Artificial Bandwidth Extension of Bone Conducted Speech using Deep Neural Networks

Master Thesis

Mikkel Heber Hahn Petersen - s123028

Tobias Toft Christensen - s123023

***Supervisor**: Tobias May*

Date: June 22, 2018

Hearing Systems

Department of Electrical Engineering

Abstract

Bone conduction microphones (BCMs) are being used in communication systems due to the low interference from outside noise sources. However BCMs imposes several challenges such as degraded and band-limited speech which reduces the perceived speech quality and intelligibility.

This Master thesis investigated the possibilities and limitations of using a deep neural network for bandwidth extension of bone conducted speech. The INVISIO X5 headset was used as a physical framework and the speech signal recorded from it's BCM was the degraded speech in focus.

The proposed solution utilizes a deep neural network to establish a mapping function from narrow band speech to wide band speech. The proposed system consists of three central building blocks; (1) feature extraction, (2) a deep neural network (DNN) mapping function, and (3) a posterior feature reconstruction.

In order to generalize better to unseen acoustic environments a multi-stage training process was utilized. The first training stage used synthetic generated data as training data. The synthetic data was constructed to reflect real world data. In the second stage transfer learning was applied to use knowledge gained from the first training stage to train and fine-tune the model using a smaller real world data set.

The network performance was evaluated by a series of experiments. The experiments were divided into measuring quality preference and intelligibility. The speech quality results showed an increase of 53.1% in preference for the processed speech, while no significant decrease in speech intelligibility. The speech intelligibility was measured by consonant recognition and confusions in normal-hearing listeners using consonant-vowel nonsense syllables presented in background noise.

Preface

This master thesis is produced and written togehter with the Hearing Systems Group under the department of Electrical Engineering at the Technical University of Denmark. It is written by Tobias Toft Christensen and Mikkel Heber Hahn Petersen in the period from January 2018 to June 2018.

This work was supervised by Tobias May from the Hearing Systems Group.

Acknowledgments

We would like to thank the people from INVISIO Communication for providing equipment, and measuring facilities for this project. Furthermore we would like to thank Johannes Zaar for providing the MATLAB GUI which made the subjective speech experiments possible.
Finally will we like to acknowledge Tobias May, Jonas Dahl, and Christoffer Ceutz for valuable feedback during the project.

Contents

Acronyms

ABE	artificial bandwidth extension
ACM	air conduction microphone
AVIL	audio visual immersion lab
BCM	bone conduction microphone
BP	band-pass
CNN	convolutional neural network
CV	consonant-vowel
DANOK	DAnsk Nonsens Ords Korpus
DNN	deep neural network
FFN	feed-forward network
GMM	Gaussian mixture model
HMM	Hidden Markov model
MCC	magnitude correlation coefficient
MOS	mean opinion score
MSE	mean square error
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
NB	narrow band
PESQ	perceptual evaluation of speech quality
ReLU	rectified linear unit
SNR	signal-to-noise ratio
SPL	sound pressure level
STOI	short-term objective intelligibility
T-F	time-frequency
WB	wide band

Chapter 1

Introduction

A robust communication system is crucial when working in stressful environments with critical noise levels. INVISIO Communication provide advanced communication and hearing protection systems, for professionals in noisy and mission critical environments. The INVISIO X5 headset uses a BCM which detects speech in the form of vibrations transmitted through the jaw bone. This improves the users communication abilities in noisy environments. The main advantage of the BCM over the traditional air conduction microphone (ACM) is the minimal signal interference caused by environmental noise sources. A major disadvantage of the BCM is it's limited bandwidth of the speech signal captured by the microphone. This leads to distorted speech quality, loss of speaker characteristics and in worst case lower speech intelligibility [Shin et al., 2012].

Artificial bandwidth extension (ABE) has been studied in the audio processing community mainly as a solution to the limited bandwidth accompanied by the telephone transmission channel [Erik Larsen, 2005]. The majority of the ABE approaches make use of the source-filter model for speech production [Rabiner and Schafer, 1978; Flanagan, 1978]. These approaches often consists of a two step procedure. The first step involves estimating the excitation signal of the source signal by findings the properties of the vocal tract. The second step involves modeling the spectral envelope for the filter. The central task for extending the bandwidth towards the high frequencies is to estimate the wide band spectral envelope. For the low frequency case the primary task is to reconstruct the missing pitch information.

Simple codebook methods [Carl and Heute, 1994; Sadasivan et al., 2016] and more sophisticated statistical models such as Gaussian mixture models (GMMs) [Nour-Eldin and Kabal, 2011] and Hidden Markov models (HMMs) [I. Katsir and Cohen, 2012; C. Yagli and Erzin, 2013; Jax and Vary, 2000] have also been thoroughly investigated. With these statistical models bandwidth extension algorithms have reached a stable baseline quality where the processed speech signals are preferred over non-processed speech signals [Erik Larsen, 2005]. Motivated by the promising performance in other signal processing areas, deep neural networks (DNNs) have been investigated as an addition to or a substitution for the statistical models like HMM and GMM in ABE applications.

DNNs have in the study area of ABE been used to approximate a non-linear regression function between narrow band (NB) speech and the missing upper band (UB) speech which has shown promising results when compared to the classification based GMM/HMM baseline. The results are among other things based on the subjective mean opinion score (MOS) test and the objective wide band (WB) perceptual evaluation of speech quality (PESQ) [Abel and Fingscheidt, 2017]. The estimation of the UB log-power spectrum using ABE algorithms have also been investigated [Li and Lee, 2015; Bukhari, 2015] together with the use of the recurrent neural network structures [Y. Gu and Dai, 2016].

The vast majority of studies applied ABE on clean speech limited to a bandwidth of 0.3-3.5 kHz, and does not consider noisy conditions such as additive background noise. Little focus have been on improving the speech quality of bone conducted speech signals with ABE methods.

Scope of Project

Based on the challenges related to the BCM the main goal is to extent the bandwidth of the band-limited speech signal in order to increase the perceived speech quality, while maintaining the speech intelligibility. This project proposes two neural network frameworks that aims to solve the problems of band-limited speech by modeling a non-linear mapping function from NB speech to WB speech. The project further investigates the benefits of ABE by the use of different objective and subjective measures. As a conse-

quence of how neural networks learn an indirect noise reduction will also take place.
Enhancing speech signals from BCMs have had a growing interest [Fingscheidi, 2012], but to our knowledge there has been very little focus on the use of DNNs. The two DNN structures which this project aims to implement are the feed-forward network (FFN) and the convolutional neural network (CNN). The FFN is the typical neural network consisting of a multi-layer perceptron like structure and the CNN is a variation of the multi-layer perceptron structure build on a shared-weights architecture.
It has been shown that CNNs obtain a high phoneme recognition performance [G. E. Hinton, 2012] and that particular CNNs have shown high accuracy in image classification[Nielsen, 2017]. Since a spectrogram can be viewed as an image, there is a high probability that especially this technique can successfully be transferred to the task of ABE, which motives the use of DNNs in the field of ABEs.

Training neural networks for regression using supervised learning techniques, requires large amount of labeled training data. To obtain a large data set this project aims to generate a synthetic speech corpus that approximates the characteristics of actual bone conducted speech. The synthetic speech corpus will be based on the TIMIT speech data set filtered with a bone conduction transfer function estimated from a corpus of real BCM speech recordings. The different DNN configurations will be trained on the synthetic data in order to capture the dynamics of a broad range of speakers. Afterwards the model will be fine tuned with the speech corpus constructed from the real recordings, by the use of transfer learning [Torrey and Shavlik, 2009]. Transfer learning is a method where knowledge gained from solving one problem can be utilized and applied to a related problem. In other words transfer learning exploits what has been learned in one setting to improve generalization in another setting.

To evaluate the ABE methods both objective and subjective metrics will be used. Speech quality has in previous studies been evaluated by the WB-PESQ measure [Fingscheidt, 2018]. It's an algorithm that analyses speech sample-by-sample of corresponding degraded and reference signals. The WB-PESQ score models the MOS which is a subjective preference scale used to assess the quality of a stimuli or system. MOS is rated on a scale from 1 (bad) to 5 (excellent).

It has been shown that a direct comparison of different ABE approaches using objective metrics is not enough, when compared to subjective listening tests [Fingscheidt, 2018; Bauer et al., 2014]. However objective measures have the advantage of saving time, and are practical when trying to get a quick performance estimate in an iterative process. The objective measure's inadequate ability to show the full picture, is the motivation to use subjective listening experiments to make a more in depth analysis of the performance of the ABE methods.
The subjective metrics will consist of speech quality and intelligibility experiments. It is important to emphasize that the need of subjective evaluation is the essential as an final evaluation.

Study show that speech intelligibility for clean narrow band speech is high. For a telephone speech signal the speech intelligibility is around 99% [Erik Larsen, 2005], indicating that the benefit from ABE is small. The speech intelligibility is on the other hand significantly reduced under noisy conditions, which makes ABE meaningful. Particularly consonants which contains energy at high frequencies suffer in the narrow band speech. This calls for a evaluation of speech intelligibility for the artificial bandwidth extended speech, in noisy conditions. This project aims to simulate real life noise environments and evaluate the speech intelligibility for NB speech and compare it to artificially bandwidth extended speech.

Structure of Report

The report is structured as follows. Chapter two provides a short review of the theoretical concepts needed to understand the considerations and reflections made in connection with the proposed method. Chapter three contains a comprehensive review and analysis of the specific steps made to implement the ABE solution. The review describes parameters such as the use of databases, synthetic data, and other resources. Chapter four provides an extensive evaluation of the performance ABE algorithm, using both objective and subjective evaluation metrics. Chapter five presents the results obtained from the evaluation metrics. The objective and the subjective evaluation results are then discussed in chapter six. Chapter seven presents the conclusion and a review of possible future directions of research.

Chapter 2

Theory

In this chapter the theoretical concepts needed to understand the principles behind the proposed algorithms will be reviewed. First the model of communication along with the different parts including speech production and transmission will be explained. Secondly different general machine learning approaches to solve the band limitation problems will be reviewed along with the essential concepts behind neural networks. Finally the different objective and subjective measures will be examined and their strengths and weaknesses analyzed.

2.1 Communication Model

Shannon's model of communication is a basic model to describe and explain speech communication [Shannon, 1948]. The model was invented to help improve telephonic communication. The model is now widely used and is considered the foundation for describing speech communication systems.

