeners, performance is ideally determined through subjective evaluation of speech intelligibility and/or speech quality [1]. These tests generally compare the listeners' evaluations of noisy speech before and after enhancement, to quantify the effect of different SE strategies.

However, since these subjective evaluations are time-consuming to perform, *objective measures* are often calculated instead. These objective measures typically quantify a degraded speech signal in comparison to a clean speech signal. For speech intelligibility, a popular objective measure is STOI, which performs well against competing intelligibility measures [9] and has a reference implementation freely available [10]. For speech quality, popular measures are PESQ [11] and its successor POLQA [12]. Although PESQ also has a downloadable reference implementation [13], licenses must be purchased to use PESQ and POLQA.

When evaluating the change in intelligibility or quality obtained with SE systems, measures based on clean speech and *unenhanced* noisy speech are calculated to establish a reference. Then, the same measures are calculated for clean speech and *enhanced* noisy speech. Comparison of these results then predicts how much the SE system affects speech intelligibility or quality.

However, these objective measures have been designed to predict intelligibility or quality for relatively simple degradations, such as additive noise, and do not necessarily perform well for more complex degradations [9], [14], [15]. DNN-based SE systems perform a complex nonlinear processing of the noisy signal, and multiple authors have found that STOI is not a reliable predictor of whether or not a given DNN-based system actually improves speech intelligibility [16], [17], [18]. Until a specific objective measure has been shown to give reliable predictions for these systems, time-consuming subjective evaluations are required to test DNN-based SE systems.

Femke B. Gelderblom and Tron V. Tronstad and Erlend Magnus Viggen are with the Acoustics Research Centre, Connectivity Technologies and Platforms, SINTEF Digital, Trondheim, Norway. Erlend Magnus Viggen is additionally with the Centre for Innovative Ultrasound Solutions, Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, NTNU – Norwegian University of Science and Technology, Trondheim, Norway, and is supported by grant no. 237887 from the Research Council of Norway.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The material consists of tables containing detailed results of the statistical analysis. Contact tronvedul.tronstad@sintef.no for further questions about this material.

When training a DNN using supervised learning techniques, we must always specify the format of the input and the target output. In speech enhancement, a common input is the logarithmic half-spectra of several adjacent semi-overlapping frames of the noisy speech signal, and a common target output is the logarithmic half-spectrum of one corresponding frame of the clean speech signal [5], [6], [17], [4], [19], [20], [21], [22], [23], [24]. In this way, the training process leads the DNN towards returning perfectly noise free speech. This approach has shown significant merit for application in ASR systems.

While SE systems often manage to reduce or even remove the presence of noise, the output speech is generally audibly degraded by this process, especially at low signal-to-noise ratios (SNRs), as SE systems cannot perfectly distinguish between speech and noise when attempting to remove only the latter [25], [24]. In fact, our previous study on one possible realization of a DNN-based SE system [17] found that this degradation significantly reduced the speech intelligibility, compared to that of the noisy speech before the enhancement was carried out. We found that the speech recognition threshold (SRT), which is the SNR at which 50 % of words are understood, degraded on median by around 4 dB, in stark contrast with the positive performance predicted by STOI.

However, this is not surprising, when put in the perspective that humans are very sensitive to degradation in speech signals, while capable of scoring 100 % intelligibility despite noisy conditions. This motivates studying training methods that look for a suitable compromise between noise reduction and speech degradation, in addition to methods that focus on finding noise free speech.

Supervised training of a DNN involves optimizing some statistical measure, called the loss function, which is based on the difference between the desired DNN output and the actual DNN output for a given input. A typical loss function in DNN-based SE is the mean-squared-error (MSE) value based on the target clean speech and the DNN's output. By iteratively adjusting the weights of the DNN to obtain a lower MSE, the training process moves the DNN's output towards the target output.

One way of shifting the "focus" of the DNN training towards speech and away from noise, in the hope of indirectly reducing speech degradation, would be to use more speech-aware loss functions. Kumar et al. proposed using a weighted squared error based on absolute thresholds of hearing, but did not report results that allow for direct performance of this loss function to a standard MSE approach [25]. Others investigated using STOI as a training target, but did not obtain improvements of such a degree that it is obvious that they will show in a subjective evaluation [26], [27], [28]. We also investigated a number of other options, such as an MSE loss function weighted according to the SII band importance weights [29] or gammatone weights inspired by the objective intelligibility measure by Dau et al. [30]. However, none of our unpublished pilot studies based on these approaches showed enough promise to warrant continuing with subjective testing on a larger scale.

Another alternative to guide the training process is to go away from using a noise free target. This article investigates using a DNN target output that is not perfectly clean speech; rather, it corresponds to the input signal at a 10 dB higher SNR. This target, which is closer to the input, ensures that noise is still significantly reduced relative to the speech, but in a less aggressive manner. This may reduce the overall speech degradation, and consequently increase the speech quality and intelligibility compared to the more aggressive clean-speech target where the noise reduction likely has a stronger negative impact on the speech [25], [24]. A 10 dB improvement in SNR is clearly perceptible, since it perceptually corresponds to a halving/doubling of the loudness of the noise/speech signal [31]. Even though intelligibility improvement rates have been shown to vary a lot between test situations (from 1 % per dB to 44 % per dB, with a mean value of 7.5 % per dB) [32], a 10 dB improvement of the SNR should always be clearly measurable in subjective testing. The optimal may both be higher (less noise) or lower (less distortion), but finding an optimal value of the target's SNR improvement is out of the scope of this study.

In the study reported in this article, we trained two DNN-based SE systems based on these two targets, as described in Section II-A. We subsequently generated a large number of sound clips where clean sentences were mixed with different background noises at various SNRs and enhanced with either of the SE systems, as described in Section II-B. Our test subjects (Section II-E) were asked to perform subjective evaluations of the speech intelligibility (Section II-C) and speech quality (Section II-D) of these clips. Additionally, we calculated STOI and POLQA scores for comparison (Section II-F). We provide our results in Section III and discuss them further in Section IV, before we conclude in Section V.

II. Method

A. Data and DNN setup

In this work, we used the same general DNN setup as in our previous work [17], which is loosely based on the system by Xu et al. [6] and implemented using Keras [33]. As the details are given in [17], we will only give the essentials here.

The clean speech for training and validation of the DNN was taken from the Norwegian speech audio dataset NB Tale [34]. This forms part of the Norwegian language library Språkbanken, and is set up similarly to the widely used English-language TIMIT dataset. Periods of silence lasting longer than 75 ms were trimmed to 75 ms where their levels were 40 dB or more below the peak of the given sentence, to capture the average dynamic range of speech [35]. The clean speech was divided into training, validation, and test sets that did not overlap in either speakers or sentences, with 1932 sentences from 137 speakers in the training set and 816 sentences from 48 speakers in the validation set. We chose to use Norwegian primarily because of our access to Norwegian native speakers as test subjects. However, we expect our results to be transferable to e.g. English, as the two are closely related Germanic languages.

During training and validation, the input was based on noisy speech constructed by combining this clean speech with noises taken from the Aurora database [36], the NOISEX-92

database [37], and Guoning Hu's collection [38]. We chose the same 104 noises for training and 15 unseen noises for validation as Xu et al. Both sets contained both stationary and non-stationary noise sources. For each set, we combined every type of noise with every sentence in that set, giving us a total of 1984 hours of training data and 98 hours of validation data. For the input data, six different SNRs uniformly spaced from −5 dB to 20 dB were used during training. Before they were combined, the speech and noise signals were downsampled to 8 kHz, the lowest sampling rate among the noise types.

The input was constructed from single-sided log-power spectra of frames of this noisy speech. Each frame was found from a 256-sample (32 ms) Hann window of the time signal. Adjacent frames overlapped by 50 % in time. These windowed frames were Fourier transformed and redundant information above the Nyquist limit was discarded, giving a single-sided spectrum. Then, the log-power spectrum was found by taking the base-10 logarithm of the magnitude of each frequency bin. The final input vectors were found by stacking 21 such log-power spectra, based on the adjacent overlapping frames, after each other. The task of the DNN was to enhance only the middle frame, and the stacking thus provided the DNN with 160 ms of past context and 160 ms of future context.

When training the DNN, we used two different training targets, leading to two different DNN models:

- **Model 1:** Here, the training target was the single-sided log-spectrum of a frame of clean speech, unaffected by noise. This is the model we reported earlier [17].
- **Model 2:** Here, the training target was the single-sided log-spectrum of a frame of clean speech mixed with the exact same noise as in the input, but at a 10 dB higher SNR.

The loss function was a standard mean squared error between the DNN output and the training target.

In both models, the DNN was a simple feedforward network with three hidden layers in addition to the input and output layer. Each hidden layer used LeakyReLU activation functions. The models were trained with 50 % dropout in the hidden layers using the Adam optimizer. We trained a number of different candidate networks over the same ranges of hyperparameters for both models. The ranges included hidden layers with 1024, 2048, and 3072 units. The final network for each model was chosen as the best epoch of all the candidate networks, according to the STOI scores that we evaluated for the validation set at 0 dB SNR after every epoch. The resulting Model 1 used 2048 nodes per hidden layer and a learning rate of 10^{-5}, while Model 2 used 3072 nodes and a rate of 10^{-2}. The final epochs for Model 1 and Model 2 were the 8th and the 33rd epochs, respectively.

In order not to change the experimental procedure more than necessary, we picked Model 2 based on STOI scores in the same way as we picked Model 1 in [17]. However, as earlier work indicates that STOI is not a robust predictor of the intelligibility of DNN-based SE systems, as explained in Section I, this approach is hardly ideal as we cannot truly expect the maximum-STOI epochs to perform best in a subjective evaluation. However, given the relatively minor performance changes reported in [39], [26], [27], [28] and the fact that we merely used STOI for selection rather than as a training target (with an expected weaker effect), we do not expect that the approach with STOI as selection criteria will have had a major impact on our results. Until one or more objective measures are identified as a robust predictor of intelligibility and/or quality, determining the best epoch or the best hyperparameters will remain problematic, as subjective evaluations of sufficient precision are generally too time-consuming to be feasible for anything other than a final test of a trained system. While an extensive study into this topic is outside the scope of this article, Section IV does compare STOI and POLQA scores with subjective evaluations of speech intelligibility and speech quality, respectively.

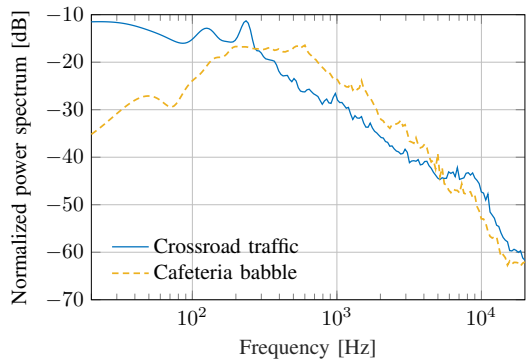


Fig. 1. Long-term average spectra of the two background noises used in the subjective evaluations. These spectra were computed using Welch's method using 2048-sample Hann windows, after the sounds had been normalized to have RMS values of 1.

When the trained network was used to enhance noisy speech, the process of reconstructing a waveform from the DNN output essentially consisted of reversing the steps used to create the input data. As the log-spectrum output does not contain phase information, we used the noisy input phase in this process. Unlike in our previous publication [17], we did not use the global variance normalization preprocessing step, for two reasons: We found then that it did not affect the results of the subjective intelligibility evaluation, and including it as a factor would double the already considerable number of tests to be performed by the test subjects.

B. Generation of test sounds

For the subjective evaluations, we generated a variety of single-channel clips of speech in noise at various SNRs. We generated clips both without enhancement and with enhancement by Models 1 and 2.

In all the clips, the base speech was a randomly generated five-word Hagerman sentence in Norwegian, generated as described by Øygarden [40]. Each sentence was built up the same way: [Name], [Verb], [Numeral], [Adjective], [Noun], with 10 possible options for each class of word. As a basis, we generated 500 reference speech clips of unique, noise-free sentences.

We then mixed these files with background noise at various SNRs, as described later in Sections II-D and II-C. Two different types of noise were used. We used one 17-second clip of traffic noise from a crossroad in Trondheim, and one 25-second clip of babble noise recorded in a university cafeteria during a lunch break. Neither type of noise was present in either the training or the validation described in Section II-A. Both noises were originally recorded with a sampling rate of 44.1 kHz. The long-term average spectra of both noises before downsampling is shown in Figure 1.

Finally, these unenhanced noisy clips were run through the two trained DNN models described in Section II-A. Thus, for each SNR, we ended up with 3000 unique degraded clips: The 500 reference clips, times two types of noise, times three types of enhancement (Unenhanced, Model 1, and Model 2). Figure 2 shows spectrogram examples of one speech clip at different points in this process.