The model is reasonably simple and consist of five blocks *(fig. 2.1)*. First block contains the sender or information source. The sender is the object which chooses the transmission channel and constructs/sends the message. Second block contains the transmitter which encodes the message, and prepare the message for the transmission channel. The third block is the transmission channel which deliver the message to the receiver. The receiver decodes the message and prepare the information for the destination. The model is shown in figure 2.1.

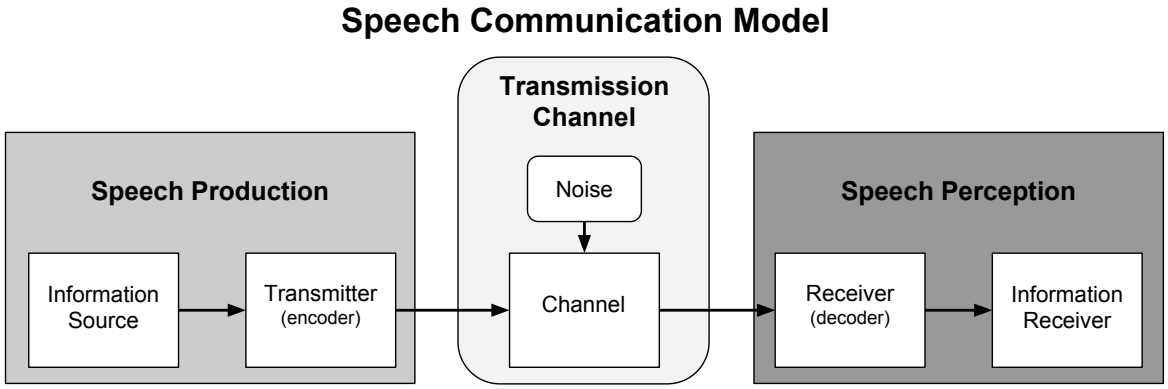


Figure 2.1: Shannon's model of speech communication [Shannon, 1948]. The information source produces a message. The transmitter prepares signal on the basis of the message which is passed through transmission channel. Transmission channel is the medium over which the signal is sent. The receiver decodes the signal back into a message. The information receiver is the destination of the message.

The noise term is included in the transmission channel block and characterize the factor which counteracts and interfere with the predictability of the communication outcome. Noise can also be added in form of articulation difficulties of the speaker. The model is important to keep in mind when considering the principles of transmission and perception of human communication.

The speech production block and the transmission channel of the communication model are the foundation of a wide range of bandwidth extension systems and are crucial to understand as they are often what is being approximated in statistical ABE methods.

2.1.1 Speech Production

Human speech is highly non-stationary and is extremely hard to model due to the constant fluctuation across time. When studying a speech signal in a shorter time frame it can on the other hand be considered partly stationary. But only when analyzing it in a 10-30 ms time frame [Loizou, 2013].

The speech signal is produced when air is pressed from the lungs through the vocal tract and on through the mouth and nose cavities [Poulsen, 2008]. The intensity of the sound is defined by the intensity of the air flow from the lungs. The tension and length of the vocal folds determines the pitch of the signal.

The timbre of an individual is unique and is determined from a wide range of different

parameters. The shape and size of the vocal tract has a large influence, but also the shape of the chest, neck and the position of the tongue can be altered to affect the timbre of an individual's speech.

The knowledge of speech production and the communication model is important if one wants to understand the properties and characteristics of a speech signal, and the elements which speech is composed of. This knowledge and understanding is favorable when investigating and evaluating speech enhancement algorithms. It can help to understand which speech elements are important if one wants to improve speech quality and/or intelligibility.

Technical Model of Speech Production

The source-filter model is a very recognized method to model speech production *(fig. 2.2)* and is motivated by the actual human speech production system. The model is, as the name suggests, a decomposition of the speech signal into two parts. The source (excitation) part and the filter part.
The source part of the model can be related to the human vocal folds from where an excitation signal is produced. The excitation signal can be periodic or non-periodic. The filter part of the model can be viewed as the vocal tract and cavities in the mouth and the nose. The vocal tract acts as a filter which shapes the spectra of the excitation signal coming from the source (vocal folds).

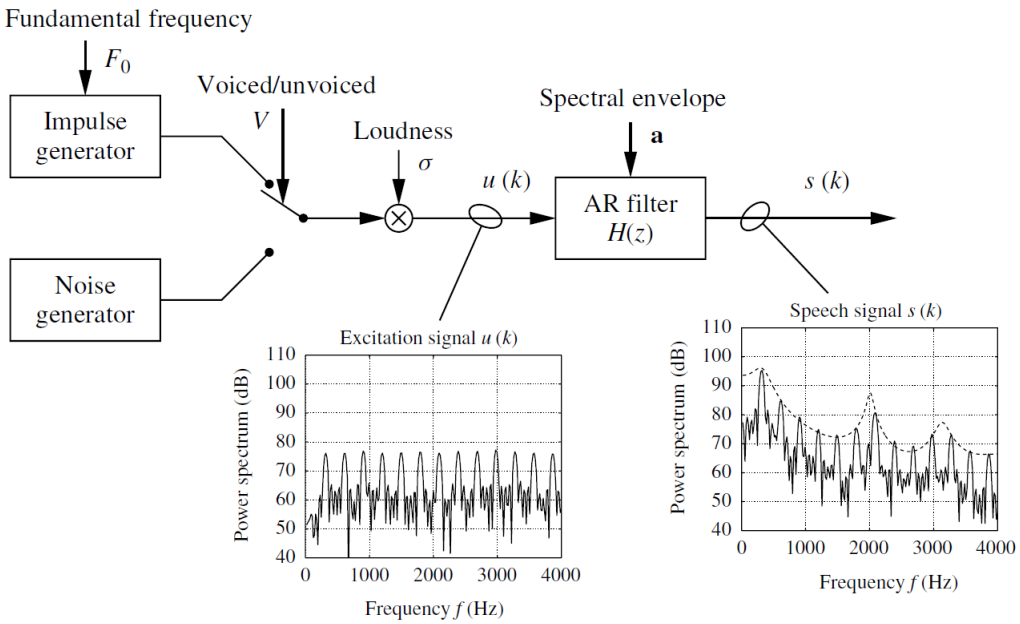


Figure 2.2: The Source Filter model of speech production. The power spectra of the excitation and output signal are shown for an idealized voiced speech sound. The speech signal is shaped by the auto-regressive (AR) vocal tract filter H(z). *(Figure from [Erik Larsen, 2005])*

The technical model have achieved a wide acceptance in application such as speech codex, and vocoders [Rabiner and Schafer, 1978; Flanagan, 1978]. This is properly due to it's simplicity. It is a simplistic linear model of speech production in which the source (vocal folds) is uncorrelated with the filter (vocal tract). Further more is it assumed that the relationship between the pressure and the velocity is linear. None of these factors is in reality so simple, and the complete model is much more complex and non-linear.

2.1.2 Speech Material

The spoken language is constructed from a relatively low number of building blocks. These building blocks are known as phonemes. A phoneme is defined as the smallest meaning-distinctive part of the speech signal. Phonemes can be decomposed into different categories. The categories are illustrated in figure 2.3.

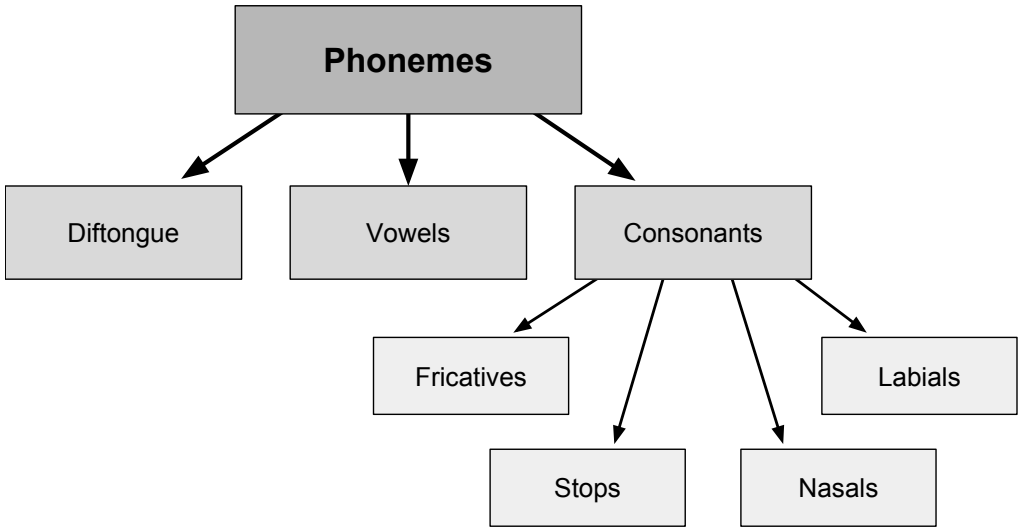
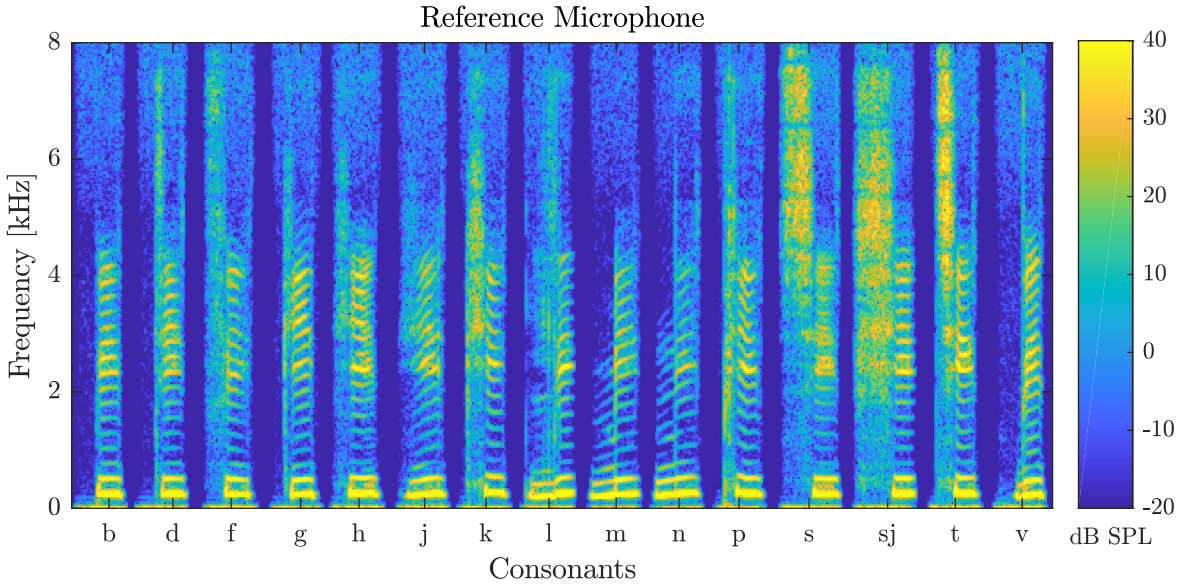