C. Speech intelligibility

The speech recognition threshold (SRT), which is a common measure of speech intelligibility [1], was determined using the same method as in our previous work [17]. The five-word test sentences were built up from five word categories as described in Section II-B. The test subjects' task was to select the words they could hear using a graphical user interface with ten possible words per category, a total of 50 words. Guessing was allowed, but the test was not forced choice.

The test subject responses were given as input to an adaptive psychometric function estimation procedure called the Ψ-method [41], which continuously estimated the SRT during the test. The final threshold estimate was found after 20 sentences (i.e. 100 words in total). All parameters used in the method were identical to the ones used in our previous study [17], i.e. a guess and lapse rate of 0.01, psychometric function based on a cumulative normal probability density function, and stimulation range of the SNR from -36 dB to 10 dB in 2 dB steps.

The method was implemented in MATLAB [42] and the sentences were presented binaurally for all test subjects. An external sound card (Edirol UA-25) was connected with USB cable to the computer. Headphones (Howard Leight Sync Stereo Headband) with sound attenuating properties were used for the playback. The test was performed in an ordinary single room office with low background noise level. The background noise level was not measured during the test, but considering the headphones' sound attenuating properties and the signal levels involved, the results should not be affected.

Since the results from our previous study [17] did not pass the normality distribution assumption, we decided to use Wilcoxon tests to decide if differences were significant.

D. Speech quality

Speech quality was assessed using the method described in ITU-T P.835 [43]. The ordinal scales presented in the recommendation were translated to Norwegian by comparing and combining the official English and French version, together with a Danish version presented by [44]. The English and Norwegian versions can be seen in Table I and II respectively. Note that the Norwegian noise scale is slightly different than the English version. Instead of using "slightly noticeable", "noticeable but not intrusive" and "somewhat intrusive" as rating 4, 3 and 2, the Norwegian version uses "noticeable but not intrusive", "somewhat intrusive" and "intrusive". The reason for changing the scale was an observation made during a pilot test for the study. Several of the participants noted that it was difficult to distinguish between "slightly noticeable" and "noticeable but not intrusive". To cope with this problem, we adapted the French version [45], which uses a slightly different scale, in the translation.

TABLE I
ENGLISH VERSION OF THE ORDINAL SCALES USED IN ITU-T P.835 [43].

Rating	Speech	Noise	Overall quality
5	Not distorted	Not noticeable	Excellent
4	Slightly distorted	Slightly noticeable	Good
3	Somewhat distorted	Noticeable but not intrusive	Fair
2	Fairly distorted	Somewhat intrusive	Poor
1	Very distorted	Very intrusive	Bad

TABLE II
NORWEGIAN TRANSLATION OF THE ORDINAL SCALES USED IN ITU-T P.835.

Rating	Speech	Noise	Overall quality
5	Ikke forvrengt	Ikke hørbar	Veldig god
4	Litt forvrengt	Hørbar, men ikke påtrengende	God
3	Ganske forvrengt	Litt påtrengende	Middels
2	Betydelig forvrengt	Påtrengende	Dårlig
1	Voldsomt forvrengt	Veldig påtrengende	Veldig dårlig

Three different signal to noise ratios (SNRs) were tested for both noise types; 0 dB, 10 dB, and 20 dB. Each combination of noise type, SNR, and enhancement (including unenhanced clips) was tested twice for different sentences, giving 36 sentences per test subject. As the subjects were asked to rate the speech, noise, and overall quality of each sentence, each subject made a total of 108 evaluations. The sound playback and the test environment was the same as in the speech intelligibility test described in Section II-C.

All participants were given an instruction before starting the test and they were allowed to adjust the sound volume to their preferred level. They were also presented examples of the sounds to be used in the test. These examples were randomly taken from all the available sentences, and they were presented to give the test participants some idea of what to expect during the test.

Since it is not certain that the rating scale used has equal steps size between all ratings (i.e. it is not necessary the case that the size of the quality change going from 5 to 4 is the same as when going from 2 to 1), an ordinal scale analysis was

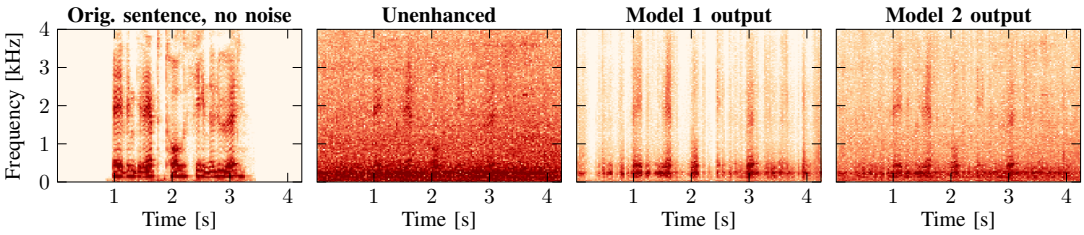


Fig. 2. Spectrograms of the utterance *"Eivind grep tolv fine luer"*, before and after added noise from crossroad traffic at SNR $= -5$ dB, and after enhancement by Model 1 and 2. Each spectrogram is plotted with a dynamic range of 50 dB.

performed to evaluate the results. A cumulative link model (*clm*) from the *ordinal* package [46] in R [47] was used to determine if the models were significantly different from the reference without SE.

E. Test subjects

The speech recognition test was performed by 12 persons, from 40 to 66 years of age (mean value 53.1). These individuals were a subset of the 15 participants from the listening test in our previous study [17]. It is assumed that the learning effect is large for the SRT test, therefore we used the same participants as last time to reduce the time needed for training.

23 persons attended the speech quality test, 8 females and 15 males, from 38 to 74 years of age (mean value 54.7). None of the listeners had performed any subjective listening tests within the last three months.

F. Objective measures

While the subjective evaluations described in Sections II-D–II-E give us the ground truth, it is still interesting to compare these results with those of objective measures. This comparison gives us more information about the reliability of the tested objective measures for DNN-based SE systems. In this work, we calculated the intelligibility measure STOI using the STOI reference code [10] and the quality measure POLQA using the implementation in the software Voice Quality Testing by GL Communications Inc. [48]. Even though PESQ has previously been used as an objective measure for the speech quality of DNN-based SE systems [5], [6], we chose to evaluate its successor POLQA due to licensing rights.

The STOI measures were calculated using the same files as in the speech intelligibility test, with SNRs from -36 dB to 10 dB in 2 dB steps. As a preprocessing step before the STOI calculations, the reference clips and degraded clips were upsampled from 8 kHz to 10 kHz. The POLQA measures were calculated from the same files used in the speech quality test, namely with SNRs of 0 dB, 10 dB, and 20 dB. The POLQA scores were calculated with the High Accuracy and Level Alignment modes activated.

III. Results

A. DNN output

The most basic way to analyze model performance is by investigating the error between the target output and the actual

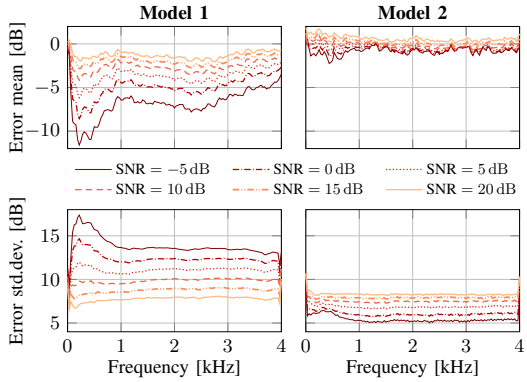


Fig. 3. Statistics for the outputs of the two models, calculated from the difference between the models' target outputs in dB and the actual outputs in dB

output. Figure 3 shows the mean and the standard deviation of the error (the difference between the target output in dB and the actual output in dB) for both models at various SNRs of the input. These statistics were calculated over each frame of the validation set. Frames where the speech signal was silent are excluded from these statistics. This means that the leftover noise shown in 2 during non-speech periods is not included in the error analyses.

We find that Model 2 generally hits its target much better (less biased and with lesser spread) than Model 1 does. This does not necessarily tell us that Model 2 outperforms Model 1 as a SE system, only that it is better at achieving its given task. We also see that Model 1 has a large negative mean error that increases with decreasing SNRs. This shows that its predicted "enhanced" output is higher than the noise-free target output, which indicates that there is still quite a lot of noise left in the output, and that this becomes increasingly true with worsening SNR. For Model 2 the statistics depend less on SNR, indicating that the Model 2 task difficulty is more similar for low and high SNRs than it was for Model 1. Indeed, the standard deviation results show the opposite behaviour with respect to SNR as the Model 1 results did. Model 2 shows less spread (i.e., performs its task with higher accuracy) at lower SNRs. Although this might seem counter-intuitive at first (a SE

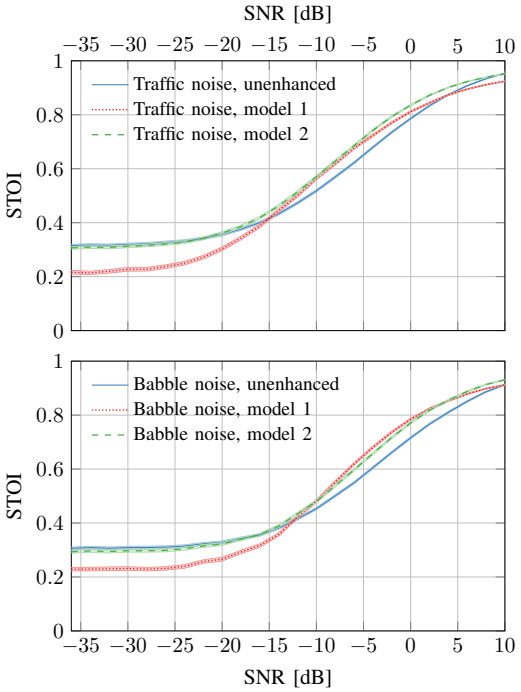


Fig. 4. STOI scores for crossroad traffic noise (upper) and cafeteria babble noise (lower). The lines indicate mean scores, and the shaded areas indicate approximate confidence intervals for the mean scores.

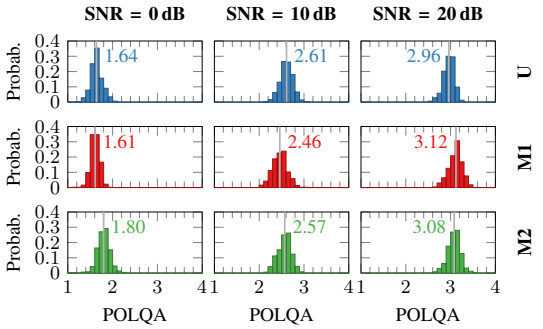


Fig. 5. Histograms of the relative probability of POLQA scores for clips with crossroad traffic noise over three different SNRs. The clips were either unenhanced (U), or enhanced with Model 1 (M1) or Model 2 (M2). The vertical gray lines and numbers represent median values.

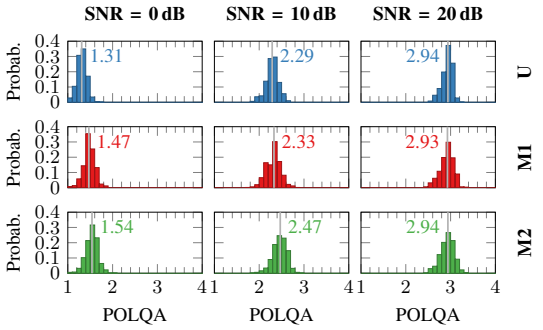


Fig. 6. Histograms of the relative probability of POLQA scores for clips with cafeteria babble noise over three different SNRs. The clips were either unenhanced (U), or enhanced with Model 1 (M1) or Model 2 (M2). The vertical gray lines and numbers represent median values.

model is generally not expected to do better at worse SNR), it makes sense from the perspective that in a situation with a lot of noise, it is easier for this noise to be identified and as such easier to be reduced by 10 dB.

B. Objective measures

As described in Section II-B, 500 clips were available from each combination of SNR, noise, and enhancement, i.e., one clip for each of the 500 original clean speech clips. Thus, we could use these various clips to calculate statistics for STOI and POLQA scores for each of these combinations.

The mean values of the STOI scores are shown as lines for each type of noise and enhancement in Figure 4. Additionally, as the STOI scores of the 500 clips for each combination of SNR, noise, and enhancement were approximately normally distributed, we calculated approximate confidence intervals for these mean values, which are also shown in Figure 4. Due to the high number of clips, the confidence intervals of the various enhancements are quite small and seldom overlap with the means of the other enhancements. Thus, the STOI values unambiguously rank the three enhancements for most SNRs.