Figure 2.3: Overview of the different phoneme categories. Phonemes can be split into three main categories. The consonant category can further be divided into four subcategories

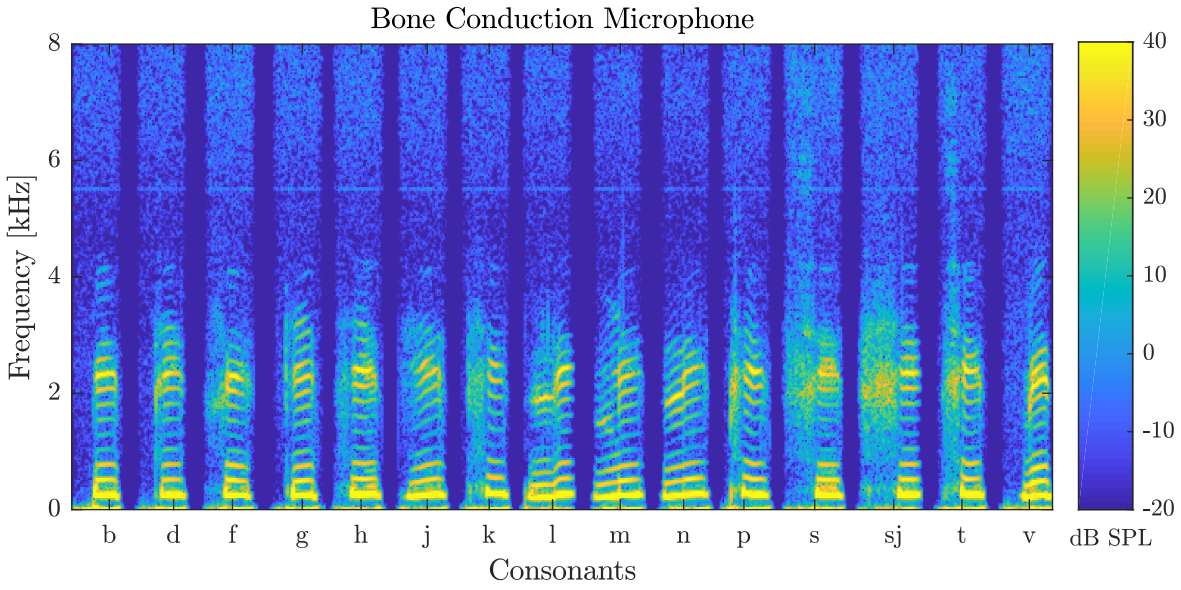
The fricatives and the stop can further be divided into voiced and unvoiced speech sounds.

Voiced and unvoiced speech are standardized way to classify speech sounds. It is a way of grouping and labeling speech sounds in phonetics and phonology. Voiced sounds are produced by vibration of the vocal tract, and includes consonants such as /b/ and /g/. The spectrum for voiced sounds follows the formant structure [Poulsen, 2008]. The formant peaks varies according to the specific sound.

Unvoiced sounds are on the other hand produced without any vibration of the vocal tract, and includes consonants like /k/, /p/, and /f/. Unvoiced sounds are instead produced by different variations of air flow modulation. For example /f/ is produced by pressing air out between teeth and the lips. /p/ is produced from a sudden opening of the air flow. The spectrum for unvoiced sounds does not follow the formant structure. Figure 2.4a shows the spectrogram of 15 consonant-vowel (CV) combinations. All consonants are followed by the vowel *i*.



(a) Spectrogram of the 15 CV combinations used in [Dau, 2017]. Every consonant is followed by the vowel *i*. The CV tokens are recorded with the reference microphone.



(b) Same CV combinations as above, but recorded with the BCM. The BCM recorded CV combinations all show large attenuation of the frequencies above 4 kHz. The frequency range below 4 kHz show signs of degradation.

Figure 2.4: Comparison between the spectrogram representations of 15 CV combinations recorded with an ACM and a BCM.

From figure (2.4a) a connection between the categorization of the phonemes and the spectrogram representations can be observed. For the consonants /s/ and /sj/ a large

amount of broadband energy is present before the vowel. These are both categorized as fricatives since the sound comes from turbulent noise produced with the lips or teeth. Turbulent noise is often broadband.

Bone Conduction

ACMs are the most commonly used microphones to capture speech signals. From a practical point of view ACMs have to be mounted close to the mouth which makes them unpractical in a dynamic environment. The so-called boom microphone which is mounted on the head is one solution to this problem. One essential problem still remains, a ACM is not very effective in noisy environments. The ACM is not shielded from unwanted sounds from the surroundings, resulting in a significantly loss of speech intelligibility at low signal-to-noise ratios (SNRs).
A BCM is a potential replacement. The BCM is attached to the users head and captures the vibrations of the skull which occurs when a speech signal is being produced. Besides the attachment which makes the BCM more optimal in dynamic environments, the most important advantage of the BCM is the minimal signal interference of unwanted noise from the surroundings.
An issue accompanied with the BCM is it's limited bandwidth of the speech signals captured by the microphone. The limited bandwidth of the speech signals often leads to a reduced speech quality, loss of speaker characteristics and in worst case a reduced speech intelligibility [Shin et al., 2012].
Compared to natural speech, band-limited speech signals have significantly degraded speech quality. Eliminating the high-frequency components above 3-4 kHz leads to a reduction of phoneme recognition. The band-limited speech is also described as 'muffled', indicating the lack of high frequency speech components. The absence of these can result in a reduced ability for the listener to identify the speaker.

2.1.3 Speech Perception

Depending on the quality of the transmission channel the desired signal can be degraded on a more or less severe grade. The degradation of the signal can impact the

intelligibility of the speech material in a way that the word are unrecognizable and the meaning and context of the speech is lost. Another aspect is the amount of cognitive effort the listener has to put into the task of understanding the speech. When noise is added to the speech signal the cognitive effort is increased and the amount of brain power that can be utilized on other tasks is decreased.

2.2 Artificial Bandwidth Extension

Speech enhancement aims at improving the intelligibility and perceptual quality of speech signals, by the use of signal processing algorithms. Artificial bandwidth extension (ABE) is a major field under speech enhancement and is largely motivated by the tele-community. A vast variety of ABE methods for telephone speech has through out the years been proposed. Methods ranging from simple signal processing techniques to more complex methods which takes the source-filter model of speech production into consideration. More modern non-linear and statistically signal processing techniques has also been widely investigated and has shown promising results [Fingscheidt, 2018; Loizou, 2013].

2.2.1 Statistical Artificial Bandwidth Extension Methods

Several statistical learning approaches have been thoroughly investigated through out the last decays. Statistical learning approaches such as Gaussian mixture model and Hidden Markov model.

Gaussian mixture models are a probabilistic model which characterizes the problem by assembling weighted multivariate Gaussian distributions and through the use of the an Expectation-Maximization iterative learning process try to classify the input data.

Speech is temporal dependent, meaning that each time frame in speech processing is dependent on the previous time frame. The Hidden Markov model exploits these temporal dependencies and incorporate these dependencies to estimate a bandwidth extension. The Hidden Markov Model is often considered as a more sophisticated method than the Gaussian mixture model, since it utilizes temporal information hidden in a speech sequence which evolves over time.

2.3 Neural Networks

Artificial neural networks have in many years been studied for the purpose of archiving human like performance in all sorts of aspects. Neural networks are as the name implies a computational model inspired by the biological neural system observed in brains. The networks are made up of simple computational elements (nodes) called neurons *(fig: 2.5)*. The neurons are connected by links with variable weights in a densely connected network structure. The computational nodes perform simple calculations like additions and multiplications and the output is passed through a non-linear term called an activation function [Lippmann, 1987]. A neural network is often used to perform a non-linear mapping between some input data and some output data. It has through studies been shown that a neural network can successfully realize and model a wide range of complex non-linear functions [Nielsen, 2017].

The neural network is presented with input data. At each neuron the the data is systematically added up and the output is calculated on the basis of a given non-linear activation function. The output from the activation function is added together with a bias, which acts as an offset. The output of a neuron is then fed to the next layer also consisting of neurons. The strength of the data fed from neuron to neuron is determent by the connection links between them, called weights. The weights and biases are changeable and are adapted during a training phase to optimize the output of the neural network.

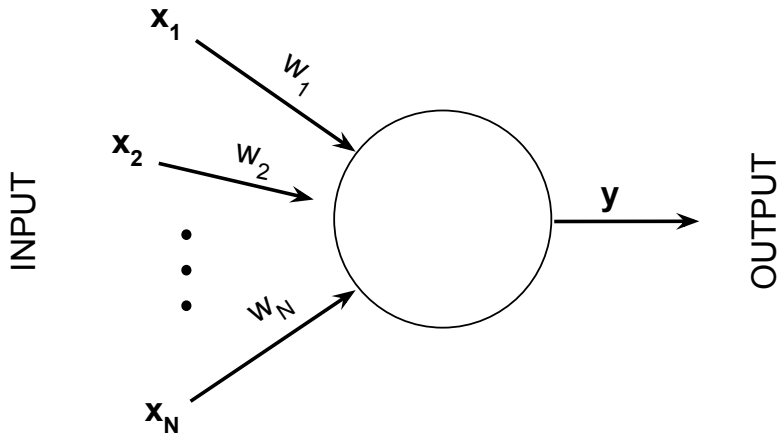


Figure 2.5: A visual illustration of the mathematical process inside a neuron. Each input (x_1, x_2, x_N) to the neuron is multiplied with a corresponding weight (w_1, w_2, w_N) and summed. A bias term (w_0) is added, before processed by an activation function **f**.

Figure 2.5 illustrates the mathematical process which occur in each neuron. The process is given by following equation:

$$y = f\left(\sum_{i=1}^{N} w_i x_i + w_0\right) \tag{2.1}$$

Where $x = [x_1, x_2, ..., x_N]$ defines the inputs and $w = [w_1, w_2, ...w_N]$ defines the weight parameters. $f(\cdot)$ is the activation function.

The activation function defines how the neuron responds to a given input. A wide range of different activation functions has been the focus of research over the years [Lippmann, 1987]. An example is the simple threshold based activation function, telling the neuron to be active if the output is greater than a threshold, and not active if the output is smaller than the threshold. This is a binary approach and is not very flexible. Another very popular activation function, especially in classification tasks, is the Sigmoid function. The Sigmoid function is a smoothed step function. Opposite to the step function it produces non-binary outputs and have a smooth gradient. The output of the Sigmoid activation function is limited in the range between zero and one which means that the outputs can't explode. The Sigmoid activation function is given

in equation 2.2

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

The most recent and very popular activation function is the rectified linear unit (ReLU) which equation is shown in 2.3. The paper [X. Glorot and Bengio, 2011] explores the use of ReLUs as an alternative to the Sigmoid activation function. They suggest that the ReLU activation function is of special interest because it is a more realistic model of a biological neuron. An important point to make is the increased sparsity in the network structure introduced by the ReLU activation function. For large network structures with a large amount of neurons, the Sigmoid function will cause nearly all neurons to activate, which means that all neuron outputs will be processed to describe the output of the network. The ReLU function outputs 0 for negative values of x, meaning that for a random weight initialization half of the neuron outputs will become zero. This reduces the computational cost of the network significantly, and allow for more sparse and efficient model.

$$f(x) = \max(0, x) \tag{2.3}$$

The ReLU activation function introduces a new issue known as "the dying ReLU problem". A ReLU neuron can "die" under the backpropagation phase *(Backprop. sec: 2.3.2)*, meaning that the weights and the bias's gets updated in such a way that the ReLU neuron ends up in a state which its unlikely to recover from. An example could be that the neuron learns a large negative bias term. The problem originates in the region for negative x values, where the activation will result in zero. This means that the weights will not get adjusted during the gradient descent phase. This problem can cause that a large amount of the neurons stops responding and that way making a considerable part of the network passive. Even though that a Sigmoid neuron can suffer from a similar problem, there will at least always be a small gradient allowing them to recover eventually.

A solution to the dying ReLU problem is to allow for a small, non-zero gradient for negative x values. Meaning that for negative values of x, the activation function will

not output zero. This solution is called a Leaky ReLU activation function. The idea is simply to allow the neuron to recover during training, by letting the gradient differ from zero. The difference between the ReLU and the Leaky ReLU activation function is illustrated in figure 2.6.

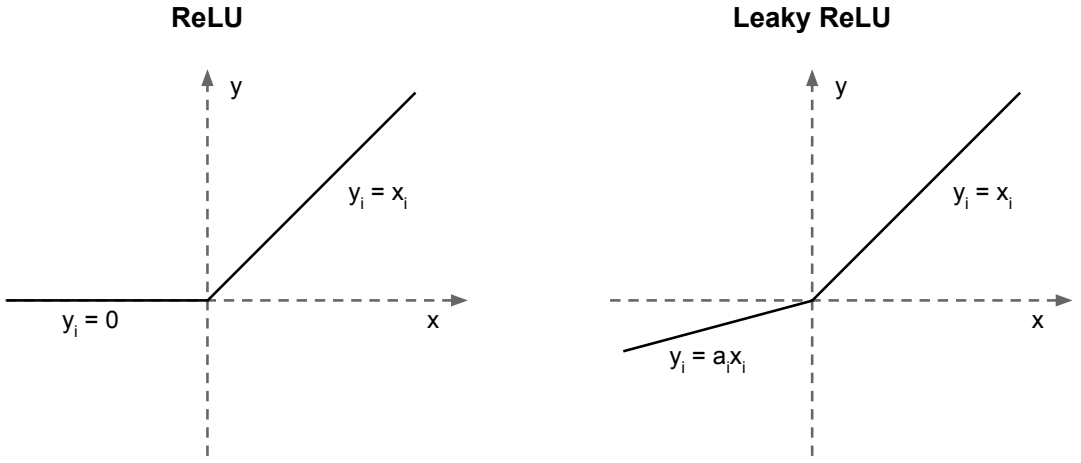


Figure 2.6: The left figure presents the ReLU function and the right figure presents the leaky ReLU function. The a_i value in the leaky ReLU function determines the size of the 'leak'.

2.3.1 Cost Function

A cost function is a function which outputs a scalar that quantifies the error of the performance of the neural network.

Choosing the right cost function is crucial to achieve the desired result of a network. The optimal cost function depends on the specific task. Choosing the right cost function affects how quickly and accurately the network learns the correct weight values. A general rule is to use a cross entropy cost function for classification and a mean squared error for regression problems. But if one has a large prior knowledge about the data, and know exactly what the neural network should achieve, a more specialized cost functions can be considered.

As this project focuses on a regression problem the mean squared error cost function is chosen:

$$C(w,b) = \frac{1}{2n} \sum_x ||y(x) - a||^2 \tag{2.4}$$

Where w is the collection of weights, b is the collection of biases, n is the total number

of inputs to the network, $y(x)$ is the target output and a is the a vector containing the predictions of the network for input vector x.

In classification tasks the link between the cost function and the performance of the network is straight forward. If the task is to classify handwritten digits the cost function usually returns the number of misclassified digits in percentage.
The task of bandwidth extension of speech is much more complex. A minimal error between the output and the target doesn't necessarily result in a perceptually better result, since low and high frequencies have a different impact on speech quality and intelligibility [Erik Larsen, 2005]. This could lead one to use a perpetually inspired function as PESQ or short-term objective intelligibility (STOI). But two assumptions needs to hold for a cost function. The first assumption that needs to hold is that the cost function should possible to write as an average over cost functions for individual training examples. The second assumption is that the cost function can be written as a function of the outputs from the neural network. Both of these assumptions holds for the mean square error (MSE) [Nielsen, 2017]. But since PESQ and STOI require time signals as inputs and the output of the network is frequency bins (see chapter 3), both assumptions would violated with those cost functions.

2.3.2 Backpropagation

The backpropagation scheme is responsible for updating the weights and biases in the neural network and that way optimizing the performance of the network. This means that the main goal of the backpropagation scheme is to update every weight and bias so that the output of the network results in the lowest possible error.
To understand the backpropagation scheme it is important to understand how changing the weights and biases in a network changes the cost function [Nielsen, 2017]. Meaning determining the partial derivative of the cost function:

$$\frac{\delta C}{\delta w_j^l} \qquad\qquad \frac{\delta C}{\delta b_j^l} \tag{2.5}$$

Equation 2.5 illustrate the partial derivatives of the cost function in relation to the weights (w) and in relation to the bias (b). l indicates the layer and j specify the

neuron.
The derivative of the cost function in a specific point gives the rate at which the function is changing with respect to the weight and biases. The backpropagation scheme analyses and exploits the rate at which the cost changes relatively to the changes on the weights and biases.
The backpropagation scheme is based on a set of fundamental equation, which together determine how the weights and biases of the network should be changed in order to minimize the performance error [Nielsen, 2017].

2.4 Neural Network Structures

2.4.1 Supervised Learning

Machine learning algorithms are often classified into two groups, supervised learning and unsupervised learning.
Supervised learning is the most common and straightforward machine learning technique, and is as the name suggest a technique where the learning process is supervised. In a supervised learning process prior knowledge regarding the ground truth is used to train the model. In a supervised regression problem the goal is to learn a function which given an input performs a mapping to a desired output.
Unsupervised learning is more towards what is known as artificial intelligence. As the name suggest unsupervised learning is a technique where the algorithm is unsupervised and learns patterns without human guidance.
Supervised learning is commonly used in regression problems. If a system is given by $f(X)$, with input X and a known output Y, supervised learning can be used to learn or approximate the system which maps the input variable to the output variable. The approximated mapping function can than after the training phase be used to predict the output for new unseen input data.

2.4.2 Feed-Forward Neural Network Structure

The Feed-Forward neural network structure is also known as a multi-layer perceptron structure [Nielsen, 2017]. Perceptrons were first described by the scientist Frank Rosen-

blatt back in 1950s. Perceptrons are strongly related to neurons which were described in the a previous section. The feed-forward structure follows a few simple characteristics. First of all the neurons are arranged in layers. The first layer, called the input layer, processes the inputs, and the last layer, called the output layer, produces the outputs. In between the input and the output layers are a series of so called hidden layers.
Second of all the information is always fed forward in the network, hence the name. Furthermore there is no connection among the neurons in the same layer. If the network is *dense* every neuron in one layer is connected to every neuron in the next layer.

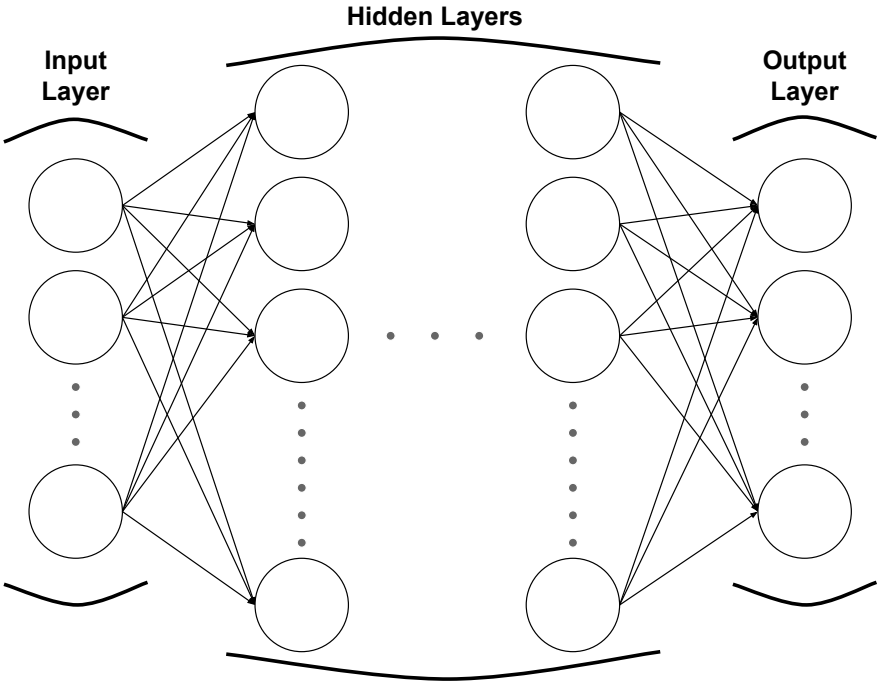


Figure 2.7: The structure of a fully connected neural network. The diagram illustrates how the inputs are connected to the neurons in the hidden layers. Two hidden layers are shown, but more could be utilized. The outputs of the last hidden layer are connected to an output layer.