The smaller number of SNRs where we calculated POLQA scores allows us to show the scores' distribution in more detail, through the histograms in Fig. 5 and Fig. 6. The median scores are shown as lines together with the median value.

Even if the distribution of the POLQA scores differ between the models with varying skewness and variance, the statistical analysis of the differences was performed using a two sample t-test. The t-test assumes normally distributed data, but it has been shown that for large sample sizes, the t-test might be more robust than the non-parametric tests when the data are a continuous variable [49]. While the mean value might not be the best descriptor for the data, the test does gives a good indication of whether the results differ or not. Note that the median has been used in the illustration if Fig. 5 and Fig. 6 as this is a slightly better descriptor for skewed data. Table III shows the results from the test performed with the function *t.test* in R. An F-test to compare variances was also performed (not shown) and used to decide if pooled variance should be used in the t-test.

C. Subjective speech quality

The results from the speech quality test are illustrated in Fig. 7 and Fig. 8. For more details about the statistical analysis the reader is referred to the supplementary material provided online. The setup for each figure is the same,

TABLE III
RESULTS FROM TWO-SAMPLE T-TEST PERFORMED ON THE POLQA SCORES FOR THE UNENHANCED SIGNAL (U), MODEL 1 (M1) AND MODEL 2 (M2). THE CONFIDENCE INTERVAL (95 % CI) MEANS THE CHANGE IN MEAN POLQA SCORE FOR THE MODELS BEING COMPARED.

Noise	SNR	Comparison	p-value	t	df	95 % CI	
Traffic	0 dB	U→M1	< .001	−4.8772	996	[−0.021	−0.050]
		U→M2	< .001	18.641	994.49	[0.139	0.172]
		M1→M2	< .001	25.336	998	[0.177	0.206]
	10 dB	U→M1	< .001	−15.591	998	[−0.168	−0.131]
		U→M2	< .001	−4.824	994.51	[−0.061	−0.026]
		M1→M2	< .001	10.752	998	[0.086	0.125]
	20 dB	U→M1	< .001	17.121	998	[0.129	0.163]
		U→M2	< .001	14.017	998	[0.101	0.134]
		M1→M2	.002	−3.1375	996.94	[−0.047	−0.011]
Babble	0 dB	U→M1	< .001	22.321	996	[0.155	0.184]
		U→M2	< .001	28.238	996	[0.209	0.241]
		M1→M2	< .001	6.5668	991.34	[0.039	0.072]
	10 dB	U→M1	.003	2.9975	998	[0.010	0.048]
		U→M2	< .001	17.891	998	[0.151	0.189]
		M1→M2	< .001	13.84	996.28	[0.121	0.161]
	20 dB	U→M1	.1528	−1.4309	998	[−0.029	0.005]
		U→M2	.8311	−0.213 34	998	[−0.019	0.015]
		M1→M2	.2807	1.0793	998	[−0.009	0.030]

presenting the different quality assessments horizontally, and different SNRs vertically. The bins consist of three groups; the unenhanced reference, Model 1, and Model 2. Each plot also indicates the significance and the direction of the change in score when going from the unenhanced signal (U) to the DNN models (M1: Model 1, M2: Model 2), as well as similarly indicating the change when going from M1 to M2. The changes' significance is indicated by asterisks, and the changes' direction is indicated with arrows. We cannot show the changes' magnitude, as the statistical test we used does not provide this information.

Both models have a negative effect on the quality of the *speech*. All the tested situations have a significant shift in the negative direction, i.e. the speech is more distorted. However, we can see from the M1→M2 comparison that Model 2 does not distort the speech as much as Model 1. This improvement in speech quality from M1 to M2 is significant ($p < .01$).

The *noise* is reduced for both models and all cases except 20 dB SNR have significant differences. For 20 dB SNR, the noise is generally evaluated as "noticeable, but not intrusive".

The *overall* quality results are more mixed. Model 1 does significantly worse for 10 dB and 20 dB SNR for both noise types, and does not have any significant difference for 0 dB SNR. The quality for the latter is not good, however, with score one ("very bad") as the most probable outcome. Model 2, on the other hand, does not have any significant differences in *overall* quality, except for 0 dB SNR with traffic noise, where there is a significant *positive* effect. The overall quality shifts from approximately equal probability for score one and two, to a most probable outcome at score two. Model 2 performs significantly better in all overall quality scores compared to Model 1 ($p < .05$).

D. Speech recognition threshold

The results from the speech recognition test are presented in Fig. 9. Each line represents results from one test subject. We should point out that the "old" reference data from our previous study [17] are similar to the ones in this study. Comparing the two reference results, using an Wilcoxon rank sum test (also known as an independent two-group Mann-Whitney U test), did not show any significant difference (Median $U_{\text{old ref}} = -9.07$ dB ($n_1 = 15$), Median $U_{\text{new ref}} = -9.14$ dB ($n_2 = 12$), $W = 89$, $p = .98$).

All the differences between the reference and the models were tested using a Wilcoxon signed rank test. Table IV shows the test statistics, and also show that all differences are significant ($p < .05$). The median and confidence interval values have been calculated using Hodges-Lehman estimators.

We also compared the two models, and the results can be seen at the bottom of Table IV. The difference between the results for the traffic noise was compared using a Wilcoxon rank sum test since the two data sets had different number of samples. For the babble noise a Wilcoxon signed rank test was used. Again, Model 2 performs significantly better than Model 1 for both noise types. The estimated improvement of the SRT from M1 to M2 is 3.0 dB for traffic noise, and 1.9 dB for babble noise.

IV. DISCUSSION

Model 1 and Model 2 were given different tasks. Where Model 1 was trained to remove noise, Model 2 was trained to only reduce noise such as to improve the SNR by 10 dB.

For both models, we used the noisy phase of the original signal during speech synthetization. Such a noisy phase may be expected to be better suited to the "less noisy" signal (from Model 2) than the "clean" signal (from Model 1) as the former

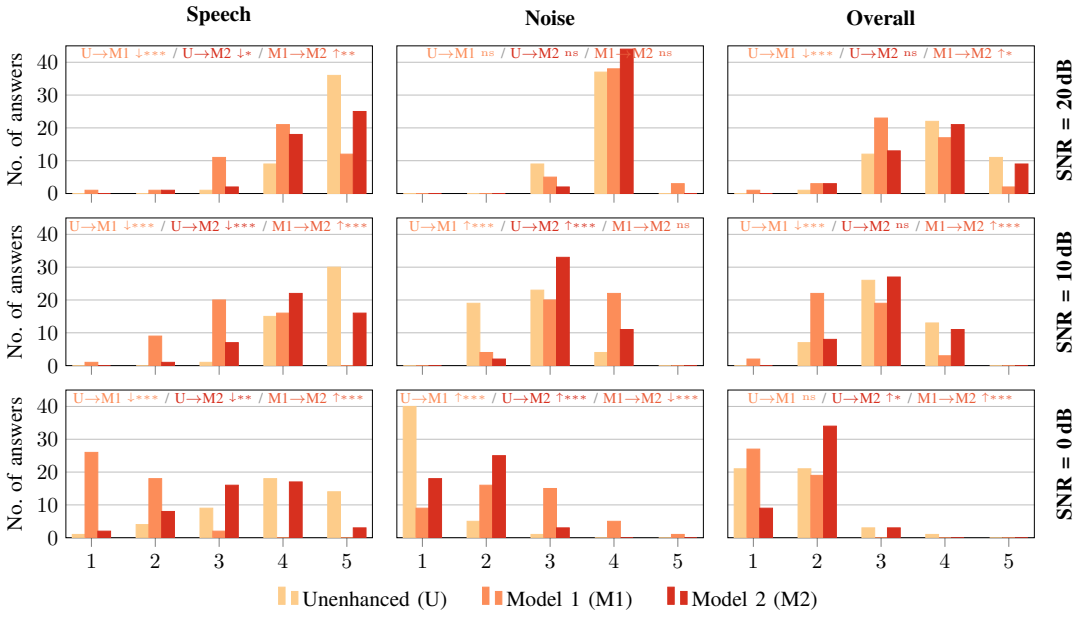


Fig. 7. Speech quality results for the speech (left), noise (middle) and overall (right) evaluation from the ITU-T P.835. The results are from the crossroad traffic noise at three different SNRs; 0 dB (lower), 10 dB (middle), and 20 dB (upper). U, M1, and M2 represent unenhanced, Model 1, and Model 2 respectively in the notation. Three stars (***) indicate $p < .001$, two stars (**) indicate $p < .01$, one star (*) indicates $p < .05$, and *ns* means "not significant". The arrows beside the stars indicate the direction of change.

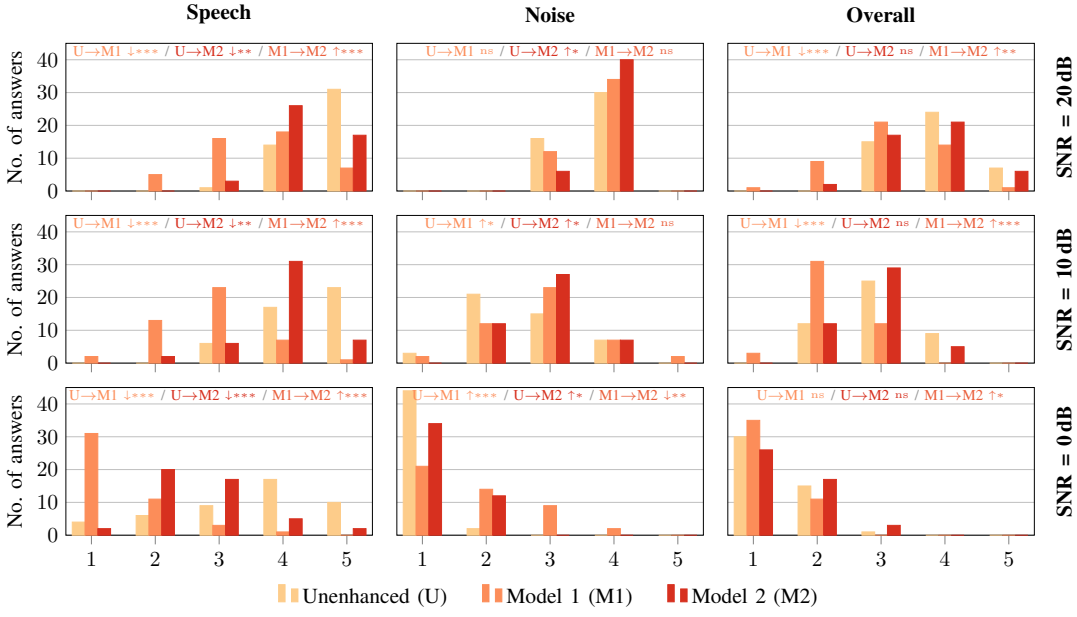


Fig. 8. Speech quality results for the speech (left), noise (middle) and overall (right) evaluation from the ITU-T P.835. The results are from the cafeteria babble noise at three different SNRs; 0 dB (lower), 10 dB (middle), and 20 dB (upper). U, M1, and M2 represent unenhanced, Model 1, and Model 2 respectively in the notation. Three stars (***) indicate $p < .001$, two stars (**) indicate $p < .01$, one star (*) indicates $p < .05$, and *ns* means "not significant". The arrows beside the stars indicate the direction of change.

TABLE IV
SPEECH RECOGNITION THRESHOLD STATISTICS FROM THE ANALYSIS OF THE RESULTS.

Noise	Comparison	n	p-value	V	Median	95 % CI
Traffic	$U_{old} \rightarrow$ M1	15	$< .001$	120	3.9 dB	[3.2, 4.8]
Traffic	U→M2	12	.002	75	1.4 dB	[0.7, 2.0]
Babble	U→M1	12	$< .001$	78	2.4 dB	[1.7, 3.2]
Babble	U→M2	12	.01	70	0.6 dB	[0.2, 1.1]
Traffic	M1→M2	15(12)	$< .001$	161^{+}	−3.0 dB	[−3.9, −1.6]
Babble	M1→M2	12	.002	75	−1.9 dB	[−2.7, −1.1]

$^{+}$: Comparison was done with Wilcoxon rank sum test since the data sets had different number of samples. The number is the observed rank sum W.

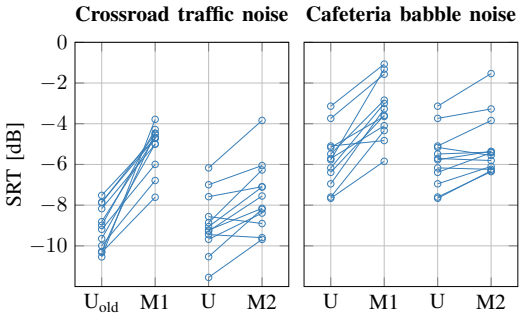


Fig. 9. Speech recognition threshold results for traffic noise and babble. The connected lines represent results from each of the test subjects from unenhanced clips (U) to clips enhanced Model 1 (M1) or Model 2 (M2). For crossroad traffic the results (U_{old} and M1) are taken from our previous study [17].

is closer to the original input from which the noisy phase was taken. Thus, different performance could possibly be the result of better/worse suitability with respect to the speech synthetization process. This in itself would be an advantage of the approach taken in Model 2: After all, the noisy phase is always readily available, whereas a clean phase would have to be approximated. However, the mean and standard deviation results presented in Figure 3 show that there is more going on. First of all, from the rather large standard deviations obtained for Model 1, one can easily argue that the resulting signal is far from "clean", and as such a clean phase won't be optimal either. Also, Model 2 performs better at its given task than Model 1: The fact that the standard deviation of the difference between targeted and obtained output is smaller, shows that the model is more accurate at reducing noise rather than Model 1 is at removing it. There is also a marked lower dependence on SNR, and the model is actually more accurate at noise reduction when the SNR gets worse. This indicates that a DNN-based SE system does indeed have less trouble with reducing noise than with removing it, making the approach worthy of investigation so long as systems aiming to remove noise entirely do not achieve ideal results.

Both models were trained with an equal variety of hyperparameters, and in each case the model with the best STOI score was selected for further subjective testing. This selection method resulted in Model 2 having 3072 nodes per hidden layer, where Model 1 only had 2048 nodes per hidden layer. As such, Model 2 has a larger capacity than Model 1, and one may argue that any differences in the results may be (partly) due to this difference, rather than the difference in noise removal/reduction strategy. However, the statistical results (not reported in this article) akin to those presented in Figure 3 of a model equal to Model 1 but with 3072 nodes per hidden layer, show the same behaviour as the chosen Model 1. During hyperparameter optimization, we also noticed that the lowest MSE obtained for models with a noisy target was generally much lower than for models with a clean target. Given this, we are confident that any performance differences obtained are not due to the different capacities of the model, but due to the different noise cleaning strategies.

As in our previous study [17] the SE did not improve the speech intelligibility. Even if STOI predicted a slight improvement for both models in the SNR range of interest, our subjective evaluation showed that both models did significantly worse than the unenhanced signal. However, Model 2 performed significantly better with respect to speech intelligibility than Model 1 for both noise types, by 3.0 dB and 1.9 dB for traffic and babble noise respectively. Compared to the unenhanced signal, however, it still has an elevated speech recognition threshold.

For the DNN models used in this study, calculated STOI scores were used to select a final model from model candidates over different sets of hyperparameters and different training epochs. The results show that this approach might not be justified as STOI does not seem to be a good predictor in our case. This means that we may have trained other models that could have performed better in our subjective evaluations, but how to identify these models is as of yet an unsolved problem.

Even though we have shown that our selected DNN-based SE systems did not end up actually improving speech intelligibility, we should point out that other authors have trained DNN-based systems that improve intelligibility to human listeners [16], [50], [51]. Our results do by no means provide evidence that DNN-based SE is not a generally promising approach worthy to be further investigated.

In addition to the speech intelligibility test, this study also evaluated the quality of the signal using the ITU-T P.835 recommendation. The results show that the models did not give a general improvement of the *overall* quality of the signal. No significant change to overall quality was found in 7 out of 12 comparisons of unenhanced and enhanced signals, and

Model 1 did actually significantly ($p < .001$) reduce the overall quality in four of the six tests performed. The only exception was for traffic noise at 0 dB SNR, where Model 2 did significantly ($p < .01$) better than the unenhanced signal.

For the evaluation of the quality of the *noise* separately, the results were as expected. The models were trained to reduce the background noise, and the results verify that they achieve this in 9 out of 12 comparisons. Only the situation with the highest SNR (20 dB), where the noise is already rated as "noticeable but not intrusive", does not show significant improvement by both models. (This comes as no surprise, as it is difficult and arguably unnecessary to improve upon a situation that already does not bother listeners.) Note that the highest noise score, "not noticeable", is almost never used. This may indicate that the step size from score four to five on the noise scale is large, and that it is difficult to show minor improvements of low-noise signals on this scale.

Another observation is that Model 2 performs similarly to Model 1 with respect to noise reduction, except at the lowest SNR (0 dB). This is surprising, since Model 2 does not try to remove the noise, only reduce it. It is, however, supported by the fact that a SNR of 30 dB often is referred to as "effective clean speech", and that people have little benefit of improving the SNR beyond 20 dB. This suggests that it might be beneficial to use variable training targets, with little noise reduction for the signals with high SNR, and progressively more reduction as the SNR gets worse. A common training target at 20 dB SNR could be a possible solution.

The evaluation of the *speech* also comply with the results from previous studies on noise reduction. Reducing noise will, in most cases, also add distortion to the speech signal. While Model 2 does perform better than Model 1 in all cases, it still does add distortion to the speech.

Objective POLQA scores were compared to the overall quality results from the subjective test to see if similar traits could be found. The general impression is, however, that POLQA does not predict the overall quality results from the ITU-T P.835 test. Even if we found significant degradation in quality for Model 1 compared to the unenhanced signal, POLQA did not show a consistent correlation. The POLQA scores were, in general, very similar within each SNR, and the largest difference found was below 0.25. Even if this is more than the theoretical accuracy for POLQA [52], such a small difference would be very difficult to detect in a subjective test. The subjective results does, however, show a significant degradation of the overall quality for Model 1, while POLQA actually shows a minor improvement in half of these situations.

Since the ITU-T P.835 recommendation was not available in a Norwegian version, the quality assessment scales were translated for this study. During the pilot test it was revealed that the initial translation was confusing for the test subjects. Several participants found it hard to differentiate between the noise being "slightly noticeable" and "noticeable but not intrusive". To solve this, we used a slightly different wording, closer to the French version of the recommendation [45]. Hence it might be difficult to compare the noise scores in this paper with other results performed with the English scale. The translation of the overall score labels might also affect the (lack of) correlation with POLQA, but this minor textual change to the scale cannot explain why the POLQA scores and the subjective results are opposite for many of the tested situations.

Another limitation of the study is the spoken material used in the test. All the sentences used, both for the intelligibility and quality test, were uttered by the same male speaker. Strictly speaking, this means that the validity of the results are limited to this speaker, and it might be possible that the models could perform better for other speakers.

Similarly to our previous study [17] the sampling frequency used was 8 kHz. This might affect the results since much high-frequency information that might be important both for speech intelligibility and quality assessment are lost. It is, however, not obvious that an increased sampling frequency would have affected the comparisons in this study since they were all done using the same sampling frequency.

In this study two different background noises were used, traffic noise from a busy crossroad and cafeteria babble. The results showed similar improvements for the two noise types, but it is possible that other types of noise could have given different results. The SRT results are otherwise in accordance with what we expect; it is more difficult to understand speech in babble noise than in traffic noise.

Another possible bias is the effect of hearing loss. The average age of the test subjects was relatively high, hence it is expected that age-related hearing loss could be a problem. None of the participants reported any problems with their hearing, or wore hearing aids, but this does not mean that they do not have an elevated hearing threshold. Such elevation could have affected the results, especially the speech intelligibility, which is known to deteriorate with increasing hearing loss. Since all the comparisons were done within each subject, it is expected that an improvement (or deterioration) of the signal would affect both those with normal hearing and those with hearing loss. It is, however, possible that a speech enhancement is perceived differently for individuals with or without hearing loss.

V. Conclusion

In this study, we compared two similar speech enhancement systems based on deep neural networks. The first system, Model 1, was trained with the target of removing all noise from a noisy speech signal, as was done in previous studies [5], [6], [17]. The second system, Model 2, was trained with the target of improving the noisy signal's signal-to-noise ratio by 10 dB.

A subjective evaluation of speech quality in terms of speech degradation, noise intrusiveness, and overall quality showed some interesting similarities and differences between the two models. From the evaluation of overall quality, Model 2 represents a significant improvement to Model 1 in all six situations tested. Both models significantly reduced the noise intrusiveness except at the highest SNR of 20 dB, with Model 1 outperforming Model 2 only at the lowest SNR of 0 dB. While both models significantly distort the speech at all SNRs, Model 2, with its less agressive training target, distorts speech

to a significantly smaller degree than Model 1 at all SNRs. This reduction in distortion may be the reason why Model 2 outperforms Model 1 by 2–3 dB in a subjective evaluation of speech intelligibility in terms of the speech recognition threshold.

For these reasons, we believe that using less aggressive training targets in DNN-based SE systems, along the lines of our Model 2, is a promising approach that warrants further investigation. However, we must point out that if we compare our subjective evaluation results for the noisy speech enhanced by Model 2 and the unenhanced noisy speech, we find that Model 2 does not perform a general improvement to the signal. Model 2 actually degrades the speech intelligibility slightly, raising the speech recognition threshold by around 1 dB. It however did make a significant improvement to the overall quality in one of the six situations tested, while not affecting performance in a statistically significant manner in the other five situations.

In order to train better DNN-based SE systems than the ones presented here, it is absolutely essential to be able to distinguish between a good system and a bad one without having to run a complete subjective evaluation, as these are prohibitively time-consuming. However, our results comparing the subjective evaluations with the objective measures STOI and POLQA indicate that these measures are not appropriate for this purpose. We found that the STOI results predicted significant improvements in intelligibility for our DNN-based SE systems while the subjective evaluations found significant reductions. We also found that the weak changes in POLQA scores failed to predict the significant changes in speech quality found by the subjective evaluations. Therefore, we must advise against solely using STOI and/or POLQA to evaluate DNN-based SE systems, either for the purpose of choosing which trained model candidate to proceed with, or for the purpose of evaluating the final system in the place of a subjective evaluation.

The studied systems are relatively simple implementations of DNN-based SE. As such, their speech enhancing ability is limited, even as indicated by objective measures. However, there is no reason to assume that there will not also be a mismatch between objective and subjective results in better and/or more complicated DNN-based SE systems. Indeed, similar mismatches have also been found elsewhere [16], [18].

Thus, we believe that we have pointed out an important issue that impedes progress for DNN-based SE systems for direct human applications like in telecommunication and hearing assistive devices. To resolve this issue, we believe that it is essential to identify or develop an objective measure that correlates well with intelligibility and/or quality even for channels with the complex nonlinear degradations that processing with a DNN-based SE system can cause. A dedicated study on this topic should be carried out.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[2] Z.-Q. Wang and D. Wang, "A Joint Training Framework for Robust Automatic Speech Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, pp. 1–11, 2016.

[3] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, 2016.

[4] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Interspeech 2014*, Singapore, 2014, pp. 616–620.

[5] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2014.

[6] ——, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[8] F. Chollet, *Deep Learning with Python*. Shelter Island, New York: Manning Publications Co, 2017.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[10] C. H. Taal. STOI – Short-Time Objective Intelligibility Measure. [Online]. Available: http://www.ceestaal.nl/code/

[11] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, ITU-T Recommendation P.862, 2001.

[12] "Perceptual objective listening quality assessment," International Telecommunication Union, ITU-T Recommendation P.863, 2014.

[13] "Recommendation P.862," https://www.itu.int/rec/T-REC-P.862.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An evaluation of objective quality measures for speech intelligibility prediction." in *Interspeech 2009*, 2009, pp. 1947–1950.

[15] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[16] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1660–1669, 2015.

[17] F. B. Gelderblom, T. V. Tronstad, and E. M. Viggen, "Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement," in *Interspeech 2017*. ISCA, 2017, pp. 1968–1972.

[18] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 153–167, 2017.

[19] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *2014 12th Int. Conf. on Signal Process. (ICSP)*. IEEE, 2014, pp. 473–477.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Cross-language transfer learning for deep neural network based speech enhancement," in *9th Int. Sym. on Chinese Spoken Lang. Process., 2014. ISCSLP-14*. Singapore: IEEE, 2014, pp. 336–340.

[21] X. Xiao, S. Zhao, D. H. Ha Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, 2016.

[22] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *9th Int. Symp. on Chinese Spoken Lang. Process. (ISCSLP)*. IEEE, 2014, pp. 250–254.

[23] ——, "Deep neural network based speech separation for robust speech recognition," in *2014 12th Int. Conf. on Signal Processing (ICSP)*. IEEE, 2014, pp. 532–536.

[24] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Commun.*, vol. 95, pp. 28–39, 2017.

[25] A. Kumar and D. Florencio, "Speech Enhancement in Multiple-Noise Conditions Using Deep Neural Networks," in *Interspeech 2016*, San Francisco, USA, 2016, pp. 3738–3742.

[26] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural Speech Enhancement using Deep Neural Networks by Maximizing a Short-Time Objective Intelligibility Measure," *arXiv:1802.00604 [cs, eess]*, 2018.