The simplicity of feed-forward networks also has a down side. In machine learning a phenomenon called the *Curse of Dimensionality* is used to describe various challenges and can also be used to describe one of the limitations of the FFN structure [Bishop, 2006]. Because the first hidden layer of the FFN is fully connected to all the input features the number of parameters can explode if the dimensionality of the input features increase [Nielsen, 2017].

2.4.3 Convolutional Neural Network Structure

The convolutional neural network architecture takes advantage of structural relations in the data. This could be the spatial relationship of pixels in an images or the temporal relations of data points in a time-series. The technique have had a large impact on reigniting the interest of neural networks.
The convolution neural network structure is in many ways very similar to the feed-forward neural network structure. It is also made up of neurons arranged in layers and each neuron have both trainable weights and biases. Each neuron receives input data which has been processed by the weights. A cost function is used to calculate the final loss which is used to update the weights and biases of the network. The major difference is that the convolution network assumes that the input data contains useful structural information, which allow for certain beneficial properties when designing the network architecture.

The convolutional neural network architecture is build on three basic ideas, local receptive fields, shared weights, and pooling *fig: 2.8 [Nielsen, 2017]*.

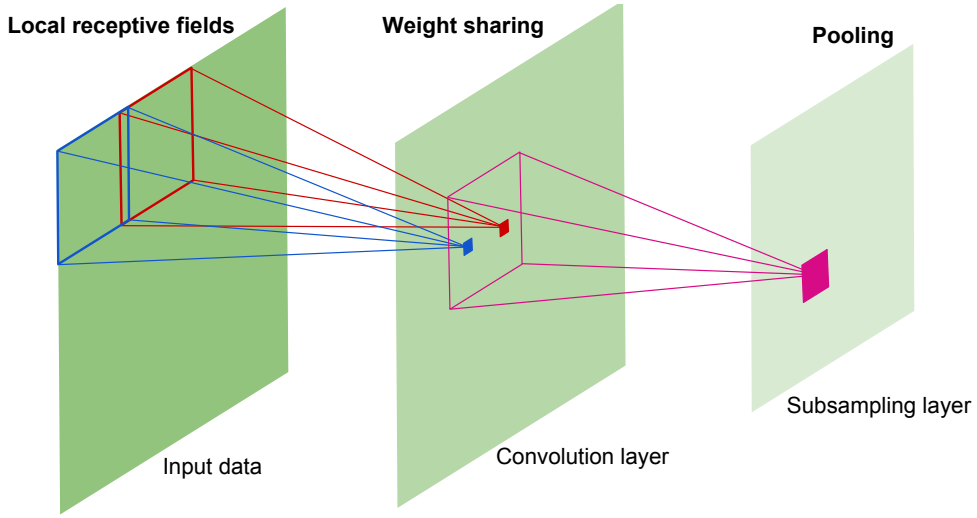


Figure 2.8: Convolution Structure. The blue and red squares visualize the local receptive fields of the convolution layer. The values of the convolutional layer are pooled together in smaller fields to condense information [Nielsen, 2017].

In the feed-forward network structure each input is connected to every neuron in the

first hidden layer, making it dense. As the dimensionality of the input features to a convolutional network often is high, it makes it impractical to use the feed-forward structure as the number of parameters will explode. In CNN every input is therefore not connected to every neuron in the hidden layer, instead the inputs are grouped together in small, localized regions which is connected then to a neuron in the hidden layer. Such a region is called a *local receptive field*. The size of the region can be changed, as well as the overlap between regions. The overlap is determined from the *stride length*. The stride length is defined as how far the local receptive field is being moved. If the overlap is large, meaning the stride length is small, a larger amount of weights has to be utilized.

Each neuron in the hidden layer is connected to it's corresponding local receptive field and in that field all weights and bias are shared. This means that even though the local receptive field changes position the weights and biases are the same. The hidden layer detect exactly the same 'feature', just in different places. The shared weights and bias are often called a kernel or a filter. By increasing the amount of filters the network can capture more features, which can be valuable when dealing with complex inputs.

The third part of the convolutional structure is the sub-sampling layer often known as a pooling layer. Pooling layers are placed after a convolutional layer and can be viewed as down-sampling layers or a way to condense the spacial information extracted from the convolutional layers. That way reducing the amount of parameters and computations in the network.

After the three described building blocks one or more fully connected layers are often used to make the network learn on a more abstract level and to integrate the global information from the convolutional layers.

2.5 Objective Metrics

To evaluate the performance of speech enhancement algorithms objective metrics are often used as a way of quantifying speech quality or speech intelligibility differences by a mathematical comparison between a reference and a degraded signal. Objective metrics, as the name implies, are impartial and are not affected by subjective opinions or interpretations.

Three different objective metrics are used in this project; Perceptual Evaluation of Speech Quality (PESQ), Magnitude Correlation Coefficient (MCC) and Short-Term Objective Intelligibility (STOI). PESQ is, as the name implies, a measure of the speech quality and has in previous studies has been used to evaluate performance of ABE algorithms. The wide band (WB-PESQ) measure was used to evaluate speech quality in [Fingscheidt, 2018]. It is an algorithm that analyses speech sample-by-sample for corresponding degraded and reference signals. The WB-PESQ framework models the mean opinion score (MOS) which is a subjective measure used to assess the quality of a stimuli or system. MOS is rated on a scale from 1 (bad) to 5 (excellent).
The MCC metric is an objective measure of the magnitude spectrum correlation between a degraded signal and a reference signal. This is an interesting metric because it allow for an correlation analysis of energy in specific frequency bins. In relation to ABE algorithms it allow us to quantify the improvement of the high frequency content of a signal.
STOI models the subjective speech intelligibility of signals and is based on the MCC method. But where MCC only calculates the correlation in TF-units, STOI is used to predict how human listeners would perceive the output of an speech enhancement algorithm.

2.5.1 Perceptual Evaluation of Speech Quality (PESQ)

The evaluation of a speech processing system is often divided into two categories. The speech quality which is categorized as the naturalness or pleasantness of speech, and the speech intelligibility which is categorized as how understandable speech is under certain conditions. PESQ is an objective perceptual method for predicting the subjective speech quality. PESQ was developed to perform automated objective assessment of speech quality through telephone transmission systems. Today it is a metric which is applied world wide in the tele-communities and in hearing research. PESQ is standardized by the Telecommunication Standardization Sector (ITU-T) as recommendation P.862 (02/01) [ITU-T, 2011a]. The mapping function (P.862.1) which maps the wide band PESQ to the MOS scale can be seen in figure 2.9 [ITU-T, 2011b].

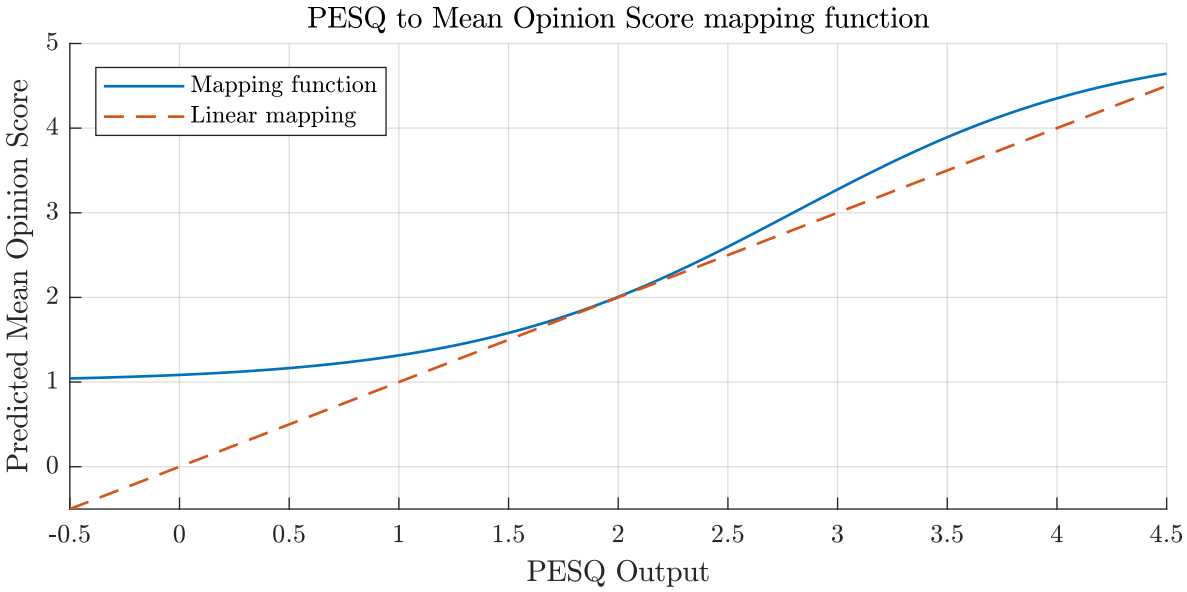


Figure 2.9: The figure shows the non-linear mapping function from the PESQ metric to the predicted mean opinion score. The PESQ metric outputs a score in the interval from -0.5 to 4.5 and the MOS metric in a range from 1 to 4.5. The correlation between PESQ and MOS was found to be ($\rho = 0.92$) [Loizou, 2013].

Even though PESQ is found to correlate very well in the case of quantifying degradation due to transmission codex, it is not found to correlate with the subjective MOS test in relation to other tasks. Tasks such as noise reduction and dereverberation. In several publications the PESQ metric is also found to be insufficient when comparing different ABE approaches to subjective listening tests [Möller et al., 2013; Pulakka et al., 2015]. [Abel et al., 2016] purpose a novel instrumental quality measure that is specifically suited to test ABE approaches. This metric is quite new and could be a important tool for evaluating not only ABE algorithms but also other speech enhancement algorithms. But at the moment it is still new, and is yet to be fully tested. We will therefore in this project not use it.

It should be stated clearly that the PESQ results should not be blindly trusted, and it is essential to compare them to subjective measurements. This makes one wonder why the PESQ metric is so frequently used by scientists to evaluating tasks where the effectiveness is questionable. The answer is often that the subjective metrics are more time consuming, and that the scientists simply does not have a better alternative.