[27] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually Guided Speech Enhancement Using Deep Neural Networks," in *ICASSP 2018*. Calgary: IEEE, 2018, pp. 5074–5078.

[28] H. Zhang, X. Zhang, and G. Gao, "Training Supervised Speech Separation System To Improve STOI and PESQ Directly," in *ICASSP*, 2018, pp. 5374–5378.

[29] "Methods for Calculation of the Speech Intelligibility Index," American National Standards Institute, Tech. Rep. ANSI S3.5-1997, 1997.

[30] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective"signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[31] S. S. Stevens, "The Measurement of Loudness," *J. Acoust. Soc. Am.*, vol. 27, no. 5, pp. 815–829, 1955.

[32] A. MacPherson and M. A. Akeroyd, "Variations in the Slope of the Psychometric Functions for Speech Intelligibility: A Systematic Survey," *Trends Hear.*, vol. 18, p. 233121651453772, 2014.

[33] F. Chollet, "Keras," GitHub, 2015.

[34] Nasjonalbiblioteket, "NB Tale - a basic acoustic phonetic speech database for Norwegian," 2015.

[35] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, p. 3387, 2009.

[36] D. Pearce, H.-G. Hirsch, and others, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." in *Interspeech 2000*, 2000, pp. 29–32.

[37] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247 – 251, 1993.

[38] Guoning Hu, "100 Nonspeech Sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[39] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, 2018.

[40] J. Øygarden, "Norwegian speech audiometry," Ph.D. dissertation, Norwegian University of Science and Technology, 2009.

[41] N. Prins and F. Kingdom, "Palamedes: Matlab routines for analyzing psychophysical data." http://www.palamedestoolbox.org, 2009.

[42] The MathWorks, Inc., *MATLAB R2017a*. Massachusetts, United States: Natick, 2017, MATLAB version 9.2.0.556344. [Online]. Available: http://www.mathworks.com

[43] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," 2003.

[44] J. Ramsgaard and S. V. Legarth, "Listening test on headset recordings applying the ITU-T P.835 with trained listeners – results from main systems under test," SenseLab, Tech. Rep. SenseLab 006-14(2), 2014.

[45] ITU-T P.835, "Recommendation P.835 (2003) Erratum 1 (05/08)," 2008.

[46] R. H. B. Christensen. (2015) ordinal–Regression models for ordinal data. R package version 2015.6-28. [Online]. Available: http://www.cran.r-project.org/package=ordinal/

[47] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017, R version 3.4.2 (2017-09-28). [Online]. Available: https://www.R-project.org/

[48] G. C. Inc. Voice Quality Testing (VQT) Software (POLQA, PESQ). [Online]. Available: https://www.gl.com/voice-quality-testing-pesq-polqa.html

[49] M. W. Fagerland, "T-tests, non-parametric tests, and large studies—a paradox of statistical practice?" *BMC Med Res Methodol*, vol. 12, no. 1, 2012.

[50] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. Wang, "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4230–4239, 2017.

[51] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2604–2612, 2016.

[52] "Application guide POLQA," HEAD acoustics, Tech. Rep. Rev0 (3/2012), 2012.

Femke B. Gelderblom received a BSc degree in Applied Physics, and a MSc degree in Biomedical Engineering from Delft University of Technology, the Netherlands, in 2012. Since then, she has been working as a research scientist at the Acoustics group of SINTEF Digital in Trondheim, Norway. She is currently also working towards a PhD degree at the Signal Processing group of the Norwegian University of Science and Technology (NTNU), under supervision of Tor Andre Myrvoll and Torbjørn Svendsen. Her main research interests include speech enhancement, deep learning, and microphone arrays.

Tron Vedul Tronstad is a research scientist at the Acoustics group at SINTEF Digital in Trondheim, Norway. He received a MSc degree in acoustics from the Department of Electronics and Telecommunications at the Norwegian University of Science and Technology (NTNU) in 2007. He received a PhD from the Department of Electronic Systems at the same university in 2018. His main research topics revolves around hearing and hearing damage.

Erlend Magnus Viggen received an MSc degree in applied physics in 2009 and a PhD in acoustics in 2014, both at the Norwegian University of Science and Technology (NTNU). While making most of his contributions to this work, he worked as a research scientist at the Acoustics group of SINTEF Digital in Trondheim, Norway. Currently, Erlend is a post-doc at NTNU, working at the Centre for Innovative Ultrasound Solutions on the topic of ultrasonic logging in petroleum wells. His current research interests include physical acoustics, computational acoustics, and machine learning.

Paper III

F. B. Gelderblom, Y. Liu, J. Kvam and T. A. Myrvoll, 'Synthetic Data For DNN-Based DOA Estimation of Indoor Speech,' in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada: IEEE, Jun. 2021, pp. 4390–4394

SYNTHETIC DATA FOR DNN-BASED DOA ESTIMATION OF INDOOR SPEECH

Femke B. Gelderblom[⋆]*, Yi Liu*[†]*, Johannes Kvam*[†]*, Tor Andre Myrvoll*[⋆]

[⋆]NTNU & SINTEF, Norway, [†]SINTEF, Norway

ABSTRACT

This paper investigates the use of different room impulse response (RIR) simulation methods for synthesizing training data for deep neural network-based direction of arrival (DOA) estimation of speech in reverberant rooms.

Different sets of *synthetic* RIRs are obtained using the image source method (ISM) and more advanced methods including diffuse reflections and/or source directivity. Multi-layer perceptron (MLP) deep neural network (DNN) models are trained on generalized cross correlation (GCC) features extracted for each set. Finally, models are tested on features obtained from *measured* RIRs.

This study shows the importance of training with RIRs from directive sources, as resultant DOA models achieved up to 51% error reduction compared to the steered response power with phase transform (SRP-PHAT) baseline (significant with $p << .01$), while models trained with RIRs from omnidirectional sources did worse than the baseline. The performance difference was specifically present when estimating the azimuth of speakers not facing the array directly.

***Index Terms*—** synthetic data, speech source localization, direction of arrival estimation, room impulse response, deep neural network, generalized cross correlation features

1. INTRODUCTION

DNN-based methods are nowadays successfully applied to many different tasks in the field of speech processing. For training such methods, there are large datasets available, containing annotated single microphone recordings of clean speech. These datasets can be converted into multichannel datasets for microphone array processing by convolving the clean speech with recorded room impulse responses (RIRs) specific for each array element and acoustic setting.

However, learning-based methods can only be expected to be widely applicable in realistic settings if they are trained for exactly that. This issue is two-fold: first of all, to ensure results apply to a wide range of rooms of varying acoustical characteristics, the training set needs to contain a similar variety [1], and secondly, the training data must approach reality as much as possible.

While recorded RIRs are a direct reflection of reality, it quickly becomes too difficult or expensive to record a sufficient number of RIRs from many different environments. Instead models can be trained on single channel recordings augmented with synthetic RIR data.

Here it is common to rely on the relatively simple image source method (ISM) room impulse response (RIR) simulation technique [2], where scattering effects that cause the late reflections of the diffuse field are ignored for simplicity. Additionally, all sources are assumed to behave in an omnidirectional manner, while a speaking person is a directive source.

This paper therefore investigates how more advanced RIR simulation methods can affect final model performance on real data. We have chosen to do this through the DOA estimation task, because of its central role in multi-channel speech processing. The ability to discriminate on where speech originates from is crucial for applications like multi-channel speech enhancement, speaker identification and automatic speech recognition.

Classic approaches to DOA estimation include multiple signal classification (MUSIC) [3], the least squares (LS) method [4], multi-channel cross correlation (MCCC) [5], and the steered response power with phase transform (SRP-PHAT) [6]. A main challenge is the multipath propagation effect where microphone sensors not only receive the direct-path signal, but also attenuated signals due to both the specular and diffuse reflections.

Inspired by the success of DNNs in many fields, several such approaches have been proposed for sound/speech source localisation (SSL) [7, 8, 9, 10, 11, 12, 13, 14].

Research based on training data generated from measured RIRs is automatically constricted to a severely limited number of rooms [7, 8]. Others rely on the simulation of just one or two acoustical environments [9, 10]. Xiao *et al.* and Perotin *et al.* simulated more varied data for DOA estimation of speech [11, 12, 13], but they, as is common practise, relied on ISM with omnidirectional sources for RIR simulation.

Only recently have researchers attempted to improve deep learning model performance in speech processing tasks, by improving the quality of the RIRs used for synthesizing data. Tang *et al.* found significant performance increases on an automatic speech recognition and keyword spotting task in [15] by using an acoustic simulation method that includes diffuse reflections. Using the same method, Tang *et al.* also observed improved performance at a DOA estimation task [14].

In this study we further investigate the effect of RIR sim-

ulation methods on final DOA model performance. Our study is unique in that we are, as far as we know, the first to investigate the effect of simulating speakers as directive sources. Like Tang *et al.* we also study the effect of diffuse reflections, but we rely on the GCC speech features and the MLP architecture proposed in [11], instead of ambisonic features and CRNN architecture. We focus only on reverberance (no noise added), and use our own dataset, which includes two test sets that allow us to differentiate between results for speakers looking directly at the array, and the more challenging situation where speakers face the array at a 90° angle.

2. DATA ACQUISITION

2.1. Synthetic RIRs for training

We simulated RIRs with four different simulation methods using the MATLAB package MCRoomSim [16]:

- **ISM-omni**: the basic RIR generated by ISM where sources are modelled as omnidirectional. No scattering and no diffuse field.
- **ISM-dir**: Like ISM-omni, but now sources are modelled as directive speakers, with either an average male or female directivity. No scattering and no diffuse field.
- **WithDiffuse-omni**: An advanced RIR with not just specular reflections, but also a diffuse field due to scattering, where sources are modelled as omnidirectional.
- **WithDiffuse-dir**: Like WithDiffuse-omni, but sources are again modelled as directive speakers.

For each method, 18 000 training and 6000 validation RIRs were simulated from three random source positions in 6000 and 2000 virtual rooms. Each room was randomly configured with parameters drawn from the uniform distributions specified in Table 1, ensuring evenly distributed target DOAs in all directions. The average absorption of a room was determined from the drawn reverberation time with Eyring's [17] algorithm with air absorption taken into account.

Table 1. Details of random virtual room configuration

Item	Parameter	Min.	Max.
Room size	width	3 m	8 m
	length	3 m	10 m
	height	2.5 m	6 m
	RT60	0.2 s	1 s
	scattering coefficient	0	1
Array position	from walls	1 m	-
	from floor	0.6 m	0.9 m
Speaker position	from walls	0.5 m	-
	from floor	1 m	1.8 m
	from array	0.5 m	-
	yaw (directive speakers only)	-180°	180°

2.2. Measured RIRs for testing

To create realistic test data, RIRs were measured manually with a 9-channel circular array (planar) with 4 cm radius, positioned on a table approximately in the middle of a typical rectangular meeting room with dimensions 4.5 x 3.8 x 2.6 m, and $RT60_{1kHz}$ of 0.3. An NTi TalkBox was used to produce the sinusoidal sweeps required for RIR measurements. This loudspeaker has human head-size like dimensions and is specifically designed for human speech measurements.

Of the measured RIRs, 47 where obtained with the speaker facing towards the array (the 'Easy' set), and 107 with the speaker rotated at 90° (the 'Challenging' set). The true DOAs were measured with an uncertainty of $\pm$ 1° at random angles uniformly distributed around the array, at a distance varying between 1 and 2 m (above critical distance).

2.3. Obtaining Speech Features

Our preprocessing steps are inspired by [11], but the specifics differ. We used 'NB Tale', a Norwegian speech database. This database contains circa 19 hours of training data and circa 5 hours of validation data from a total of 380 speakers.

First the speech files were passed through the open source voice activity detector from WebRTC with a hop length of 30 ms, zero minimum silence length and strength 3. They were then convolved with (simulated or measured) RIRs to create a reverberant multichannel speech sample, which was resampled from 48 kHz to 16 kHz. We then selected a random 1 s long segment.

Lastly, GCC vectors with PHAT weighting were obtained for each pair of microphone channels. For our array, the maximum distance between a pair of microphones is 8 cm, which represents a maximum delay of 4 (0.08 m / 340 m/s $\times$ 16 000 Hz) time samples of each GCC vector. Hence, the GCC vector was truncated to the 9 centre time samples for each microphone pair. From the 9 channels, we have 36 possible microphone pairs, giving us 36 GCC vectors. Each of the vectors was scaled so that its max value became 1, and then stacked to obtain a single model input sample.

Due to the random selection of the speech segment, diffuse reflections of earlier speech affect the model input sample, even if vector truncation removes later reflections. This can be seen in Figure 1, which shows examples of the synthetic input training samples for each simulation method, given the same room size, source and array location. Less aggressive truncation did not improve final model performance.

Using the above procedure, we created *synthetic* training and validation sets for each of the RIR simulation methods, with 18 000 training and 6000 validation samples per set. The same procedure was also applied using the two types of *recorded* RIRs to create two *measured* test sets called 'Easy' (speaker facing directly towards the array) and 'Challenging' (speaker at a 90° angle away from the array). The final test sets had 517 'Easy' and 1177 'Challenging' input samples.

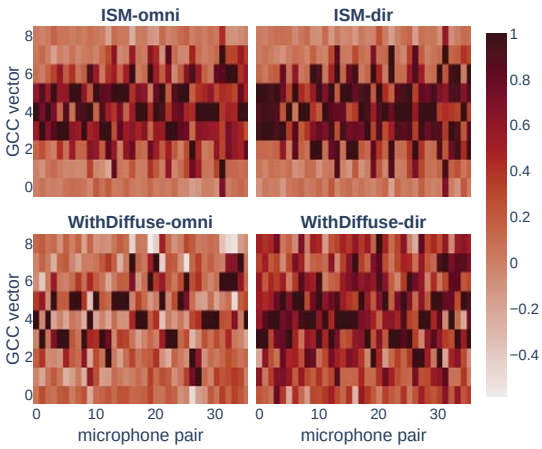


Fig. 1. Examples of the GCC input feature for each method

3. DOA ESTIMATION MODEL

The DOA estimation task is most intuitively formulated as a regression task where the continuous azimuth variable is directly predicted from the input features. However, others have noted advantages from converting the task into classification, where possible azimuths are separated into discrete bins [11, 14]. In this paper, we include both.

For the regression formulation, we investigated two loss functions, which we call the angular mean square error:

$$\mathrm{MSE}_{\angle} = \frac{1}{N}\sum_{n=1}^{N}\Big(\,\mathrm{atan2}\left(\sin\left(\hat{y}-y\right),\cos\left(\hat{y}-y\right)\right)\Big)^2, \tag{1}$$

and the angular mean absolute error:

$$\mathrm{MAE}_{\angle} = \frac{1}{N}\sum_{n=1}^{N}\Big|\,\mathrm{atan2}\left(\sin\left(\hat{y}-y\right),\cos\left(\hat{y}-y\right)\right)\Big|, \tag{2}$$

where $\hat{y}$ and y are the true and estimated DOA respectively, and the $\mathrm{atan2}$ operator computes the arctangent of the element-wise division of its first and second argument, respecting signs of the arguments.

These are based on the general mean squared error (MSE) and mean absolute error (MAE) loss functions, but ensure the calculation is always based on the minimal error between two angles, be it clockwise or anticlockwise. Output layers of both regression formulations were given linear activation.

For the classification formulation, we used the standard categorical crossentropy loss with either 72 (5° per bin) or 360 (1° per bin) classes. Both classification models were given an output layer with softmax activation.

As model we chose the MLP neural network. First a wide hyperparameter search was conducted for all datasets using the tree-structured parzen estimator (TPE) approach [18], to determine a single model topology that worked well for all datasets. This search included varying the number of hidden layers, number of nodes per layer, type of activation, rate of dropout, $\ell 1$ or $\ell 2$ regularization, batch normalization and learning rate for the Adam optimizer.

From this, a general model with 3 hidden layers, each with 3072 hidden nodes and relu activation, was chosen for all datasets and problem formulations. No batch normalisation was applied. A new optimisation process was then started for each combination of the 4 datasets and 4 loss functions. Now only the learning rate and level of dropout was varied to find the best model for each set, to ensure that results would be directly comparable. Classification models converged best with high levels of dropout (circa 0.8), while regression models did best without dropout.

Table 2 shows the MAE results for all model types and all simulation methods, obtained for a validation test set specific for each simulation method. These errors do not reflect real-life performance, but performance on synthetic validation set that was created in the same way as the training set used to train each model. Therefore, the consistently lower MAE for methods with omnidirectional sources merely shows that these tasks are easier to learn, but it is not an indication of how the resulting MLPs will deal with real data.

Table 2. MAE for each method's *synthetic* validation set

	Regression		**Classification**	
	$\mathrm{MSE}_{\angle}$	$\mathrm{MAE}_{\angle}$	1° bins	5° bins
ISM-omni	2.4°	1.8°	1.6°	2.3°
ISM-dir	5.5°	5.0°	4.6°	4.7°
WithDiffuse-omni	2.0°	1.4°	1.1°	2.0°
WithDiffuse-dir	6.3°	4.3°	4.0°	4.4°

4. RESULTS

All final models were tested with the exact same two *measured* test sets ('Easy' and 'Challenging'), and performance was evaluated with MAE for all models (independent of the training loss function used!), to allow for direct comparison. For Table 3, test samples are based on RIRs where the speaker was facing directly towards the array. Table 4 shows the results for RIRs where the speaker faced past the array at a 90° angle. Testing with MSE or accuracy within 5° or 10° instead of MAE resulted in the same trends, and are therefore not included in this paper.

In our application, the variance of the error from the true direction indicates system performance (assuming zero mean error). We therefore apply the Brown-Forsythe statistical test [19], which tests the variance of the distributions without a strong assumption of normality. We report the test's probability results p, for relevant pairs of systems, in Section 5.

Table 3. MAE for the 'Easy' test set, where speakers face directly towards the array

	Regression		Classification	
	$MSE_{\measuredangle}$	$MAE_{\measuredangle}$	1° bins	5° bins
SRP-Phat	1.5°			
ISM-omni	2.2°	2.1°	1.4°	1.3°
ISM-dir	3.0°	2.1°	1.5°	1.5°
WithDiffuse-omni	2.8°	1.1°	1.3°	1.4°
WithDiffuse-dir	3.8°	1.4°	1.1°	**0.9**°

Table 4. MAE for the 'Challenging' test set, where speakers face 90° away from the array

	Regression		Classification	
	$MSE_{\measuredangle}$	$MAE_{\measuredangle}$	1° bins	5° bins
SRP-Phat	16.5°			
ISM-omni	18.2°	18.2°	19.1°	18.8°
ISM-dir	12.7°	11.5°	8.9°	**8.1**°
WithDiffuse-omni	19.7°	19.6°	18.6°	17.9°
WithDiffuse-dir	13.0°	10.5°	9.9°	10.1°

5. DISCUSSION

From Table 3 we observe that for the relatively easy task of finding the correct azimuth of a speaker facing the array, all models are able to estimate the DOA with high accuracy.

The training data simulation method starts to matter when testing with samples where speakers looked past the array, giving increased confounding reflections. In this case (see Table 4) all directional data based MLPs outperformed their omnidirectional equivalents and the SRP-Phat baseline method significantly ($p << .01$). Simulating with directional sources also increased the difficulty of the task given to the SSL method as evident from the increase in validation error (see Table 2). As such, results show that the MLPs were able to learn relevant information from the directional simulations that turned out to be applicable on measured data.

This is crucial given that we found no studies that simulated directive sources to train learning-based SSL models. Also, given the importance of localisation for many other speech processing tasks like speech recognition and speech enhancement, the conclusion may be valid for many other multichannel speech applications.

We observe that for each DNN topology, either the simulation methods ISM-dir or WithDiff-dir leads to the highest performance, and overall the performance difference between the two was insignificant ($p > .01$). Adding a diffuse field when simulating sources as omnidirectional also did not have a significant effect ($p > .01$).

As such, in contrast with [14], we do not find benefit (nor deterioration) from adding the diffuse field. However, this may simply be because the chosen preprocessing steps to generate speech features may have stopped the models from learning relevant information from the diffuse field. We also have to be careful to draw conclusions based on measurements taken in a single meeting room, as its diffuse field is not representative for all meeting rooms.

Observed trends are independent of the choice of loss function and whether the problem is formulated as a regression or classification task. This provides evidence that the obtained differences are indeed due to the different datasets used for training, and not due to effects of biased hyperparameter tuning.

Like others [11, 14], we note that defining the DOA estimation task as a classification task is advantageous as this formulation resulted in our best performing models. Especially the directive training sets contain samples that are too challenging for the network to learn. The regression network with $MSE_{\measuredangle}$ loss penalises large errors harshest, and as such the learning process focuses most on these outliers. The classification networks are on the other end of the spectrum - penalising all predictions outside the target bin equally, and as such, their training focuses on the more informative samples. Additionally, all classification networks required high levels of dropout, indicating that smaller networks may work equally well for this task formulation.

The focus of this study was on the effect of using more advanced RIR simulation techniques for generating better training data, and not on finding the best DOA estimator.

6. CONCLUSION

We synthesized different training sets to train MLP models for a DOA estimation task from 4 different RIR simulation techniques. The model trained on data from RIR simulation techniques with directive sources, achieved up to a 51% lower mean absolute error on a measurement-based test set than the industry standard SRP-PHAT method, while equivalent models trained on the standard image source method with omnidirectional sources performed worse than this baseline.

Results show that, for improved real-life performance, sources should be modelled as directive speakers, rather than omnidirectional sources, especially for the situation where the speaker is not directly looking at the array. This is an important conclusion given the widespread use of simple ISM RIRs, indicating that the complexity of the RIR simulation technique has been undervalued as a source of performance gain for learning-based SSL. We further speculate that the conclusion may hold true for other applications within multichannel speech processing.

7. ACKNOWLEDGMENTS

We thank the Research Council of Norway and Huddly for their support through project '256753 - Meet Easy'.

8. REFERENCES

[1] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 379–383.

[2] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] Yiteng Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[5] Jacob Benesty, Jingdong Chen, and Yiteng Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 509–519, 2004.

[6] Joseph Hector DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown University, Providence, Rhode Island, USA, 2000.

[7] Ryu Takeda and Kazunori Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016, pp. 405–409.

[8] David Diaz-Guerra and Jose R. Beltran, "Direction of Arrival Estimation with Microphone Arrays Using SRP-PHAT and Neural Networks," in *IEEE 10th Sensor Array and Multichannel Signal Processing Workshop*, Sheffield, UK, 2018, pp. 617–621.

[9] Zhaoqiong Huang, Ji Xu, and Jielin Pan, "A regression approach to speech source localization exploiting deep neural network," in *IEEE Fourth International Conference on Multimedia Big Data*, Xi'an, China, 2018, pp. 1–6.

[10] Soumitro Chakrabarty and Emanuel A. P. Habets, "Broadband DOA estimation using Convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York City, USA, 2017, pp. 136–140.

[11] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, 2015, pp. 2814–2818.

[12] Laureline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guerin, "CRNN-based Joint Azimuth and Elevation Localization with the Ambisonics Intensity Vector," in *International Workshop on Acoustic Signal Enhancement*, Tokyo, Japan, 2018, pp. 241–245.

[13] Laureline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guerin, "CRNN-Based Multiple DoA Estimation Using Acoustic Intensity Features for Ambisonics Recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[14] Zhenyu Tang, John D. Kanu, Kevin Hogan, and Dinesh Manocha, "Regression and Classification for Direction-of-Arrival Estimation with Convolutional Recurrent Neural Networks," in *INTERSPEECH*, Graz, Austria, 2019, pp. 654–658.

[15] Zhenyu Tang, Lianwu Chen, Bo Wu, Dong Yu, and Dinesh Manocha, "Improving Reverberant Speech Training Using Diffuse Acoustic Simulation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, 2020, pp. 6969–6973.

[16] Andrew Wabnitz, Nicolas Epain, Craig Jin, and André van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *International Symposium on Room Acoustics*, Melbourne, Australia, 2010.

[17] Carl F. Eyring, "Reverberation time in "Dead" rooms," *The Journal of the Acoustical Society of America*, vol. 1, no. 2A, pp. 168–168, 1930.

[18] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl, "Algorithms for hyper-parameter optimization," in *24th International Conference on Neural Information Processing Systems*, Granada, Spain, 2011, NIPS 2011, pp. 2546–2554.

[19] Morton B. Brown and Alan B. Forsythe, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.

Paper IV

F. B. Gelderblom and T. A. Myrvoll, 'Deep Complex Convolutional Recurrent Network for Multi-Channel Speech Enhancement and Dereverberation,' in *IEEE International Workshop on Machine Learning for Signal Processing*, Gold Coast, Australia: IEEE, Oct. 2021, pp. 1–6

DEEP COMPLEX CONVOLUTIONAL RECURRENT NETWORK FOR MULTI-CHANNEL SPEECH ENHANCEMENT AND DEREVERBERATION

Femke B. Gelderblom and Tor Andre Myrvoll

NTNU & SINTEF, Norway

ABSTRACT

This paper proposes a neural network based system for multi-channel speech enhancement and dereverberation. Speech recorded indoors by a far field microphone, is invariably degraded by noise and reflections. Recent single channel enhancement systems have improved denoising performance, but do not reduce reverberation, which also reduces speech quality and intelligibility. To address this, we propose a deep complex convolution recurrent network (DCCRN) based multi-channel system, with integrated minimum power distortionless response (MPDR) beamformer and weighted prediction error (WPE) preprocessing.