2.5.2 Magnitude Correlation Coefficient (MCC)

Even though it appears that there is a correlation between the speech quality and the speech intelligibility, PESQ was found to correlate really poorly with speech intelligibility Loizou [2013]. This motivates the use of other objective metrics to evaluate the specific aspects of the ABE algorithm.

The Spectral magnitude correlation coefficient (MCC) is a measure based on how similar the spectrum's of the reference and the degraded signal is. The equation for the MCC is presented in equation 2.6.

$$d_{MCC}(x,y) = \frac{\sum_{j=0}^{J-1}(\Gamma_x(j) - \mu_{\Gamma_x})(\Gamma_y(j) - \mu_{\Gamma_y})}{\sqrt{\sum_{j=0}^{J-1}(\Gamma_x(j) - \mu_{\Gamma_x})^2 \sum_{j=0}^{J-1}(\Gamma_y(j) - \mu_{\Gamma_y})^2}} \tag{2.6}$$

Where $\Gamma_x(j)$ and $\Gamma_y(j)$ denotes the energy within the jth critical band of the spectrum x_m and μ_{Γ_x} and μ_{Γ_y} denote the sample mean of the clean and processed (degraded) critical band values

The MCC metirc uses a perceptually motivated frequency analysis by the means of a DFT based critical-band decomposition. A critical-band decomposition describes the decomposition into frequency bands with a bandwidth estimated from the auditory filters inspired by the human cochlea. It is implemented by applying an l_2-norm on the critical-band filtered DFT spectrum *(eq. 2.7)*.

$$\Gamma_{x_m}(j) = \sum_{k=0}^{K/2} \mid H_j(k)X_m(k) \mid^2 \tag{2.7}$$

In a paper by [C. H. Taal and Jensen, 2011] 17 different objective measures were evaluated. Out of all 17 measures the highest correlation with respect to speech intelligibility was obtained for the frame based MCC measure, despite it's simplicity. The correlation were between the objective metrics and subjective listening experiments. The subjective listening experiments consisted of word recognition rate.

The paper also conclude that the advanced objective speech quality metric PESQ has a low correlation with speech intelligibility, which emphasizes the need for multiple metrics.

The MCC metric will in this project primarily be used as a simple spectral correlation measure, which allow us to measure how much the processed speech signal correlates with the clean reference speech signal. The MCC metric is an interesting analyzing tool in relation to ABE methods, as it can analyze the correlation in specific frequency bands. This is valuable when one wants to isolate and analyze the correlation improvement specifically for the high frequency bands.

2.5.3 Short-Term Objective Intelligibility (STOI)

The STOI algorithm was developed as an improved objective metric that correlated well with the subjective intelligibility score [Cees H. Taal and Jensen, 2010]. The algorithm is a further development of the MCC metric and is aimed at evaluating speech signals enhanced by noise reductions algorithms. The method consists of calculating the TF-representation of a processed speech signal and the corresponding clean reference signal. The TF-representations are calculated from 400ms, 50% overlapping, Hanning-windowed DFT frames, which are grouped into 15 1/3-octave bands with the lowest center frequency set to 150 Hz.

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k,m)|^2}, \tag{2.8}$$

Equation 2.8 present the equation for the TF-representation of the clean reference signal $X_j(m)$, where k_1 and k_2 denote one-third octave band edges which are rounded to the nearest DFT-bin. The TF-representation of the processed speech is obtained similarly and is denoted $Y_j(m)$.

The TF-representation is used to calculate a intermediate speech intelligibility *(eq: 2.9)*, which is a linear correlation coefficient between the clean and processed TF-units.

$$d_j(m) = \frac{\sum_n \left(X_j(n) - \frac{1}{N} \sum_l X_j(l) \right) \left(Y_j'(n) - \frac{1}{N} \sum_l Y_j'(l) \right)}{\sqrt{\sum_n \left(X_j(n) - \frac{1}{N} \sum_l X_j(l) \right)^2 \sum_n \left(Y_j'(n) - \frac{1}{N} \sum_l Y_j'(l) \right)^2}} \tag{2.9}$$

where $l \in \mathcal{M}$ and $Y' = \max(\min(\alpha Y, X + 10^{-\beta/20}X), X - 10^{-\beta/20}X)$, where Y' is the normalized and clipped TF-unit, β is the lower signal-to-distortion ratio and α is a factor such that the energy equals the clean speech energy, within that TF-region.

The final objective intelligibility measure is the mean of the intermediate measure *(eq: 2.10)*.

$$d = \frac{1}{JM} \sum_{j,m} d_j(m) \tag{2.10}$$

where M is the total number of time frames and J is the number of frequency bins. STOI has been found to correlate very well with subjective measurements ($\rho = 0.95$) [Cees H. Taal and Jensen, 2010].

2.6 Subjective Metrics

Objective measures have the advantage of saving time, and is practically feasible when trying to get a quick performance estimate of speech enhancement algorithms. The objective measures are also favorable when comparing different variants of the same ABE method to each other. However it has been shown that a direct comparison of different ABE approaches using objective assessments is not optimal when compared to subjective listening tests [Fingscheidt, 2018; Bauer et al., 2014].
This makes subjective evaluation essential and necessary to obtain a complete picture of the actual performance. Subjective evaluations are normally performed by listening tests and/or conversational tests.

2.6.1 Speech Quality

To this day subjective listening tests are the most recognized way of categorize the effect of different algorithms on the sound quality [BS.1534-2, 2014]. Listening tests needs to be carefully designed to limit the effects of uncontrolled factors. The tests must be designed so that the degradation of the audio are easy to assess on a quality scale. This means that the scale of preference needs to reflect and reveal the degradation of the audio signal. There should be a audible difference between test signals. Listening tests must also be designed so that the subjects don't get exhausted by the duration of the

complete test, which could lead to noisy results.
Another crucial factor for obtaining reliable and repeatable test results is a subject screening process. A screening process can include an audiogram which determines subjects hearing ability or an expertise gauge, that is intended to measure subjects experience and the subjects ability to listen critically.

To fulfill these requirements the MUSHRA: ITU-R Recommendation BS.1534-2 has been chosen for this project. In the Recommendation, guidelines for all of these consideration has been provided. A more thoroughly review of the MUSHRA procedure is written in the evaluation section.

2.6.2 Speech Intelligibility

Speech intelligibility has been the focus of investigation for a long time and is defined as how understandable a speech signal, from a speaker, is for a listener. The speech intelligibility is dependent on a variety of parameters. Background noise, reverberation and the quality of the given transmission system all have a large impact on the speech intelligibility and the listeners ability to correctly perceive the content of the speech signal.
The influences of intonation, pauses etc. in speech also determines how a listener understands and interprets what he/she understands from a speaker.
It is important to evaluate speech intelligibility as it's an important factor in the basic model of communication (Shannon's model - 1948). It involves the receivers ability to decode the message. The ability to understand and interpret speech signals are essential to human communication.
Studies show that speech intelligibility for narrow band speech signals is high [Erik Larsen, 2005]. In clean telephone speech it is around 99%, indicating that the benefit from ABE is small. The speech intelligibility is on the other hand significantly reduced for narrow band speech under noisy condition, which makes ABE meaningful. Particularly fricatives which contains most of energy at high frequencies suffer from the band limitation. Thus, testing speech intelligibility in background noise is interesting as an extended evaluation of the ABE algorithm.

One of the most simple and recognized evaluation techniques for speech intelligibility is the percentage of correct identified words or complete sentences in the presence of interference. The listener is presented with a list of words or sentences that have a known signal to noise ratio (SNR). For each of theses SNRs, a number of correctly recognized words or sentences are counted and divided by the number of presented words or sentences to find the probability of correct answers. These values for each SNR are referred to as intelligibility scores and are plotted as a function of SNR which normally results in a S-shaped psychometric function.
Usually, it's easier and more practical to present the speech intelligibility with a single value. To do that, the concept Speech Reception Threshold (SRT) is used [Hagerman, 1982]. This threshold is defined as the SNR level for which 50% of the spoken corpus can be correctly recognized and understood by the listener.
This type of speech intelligibility test provide an informative macroscopic evaluation. However the results of such a method may be influenced by the listener's cognitive abilities, such as the ability to identify certain inaudible words in the context of a complete sentence.

A way to avoid this could be to study the speech intelligibility on a more microscopic phoneme level. More specifically a consonant recognition measurement could be very useful in evaluating and analyzing the speech intelligibility of the artificial bandwidth extended speech. The consonant perception measurement allows for a more detailed investigation and provides information on which phonemes gets confused and which phonemes that doesn't. Another important advantage of this approach is that many consonant speech cues contain most energy at high frequency. Studies show that conflicting cue regions in natural speech are often correlated with consonant confusion, which particularly arises under noisy conditions or for bandwidth-limited speech [Dau, 2017; Allen, 2012]. This is a motivating reason to use this method for ABE evaluation, where the purpose is to reconstruct the high frequency content.

Chapter 3

Artificial Bandwidth Extension

In this chapter the different building blocks of the algorithm is presented. First the complete structure of the algorithm is briefly explained followed by a detailed rundown of each building block. Second the procedure of training a complex neural network is explained along with the reasoning behind the iterative process.

The proposed artificial bandwidth extension algorithm is shown in figure 3.1 as a block diagram depicting the high level building blocks which makes up the complete system. The algorithm consists of three primary parts. (1) The feature extraction module which as the name suggest extracts the relevant information by converting the raw input speech signal to useful features. (2) The deep neural network (DNN) module which is the core of the algorithm and computes the mapping between the input features $|\mathbf{X}|$ and the output features $|\dot{\mathbf{X}}|$. (3) The last part is the feature reconstruction module, which reconstructs the enhanced speech signal by combining the output features with the unprocessed noisy phase obtained directly from the input features.

3.1 Pre-processing

In the pre-processing step the goal is to compute an informative feature representation of the input data.

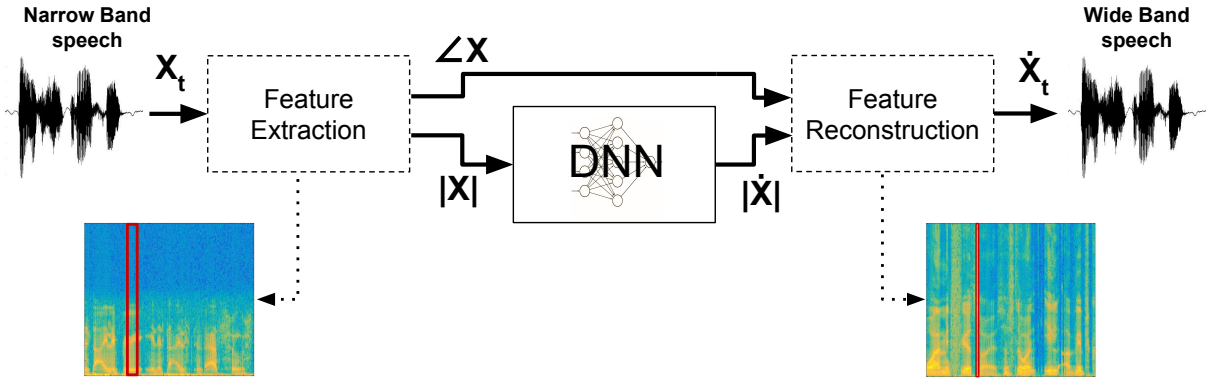


Figure 3.1: Block diagram depicting the high level framework for the complete artificial bandwidth extension system. The algorithm consists of three essential building blocks. A feature extraction and feature reconstruction block to extract and reconstruct the features and a deep neural network block as the core engine of the algorithm

3.1.1 Feature Extraction

Feature extraction is the process of analyzing and extracting relevant information from the input speech signal. It's a method of guiding the network towards solving a specific problem by choosing relevant information which the network should evaluate. Choosing the features that contain relevant information always depend on the task at hand. Feature selection for audio processing tasks often results in the utilization of a spectral representation of the audio signal.

A digitized audio signal is a series of discrete values that represents the sound pressure level at the location of the microphone. It would be challenging for a neural network to learn the relationship between sound pressure level and missing frequency components, which would be necessary for the task of artificial bandwidth extension. But if the network is fed a spectral representation of the audio signal, the task of reconstructing the high frequency components becomes more intuitive.

Figure 3.2 presents a closer look of the feature extraction building block. The building block receives an input which is a raw audio signal sampled at 16 kHz. The audio is analyzed by short-time Fourier transform with 512 FFT points, 75% overlap and a Hanning window to extract the spectral representation. Speech is highly non-stationary, but can be considered as fairly stationary, when looking at a sufficiently short time period (10-35 ms). The FFT size of the spectral transformation is selected to be 512 FFT points on

this bases. With a FFT size of 512 and sampling rate of 16k Hz the time period of a single STFT frame is $512/16e3 = 32ms$. It was verified that the STFT parameters met the constant overlap add constrains, that ensures that a perfect reconstruction can be done.

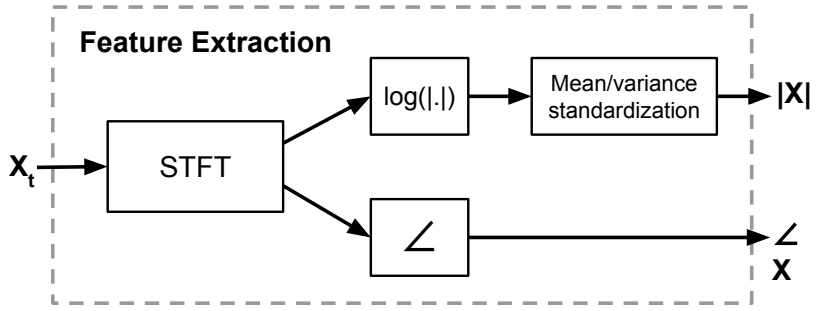


Figure 3.2: Diagram of the feature extraction module. The time signal is processed by a STFT procedure and split into magnitude and phase components. Before the magnitude features enters the neural network, the features are standardized to zero mean and unit variance. The phase component are untouched and passed to the reconstruction stage.

The output of the spectral transformation is split into magnitude and phase components. The logarithm (base-10) operation is performed on the magnitudes and the magnitudes are afterwards standardized.
The phase from the spectral representation output is unprocessed and passed directly to the reconstruction stage.

Feature Standardization

An important feature preparation that needs to be completed before training a neural network is to normalize or standardize the input features. Normalization is often linked to the operation of shifting the data distribution of the features into a range of either [0,1] or [-1,1]. Standardization on the other hand is often linked to the operation of making a distribution more Gaussian. This is done by subtracting the mean and dividing by the standard deviation. In order for the neural network to converge faster the numeric range of the features needs to be in a suitable range for the network. A suitable range is determined by the feature selection and type of activation function used in the network. The most common feature standardization is the zero mean and unit variance

standardization which is found to work great for a wide range of different network architectures and especially architectures based on the gradient decent optimization scheme [Nielsen, 2017]. When deciding on the kind of normalization/standardization to use, it's crucial to inspect the activation function used in the layers. In this project the leaky ReLU activation function was utilized. Because The leaky ReLU doesn't saturate as opposed to the Sigmoid function, it isn't necessary to limit the numeric range of the features. On this basis the standardization procedure was chosen.

The distribution of magnitude features is visualized in each feature extraction step in figure 3.3.

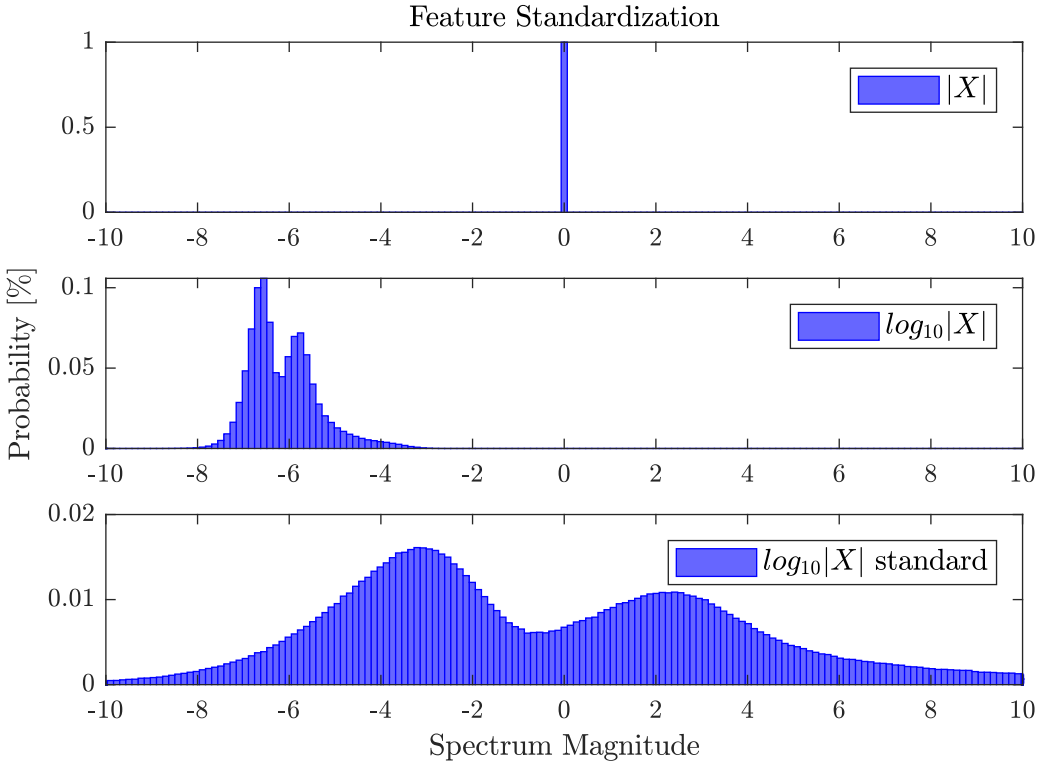


Figure 3.3: Histogram representation of the magnitude values across all frequency bins in each step of the feature standardization.

Before the standardization by zero mean, unit variance a log_{10} operation was performed. It was inspired by the human auditory system, which weights the frequencies similar to a log-scale.The distribution of the magnitude features before the log operation is presented in the top panel of figure 3.3. The log_{10} operation increases the dynamic

range of magnitude values, illustrated in the second panel. By then performing the standardization as seen in the bottom panel of the figure, the dynamic range was further increased.

3.2 Deep Neural Network

The deep neural network (DNN) is the central engine of the algorithm. The DNN performs the actual non-linear mapping between the input and the output features. Two different network structures were used; the feed-forward and the convolutional. The different network structures and the training process is further explained in section 3.4.

3.3 Post-processing

In the post-processing step the goal is to transform the output of the neural network back into the same domain as the input of the system, in this case the time domain. Other processing techniques that enhances the outcome will additionally be utilized here.

3.3.1 Feature Reconstruction

In figure 3.4 a block diagram describing the feature reconstruction procedure is presented. The output of the network is a 257 FFT-point STFT representation of the processed speech signal.

The first step of the reconstruction process is to shift the magnitude values of the distribution back in the original range. This is done by scaling the variance and then inverting the feature standardization operation.

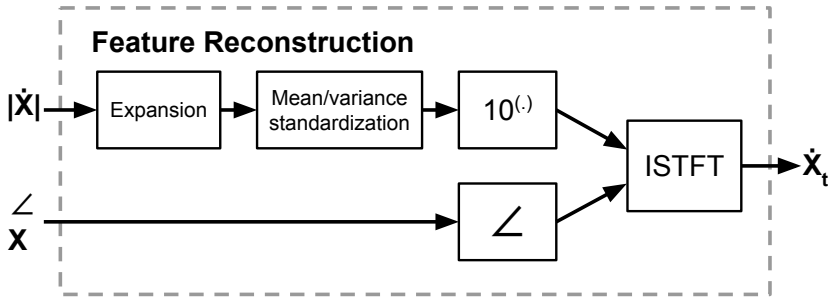


Figure 3.4: Diagram of the feature reconstruction module. To transform the output of the network to a time signal, the procedures of the feature extraction as to be inverted. First the output needs to be transformed into the original distribution of magnitude values. This is accomplished by applying variance scaling (expansion) and inverting the feature standardization. Next the inverse operation of log_{10} is applied and finally the inverse-STFT is performed. The resulting outcome is then a time signal.

Second step is to inverse the log_{10} operation. Final step is to calculate the real and imaginary parts to the inverse Fourier transform from the phases of the original input and the corresponding processed magnitudes.
What is left is an enhanced reconstructed version of the input signal.

Variance Scaling - Expansion

Variance scaling is applied to shift the numeric values of the neural network output to the same natural variance as the target speech signal.
After the inverse standardization it was observed that the variance in the distribution of output magnitude values were lower than the variance in the distribution of the target speech. This discrepancy motivated a scaling of the variance, which was applied on the magnitude distribution before performing the inverse standardization. The variance scaling can be seen as a data distribution expansion, and is simply computed by multiplying the magnitude values with a scaling factor of 1.3. Meaning that a magnitude level of 1 dB was amplified to having a magnitude level of 1.3 dB and a magnitude level of -1 dB was attenuated to having a magnitude level of -1.3 dB. The multiplication factor can also be operated as a exponential factor in the linear spectrum domain.

A study from [Xu et al., 2014] observed that the global variances of the estimated features were smaller than those of the target features. They stated that this indicates

a over-smoothing problem during the DNN training process. They proposed a global equalization factor which was multiplied by the network output log spectrum. The global equalization factor was calculated by the difference between the variance of the estimated features and the variance of the target features.

We found that applying the same global equalization method resulted in a too severe equalization, which decreased the objective speech quality results. We instead found the specific scaling factor empirically using objective experiments in which the speech quality (PESQ) was measured. The scaling factor was chosen on the basis of obtaining the highest PESQ score.

The scaling method was globally applied and resulted in a general attenuation of a large portion of the noise components, due to their low magnitude level. Speech components which have a high magnitude level was on the other hand amplified.
The method blindly enhances the modulations in the entire speech mixture, and does not distinguish between time frames dominated by noise or speech. This can possibly lead to an attenuation of speech components or a possible amplification of noise components with a high magnitude level. Both of which isn't favorable.

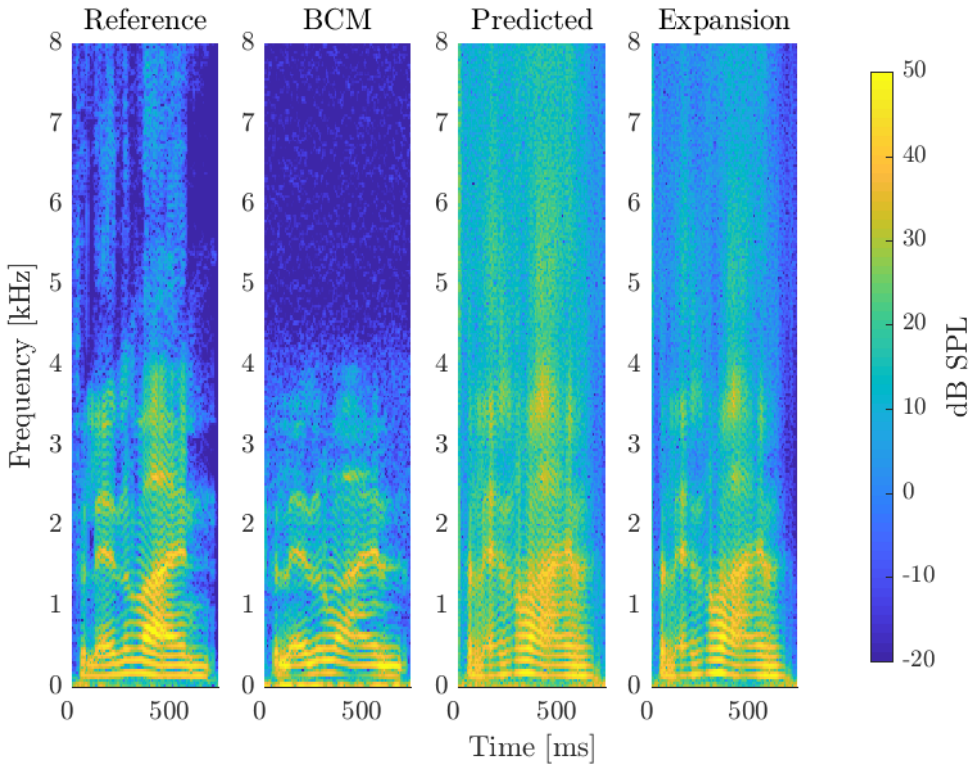


Figure 3.5: Time-frequency representation of a speech excerpt. The speech excerpt is acquired from a real BCM recording. The figure present four different versions of the same excerpt; clean reference, unprocessed BCM, processed (predictions) and processed with expansion. The energy of the background noise was lower for the expansion compared to the predicted, while keeping the same energy level for the speech components.

The process of variance scaling is visualized in figure 3.5. The spectrograms present the variations between the network outputs before and after the variance scaling. The spectrograms show that the processed speech (predicted) without expansion have enhanced a large amount of unwanted noise components. Noise components with a low amplitude, but still audible. By applying the variance scaling and expanding the output magnitude distribution a large portion of the introduced noise components are attenuated. This is illustrated in the expansion spectrogram by the reduced energy of the background noise, while maintaining the same energy level for most of the speech components. Especially the speech harmonics obtain a higher contrast and the distinction between speech and noise components increases. Similar effects were found in [Xu et al., 2014], which reported an effectively sharped of formant peaks of the recovered speech and suppression of residual noise.

These different effects increase the similarity between the clean reference speech signal and the processed speech signal with expansion which results in an increased PESQ score.
Variance scaling is through out the report sometimes denoted as Expansion.

3.4 Training Process

The training process for the neural network models is divided into three main steps. Step one is based on a simplification of the project task. It functions as a proof of concept and has the purpose of determining viable methods for future steps. In the second training step the task is more complex. The task here is to explore and solve a more complex problem based on the viable methods determined and evaluated in the first step. In step three the generalization of the model developed in the second step is explored, and the use of transfer learning is implemented to further improve the performance of the model. The third training step steers the final model towards real data inputs.
An overview of the training steps is shown in figure 3.6, and each step is further described below.

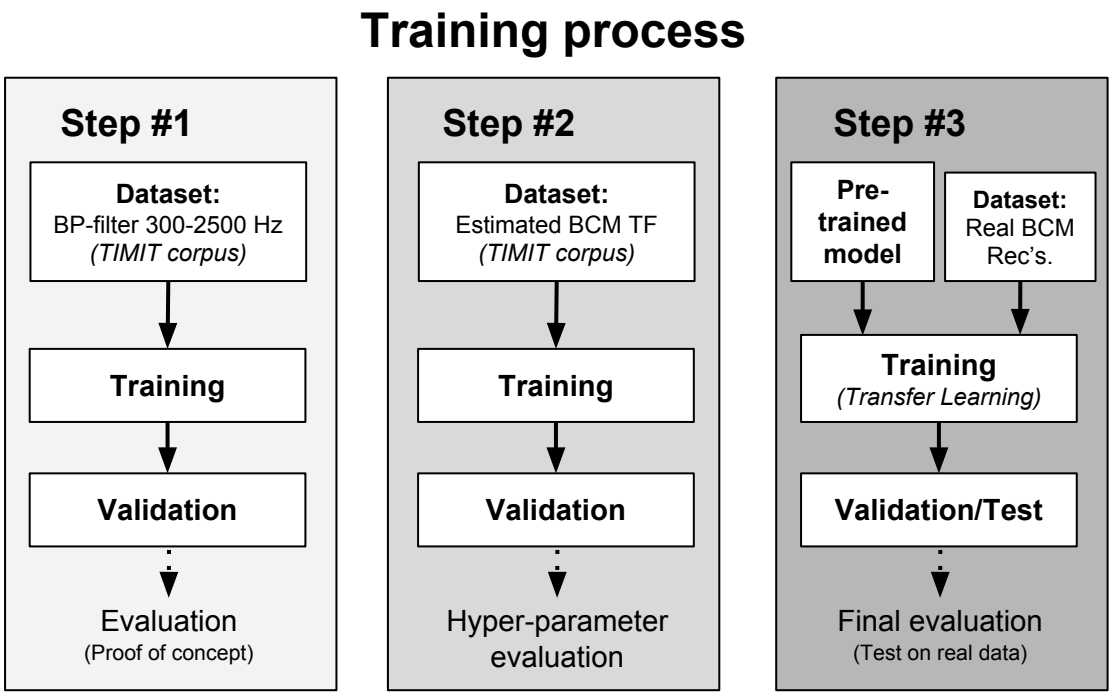


Figure 3.6: The three iterative steps of the training process in the implementation. The first step was a simplified problem and was used as a proof of concept. The second step was more realistic and involved training on a synthetic data set, based on the TIMIT corpus, which approximated the characteristics of bone conducted speech. In the third step transfer learning was applied to use knowledge gained from the first training stage to train and fine-tune the model using a smaller real world data set.

3.4.1 Step One - Initial Training Phase

In the initial step a simplified task was defined from the overall project goal. The task was to train and evaluate simple feed-forward and convolutional networks on a degraded speech corpus.

The speech corpus used was the TIMIT data set. The TIMIT data set has multiple advantages. First of all it has a very usable size it consist of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences [**?**]. It is large enough to get reliable results, while small enough to make it convenient to try many variations of methods. Second of all the TIMIT data set is widely known and acknowledged and has been used to benchmark many existing speech enhancement techniques.

The degradation of the speech corpus was done by applying a simple Butterworth band-pass (BP) filter. The BP filter transforms the speech corpus into a narrow band speech corpus, similar to simulating a telephone transmission line. The BP filters cut-off frequencies were defined as 300 Hz and 2.5 kHz and the filter order was 8. The idea of starting with this simple task was to make a proof of concept and verify that it was possible to artificially expand the bandwidth of the degraded speech signal, and reconstruct the missing frequency components.
Starting with a simplified problem allows for a initial study of the capabilities of a neural network and the influence of basic parameters like size of the training set, feature selection and general network structures. These are all basic parameters which have a crucial influence on the performance of a neural network. The study of parameters in this simplified task, can help determine the correct parameters for the more complex task.

Even though the TIMIT data set is a simple and convenient way to evaluate speech enhancement methods, it should be stated that performance on the TIMIT set does not always reflect a high performance on other data sets. However TIMIT is a good starting point for developing new speech enhancement algorithms.

With the simplified task defined the main goal of the training process was to find the weights and biases that minimized the error between the output of the network and the target speech. The regression problem was solved by processing the input features through the neural network, and afterwards evaluating the output using a cost function. The weights and biases were updated according to the impact they have on the cost function. The mean squared error cost function was used.

The validation was performed after each epoch on a validation data set extracted from the TIMIT data set. An epoch is when the complete training set has been processed. Before selecting a validation data set it is crucial to recall the difference between a training data set and a validation data set. Opposite the training data set which is used adjust the weights and biases of the neural network, the validation data set is not

used to adjust anything. The validation data set is used to verify that the network's performance on training data correlates with the performance on unseen data. Selecting a validation data set which consist of features that the network has not seen before is essential, and is the best way to predict overfitting.