PESQ and STOI performance is evaluated on a test set of room impulse responses and noise samples *recorded* by the same setup. The proposed system shows a statistically significant improvement ($p \ll 0.05$) over competitive systems.

***Index Terms*—** speech enhancement, microphone arrays, deep neural networks, dereverberation, beamforming

1. INTRODUCTION

The field of speech enhancement (SE) has undoubtedly been revolutionized by deep learning techniques. Now that the whole world has been forced to adapt to online meetings at an unseen rate, the topic is also more relevant than ever.

Rapid developments in the related field of automatic speech recognition (ASR) have inspired many source separation and denoising systems. However, over the course of only the past year, Microsoft has organized three SE challenges, where the focus was on enhancing single channel signals specifically for human listeners [1, 2, 3]. Additionally, the challenge ConferencingSpeech 2021 targets multi-channel speech enhancement for video conferencing [4].

Most results of these challenges are not yet available. However, top performing systems of the first deep noise suppression (DNS 2020) challenge, demonstrate remarkable performance increases with respect to removing additive noise from speech recordings.

Isik *et al.* proposed PoCoNet; a 2D UNet (with DenseNet blocks and self-attention) with small kernels [5]. They also utilized a semi-supervised method to increase the amount of training data and investigated the effect of different augmentation techniques. Their proposed system with approximately 50M parameters won first place in the non-real-time track.

Hu *et al.* proposed the deep complex convolution recurrent network (DCCRN) [6]. The DCCRN also follows the UNet structure, but uses complex-valued convolutional encoders and decoders, and LSTMs to model the context dependency. With only 3.7M parameters, the DCCRN models ranked first for the real-time-track and second for the non-real-time track. The lower complexity of this network, combined with the fact that it was trained on less data, while obtaining such competitive performance, makes it an ideal candidate for further research.

However, speech quality and intelligibility is also negatively affected by the presence of reverberance [7, 8]. The DCCRN system does not attempt to remove reverberance at all, and PoCoNet only attempted partial dereverberation.

From the field of multi-channel speech enhancement, we know that there lies a huge potential in relying on multi-channel signals as input, and in applying beamforming techniques [9]. Heymann *et al.* proposed a system where a DNN estimates an ideal binary mask (IBM) to deduce the cross-power spectral densities of the target speech and noise. These are then used for beamforming with a generalized eigenvector (GEV) beamformer [10]. Their system did really well on the CHiMe-3 challenge for robust ASR, but as their network estimates the IBM, and not the target signal, performance is inherently capped. We also observe that, despite its definite merits over earlier single-channel systems, the system proposed by Heymann *et al.* struggles to outperform the single channel DCCRN on our test set, even if we rely on oracle IBM masks (see Table 2 in Section 5). Erdogan *et al.* proposed a similar masked based MVDR system with a spectrum magnitude based loss [11]. However, their final system performance was lower.

In this paper, we therefore propose a far-field multi-channel neural network for simultaneous speech dereverberation and enhancement that combines the recent advancements in single channel speech enhancement for human listeners, with mask based neural beamforming from the domain of multi-channel speech enhancement. We integrated the DCCRN with a minimum power distortionless response (MPDR)

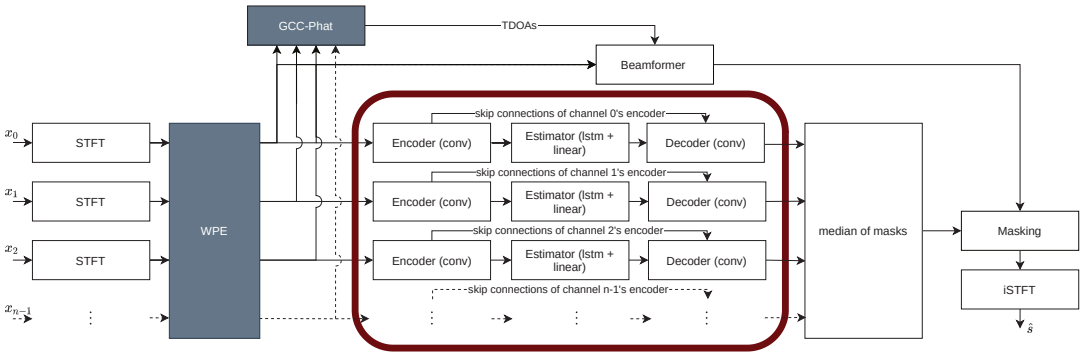


Fig. 1: Overview of the proposed speech enhacement and dereverberation system. The highlighted WPE and GCC-Phat boxes are only employed during inference. The red frame contains all blocks with trainable parameters, where each Encoder-Estimator-Decoder structure represents a single channel DCCRN.

beamformer [12] and added weighted prediction error (WPE) speech dereverberation [13] to the processing pipeline. As such, our system crucially differs from the other mask based beamformering approaches, by relying on a time domain loss, and having complex spectral input features and complex network layers. Furthermore, the usage of the MPDR separates the steering vector estimation from the mask estimation process. The proposed system only requires the corrupted multi-channel speech signal during inference, and as such does not need estimates of noise statistics, or information on microphone layout.

We evaluate the system on two highly realistic test sets. These sets were obtained by combining our own *recorded* multi-channel room impulse responses (RIRs) with clean speech from an open database, and our own multi-channel noise recordings. The latter were recorded by the same array placed at the same location in the same room as where the RIRs were obtained. This setup allows for objective testing with a clean reference signal, while simultaneously avoiding the need for synthetic RIRs that would reduce the realism of the test sets. Furthermore, with this setup we can differentiate between results for speakers looking directly at the array, and the more challenging situation where speakers face the array at a 90° angle.

We compare the system to three state-of-the art baseline systems; i) single channel DCCRN, ii) multi-channel baseline system of the ConferencingSpeech 2021 Challenge, and iii) mask based GEV beamformer with blind analytic normalization postprocessing by Heymann *et al.*

2. THE SYSTEM

2.1. Overview

Figure 1 shows an overview of the proposed system. A multi-channel noisy and reverberant speech signal x is transferred to the frequency domain by a short time Fourier transform (STFT) operation. The resultant signals are fed into the weighted prediction error (WPE) block for deverberation. A DCCRN neural net estimates masks for each channel, where neural network weights are shared across channels (but input is not). All resultant masks are then combined into a single mask using the median operator, because of its resilience to outliers. This mask is then applied to a beamformed result of the dereverberated signal. Lastly, the enhanced signal is taken back to the time domain by an inverse STFT (iSTFT).

The beamformer requires time difference of arrival (TDOA) estimates to obtain an appropriate steering vector. During training, this information is obtained from the known true speaker direction. During the prediction stage, this information is estimated directly from the WPE's output, using generalized cross correlation with phase transform (GCC-PHAT). As such, the final system only requires the corrupted signal as input.

The next subsections provide further processing details.

2.2. Short Time Fourier Transform

Adhering to the original DCCRN paper, we use a Hann window, a FFT length of 512 samples, a window length of 25 ms (400 samples at 16000 Hz) and a hop size of 6.25 ms (100 samples at 16000 Hz) to obtain a complex-valued STFT [6].

2.3. WPE dereverberation

The idea of WPE is to estimate the reverberation tail of the signal and subtract it from the observation with a maximum likelihood approach [13]. We have tested our system with one iteration (using the Nara-WPE implementation [14]), as this has been shown to already provide significant benefit, while multiple iterations quickly become highly time consuming.

2.4. Beamforming

Beamforming is a signal processing technique, where the channels of a multi-channel signal are delayed, weighted, and then combined into a single signal that is steered towards a specific source/direction. Depending on the chosen algorithm, a beamformer can both denoise and dereverberate a multi-channel signal.

One popular beamformer, is the minimum variance distortionless response (MVDR) beamformer. It requires statistical noise characteristics, which are particularly difficult to obtain when the noise is non-stationary as well as mixed with the signal of interest.

One implementation of the MVDR-related algorithm avoids this problem, by deriving the distortionless filter for a specified steering direction that minimizes the mean square output power, and as such only requires the corrupted input signal. Although this implementation is often referred to as an MVDR in the literature, we comply with Van Trees' practice of referring to it as the minimum power distortionless response (MPDR) beamformer for unambiguity [12].

The weights of the MPDR beamformer are obtained as follows:

$$\mathbf{w}_{\mathrm{mpdr}}^{H} = \frac{\mathbf{v}^{H}\mathbf{X}^{-1}}{\mathbf{v}^{H}\mathbf{X}^{-1}\mathbf{v}} \tag{1}$$

where $\mathbf{X}$ is the spectral matrix of the entire input, and $\mathbf{v}$ the steering vector.

When the steering direction is equal to the desired signal direction, the MPDR beamformer reduces to the standard MVDR beamformer [12]. As the target direction is known during training, we effectively train the algorithm with an MVDR beamformer. During inference, the target direction has to be estimated as discussed in Section 2.5.

2.5. GCC-Phat

During inference, one cannot expect the true azimuth of speakers to be available and once the steering vector starts to deviate from the signal vector, the performance between an MPDR and MVDR may differ significantly.

There are many DOA estimation techniques available, both traditional [15, 16], and neural network based [17]. We leave the problem of estimating the azimuth largely outside the scope of this study, but present the results for the final system, both for the ideal situation where the speaker azimuth is known, and for an estimated azimuth using generalized cross correlation with phase transform (GCC-PHAT) [15]. This method allows us to estimate the steering vector without needing to provide the microphone layout.

2.6. DCCRN single channel speech enhancement

The DCCRN single channel SE system was first proposed in [6]. Its goal is to estimate a complex ratio mask (CRM) for the complex-valued STFT. The DCCRN therefore receives both real and imaginary information. This in contrast with SE systems that try to enhance the magnitude of a signal, but rely on the noisy phase.

The DCCRN network can be structured into three parts: the encoder, the estimator and the decoder.

The encoder and decoder contain 6 encoder/decoder blocks each. Each of these blocks consist of a 2D complex convolutional (or deconvolutional) layer, followed by real-valued 2D batch normalization (BN) and leaky ReLU activation. Encoder and decoder blocks (with output channels [32, 64, 128, 128, 256, 256]) are furthermore connected through skip connections.

The encoder extracts high-level features from the input, while the symmetric encoder-decoder architecture ensures that the decoder takes these features (after the estimator stage) back to the same shape as the input. Skip connections between encoder and decoder blocks, make that the noisy input (translated into the corresponding feature spaces), are available during decoding.

At the estimator stage, the network needs to identify the desired signal from the noise, to construct a mask like structure in the encoded feature space. For this, it is important to leverage long-term contexts, which the DCCRN does with LSTM layers. The estimator therefore consist of two real valued LSTM layers (not bidirectional, and each with 256 nodes) followed by a linear layer (1024 nodes). We relied on the polar coordinate masking approach (DCCRN-E).

3. TRAINING

3.1. Setup

We first trained a single channel DCCRN SE model as a pretraining step. This model also functions as one of the reference systems. We then initialize the multi-channel system with the obtained weights.

Both single channel and multi-channel systems were trained with the SI-SNR loss function [6] and the Adam optimizer. While the DCCRN model itself has been kept equal to the original, we made changes to the data synthesis process, updated to the newer 2021 dataset for training, and changed the learning rate; all for improved performance. We used a learning rate of .002, and .0005 during single channel pretraining and multi-channel fine-tuning, respectively.

3.2. Training Data

3.2.1. Single channel pretraining dataset

The DNS Challenge 2021 speech and noise data was used during the pretraining stage, but we relied on the ISM-dir dataset described in [17] for the RIRs. RIRs in this set are simulated with the image source method (ISM) where speaker sources are modelled as directive sources with an average speaker pattern directivity.

For 80% of the time, reverberant speech was obtained from combining clean speech with a random single-channel RIR. For the remaining 20%, speech was left non-reverberant. Noise (always non-reverberant) was then added to obtain the noisy input of SNR within the -5 to 20 dB range. We trained the single channel model using reverberant speech as the target, as training to a clean reference did not improve performance.

3.2.2. Multi-channel fine-tuning dataset

For the multi-channel system, also the noise was made multi-channel and reverberant using synthetic RIRs, but here sources were modelled as omnidirectional during simulation. Speech and noise sources were simulated as if from the same room, but at different random locations. The multi-channel system was trained to a clean (non-reverberant) target, by combining above RIRs with the DNS Challenge 2021 speech and noise.

4. EVALUATION

4.1. Testing setup

We test the performance of our system with PESQ, an objective measure of speech quality, and STOI, an objective measure of speech intelligibility. When calulating these objective measures, it is important to compare to the right reference signal. A dereverberating system will *appear* to have worse performance when a reverberant reference is used, as it is 'punished' for dereverberating the input, bringing the enhanced output away from the reference it is tested against. However, the single channel systems from the literature were tested against the reverberant speech signal. Therefore we switch from using a reverberant reference signal for the single-channel system (allowing for fair comparison), to the clean non-reverberant target for the multi-channel system (to take the dereverberation into account).

4.2. Testing Data

4.2.1. Single channel test set

To anchor the performance of the single channel SE system to a known test set, we test it with the DNS Challenge 2020 test set.

4.2.2. Multi-channel test sets

To create realistic multi-channel test data, RIRs were measured manually with a 9-channel circular array (planar) with 4 cm radius, positioned on a table approximately in the middle of a typical meeting room. See [17] for further details. Two types of RIRs were measured: i) speaker facing towards the array (the 'Easy' set), and ii) the speaker rotated at 90° away from the microphone (the 'Challenging' set).

Obtained RIRs were combined with random speech samples from 'NB Tale', an open Norwegian speech database. None of the training sets contained Norwegian speech.

Additionally, we recorded typical meeting room like noises (see Figure 2) in the same room, using the same array at the same location, as where the RIRs were measured. This means that all recordings also contained more general background noise, like the room's ventilation system.

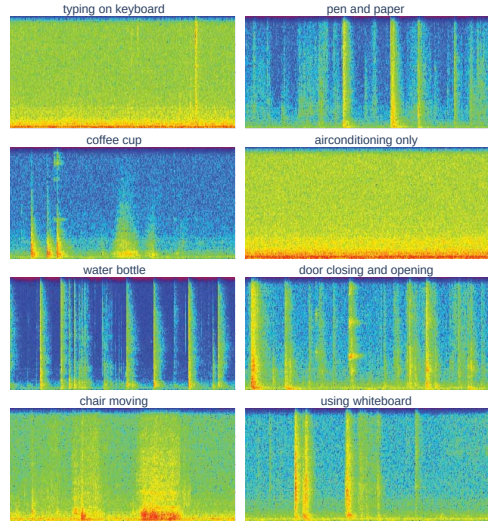


Fig. 2: Sample spectograms of recorded test noises (only the first channel is shown)

The true DOAs were measured with an uncertainty of $\pm 1°$ at random angles uniformly distributed around the array. As such, it was also possible to test using the oracle steering direction for the beamformer, which normally isn't available during inference.

4.3. Reference systems

We compare results to the performance of three reference systems from the literature, and an alternative to our proposed system:

1. **ConferencingSpeech 2021 baseline**: The multi-channel SE system described in [4], trained with our own multi-channel training set.
2. **Single channel DCCRN**: The pretrained single channel DCCRN model, where we ignore all but the first channel of our test data.
3. **GEV (oracle IBM mask) with BAN**: Mask based GEV beamformer, where the IBM mask is not estimated by a DNN, but obtained directly from the known target/noise signals.

4. **MPDR (oracle TDOAs) + Single channel DCCRN**: Here the MPDR beamformer (suplied with oracle TDOAs) is added as a standalone preprocessing step for the single channel DCCRN.

All of these systems are applied to the noisy signal directly, or to a signal that has first been preprocessed by a standalone WPE block.

5. RESULTS AND DISCUSSION

Table 1 shows the PESQ results for the pretrained DCCRN system. Our single channel system performs on par with the two winning systems, when looking at PESQ scores for the non-reverberant test set. Furthermore, the changes to the training setup give it superior performance on the reverb set, when compared to the original DCCRN-E, and also possibly when compared to PoCoNet, depending on the standard deviation of their test scores (not published). From these results we are confident that our DCCRN acts as a competitive baseline system for our multi-channel results.

Table 1: Narrowband and wideband PESQ results for the DNS Challenge 2020 channel dataset. Reverberant signal used as reference.

	PESQ nb		PESQ wb	
	No reverb	**Reverb**	**No reverb**	**Reverb**
Noisy	2.16	2.52	1.58	1.82
PoCoNet [5]	-	-	2.75	2.83[a]
DCCRN-E [6]	3.27	3.08	-	-
Our DCCRN	3.28	3.44	2.76	2.94

[a]Result without partial dereverberation, for unbiased comparison

Table 2 shows the PESQ and STOI results for the multi-channel testsets. Generally speaking, we obtain much lower PESQ scores than those observed in Table 1, despite similar SNRs in both test sets. This is because we are now calculating PESQ with respect to the clean (instead of the reverberant) speech signal.

Independent of the test set used, we see that all enhancement systems benefit from the WPE preprocessing step, even if for STOI scores the difference isn't always significant. This shows that although all systems are trained with reverberant data, they do not learn to deal with it as effectively as WPE.

The independent two-sample t-test shows that all our three systems have statistically significant higher performance than the three reference systems ($p \ll 0.05$). This is true, both when measuring performance in PESQ, or in STOI.

Table 2 furthermore shows that when the speaker is looking at the array ('Easy' set), there is no statistically significant difference in performance, between integrating the MPDR in the training loop, or simply adding it as a preprocessing step to the single channel DCCRN. The same comparison does however find a significant performance difference for the challenging dataset for the SNRs of 5 and 10 dB. Here the alternative to the proposed system (where the MPDR is added as a standalone preprocessing step before the pretrained DCCRN) performs statistically significant worse ($p < 0.05$). This suggest that integrating the MPDR into the training loop, actually allows the enhancement system to learn information that makes it better equipped to deal with a speaker looking in the wrong direction, than the MPDR is capable of on its own, unless there is too much noise.

The performance decrease from moving from *oracle* TDOAs to *estimated* TDOAs is statistically significant for lower SNRs, as expected. At low SNRs, the estimated

Table 2: Wideband PESQ and STOI results for the different multi-channel datasets. Clean signal used as reference. Best scores per SNR are shown in bold, where multiple highlighted values indicate that the difference was not statistically significant.

		Easy (looking towards array)						Challenging (looking away at a 90° angle)					
	WPE	**PESQ wb**			**STOI**			**PESQ wb**			**STOI**		
SNR [dB]		**0**	**5**	**10**	**0**	**5**	**10**	**0**	**5**	**10**	**0**	**5**	**10**
No enhancement	No	1.25	1.33	1.39	0.69	0.72	0.74	1.22	1.29	1.35	0.60	0.62	0.63
	Yes	1.33	1.44	1.56	0.72	0.76	0.78	1.27	1.36	1.46	0.18	0.66	0.68
ConferencingSpeech	No	1.33	1.36	1.48	0.68	0.72	0.73	1.27	1.31	1.41	0.59	0.61	0.62
2021 baseline [4]	Yes	1.40	1.46	1.63	0.71	0.75	0.77	1.33	1.39	1.52	0.63	0.66	0.67
Single channel DCCRN,	No	1.46	1.49	1.51	0.73	0.75	0.75	1.41	1.44	1.46	0.64	0.64	0.65
by Hu *et al.* [6]	Yes	1.64	1.71	1.76	0.77	0.78	0.79	1.55	1.61	1.66	0.68	0.69	0.70
GEV (oracle IBM mask) with	No	1.48	1.59	1.60	0.77	0.78	0.79	1.41	1.46	1.52	0.61	0.66	0.67
BAN, by Heymann *et al.* [10]	Yes	1.58	1.75	1.80	0.78	0.80	0.81	1.49	1.58	1.67	0.68	0.69	0.71
MPDR (oracle TDOAs)	No	1.68	1.73	1.76	**0.80**	**0.81**	**0.81**	1.54	1.59	1.62	0.71	0.72	0.73
+ Single channel DCCRN	Yes	**1.89**	**1.98**	**2.04**	**0.81**	**0.82**	**0.83**	**1.71**	1.79	1.85	**0.74**	0.74	0.75
Proposed system	No	1.68	1.86	1.88	**0.80**	**0.82**	**0.83**	1.61	1.73	1.78	**0.75**	**0.76**	**0.77**
(oracle TDOA)	Yes	**1.80**	**2.02**	**2.06**	**0.80**	**0.83**	**0.83**	**1.74**	**1.89**	**1.94**	**0.76**	**0.77**	**0.78**
Proposed system	No	1.60	1.80	1.85	0.78	**0.81**	**0.82**	1.50	1.62	1.69	0.72	0.73	0.75
(estimated TDOAs)	Yes	1.74	**1.95**	**2.04**	**0.79**	**0.82**	**0.83**	1.63	1.79	**1.88**	0.73	**0.75**	**0.77**

TDOAs are more likely to cause the MPDR to point towards the noise, and this even more likely to happen when the speaker is not looking at the array, weakening the direct signal. However, the TDOA estimation method used, leaves a lot of room for improvement to bring the performance closer. As such, it is very promising that the effect size of the performance degradation is this limited. It establishes the MPDR beamformer as a valid candidate for speech enhancement, especially for challenging noise types where the noise statistics are difficult to estimate.

6. CONCLUSION

We proposed a neural network-based system for multi-channel speech enhancement and dereverberation, based on WPE dereverberation, the MPDR beamformer and the DCCRN denoiser. The proposed model outperforms state of the art reference systems with respect to speech quality as measured with PESQ and speech intelligibility measured with STOI.

Future work will include improving the estimation of TDOAs by investigating other methods, and exploring opportunities within the system. Furthermore, we plan to evaluate the systems subjectively.

7. REFERENCES

[1] Chandan K A Reddy, Ebrahim Beyrami, Harishchandra Dubey, Vishak Gopal, Roger Cheng, Ross Cutler, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Sriram Srinivasan, and Johannes Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework," in *INTERSPEECH*, Shanghai, China, 2020, p. 5.

[2] Chandan K. A. Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Interspeech 2021 Deep Noise Suppression Challenge," in *INTERSPEECH*, Brno, Czechia, 2021.

[3] Chandan K. A. Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "ICASSP 2021 Deep Noise Suppression Challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, June 2021, pp. 6623–6627.

[4] Wei Rao, Yihui Fu, Yanxin Hu, Xin Xu, Yvkai Jv, Jiangyu Han, Zhongjie Jiang, Lei Xie, Yannan Wang, Shinji Watanabe, Zheng-Hua Tan, Hui Bu, Tao Yu, and Shidong Shang, "INTERSPEECH 2021 ConferencingSpeech Challenge: Towards Far-field Multi-Channel Speech Enhancement for Video Conferencing," in *INTERSPEECH*, Brno, Czechia, Apr. 2021.

[5] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, "PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss," in *Interspeech 2020*, Shanghai, China, Oct. 2020, pp. 2487–2491.

[6] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *INTERSPEECH*, Shanghai, China, 2020, pp. 2472–2476.

[7] Amaro A. de Lima, Sergio L. Netto, Luiz W. P. Biscainho, Fabio P. Freeland, Bruno C. Bispo, Rafael A. de Jesus, Ronald Schafer, Amir Said, Bowon Lee, and Ton Kalker, "Quality Evaluation of Reverberation in Audioband Speech Signals," in *E-Business and Telecommunications*, Joaquim Filipe and Mohammad S. Obaidat, Eds., vol. 48, pp. 384–396. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[8] Karen S. Helfer and Laura A. Wilber, "Hearing Loss, Aging, and Speech Perception in Reverberation and Noise," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 1, pp. 149–155, Mar. 1990.

[9] DeLiang Wang and Jitong Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[10] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.

[11] Hakan Erdogan, John R. Hershey, Shinji Watanabe, Michael I. Mandel, and Jonathan Le Roux, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *INTERSPEECH*. 2016, pp. 1981–1985, ISCA.

[12] Harry L Van Trees, *Optimim Array Processing*, Wiley, 2002.

[13] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sept. 2010.

[14] Lukas Drude, Jahn Heymann, Christoph Boeddeker, and Reinhold Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[15] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, vol. 1, pp. 375–378.

[16] Joseph Hector DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown University, Providence, Rhode Island, USA, 2000.

[17] Femke B. Gelderblom, Yi Liu, Johannes Kvam, and Tor Andre Myrvoll, "Synthetic Data For Dnn-Based Doa Estimation of Indoor Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, June 2021, pp. 4390–4394.

Paper V

F. B. Gelderblom, T. V. Tronstad, T. Svendsen and T. A. Myrvoll, 'On the Predictive Power of Objective Intelligibility Metrics for the Subjective Performance of Deep Complex Convolutional Recurrent Speech Enhancement Networks,' [Submitted]

This paper is submitted for publication and is therefore not included.

ISBN 978-82-326-5791-9 (printed ver.)
ISBN 978-82-326-6863-2 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology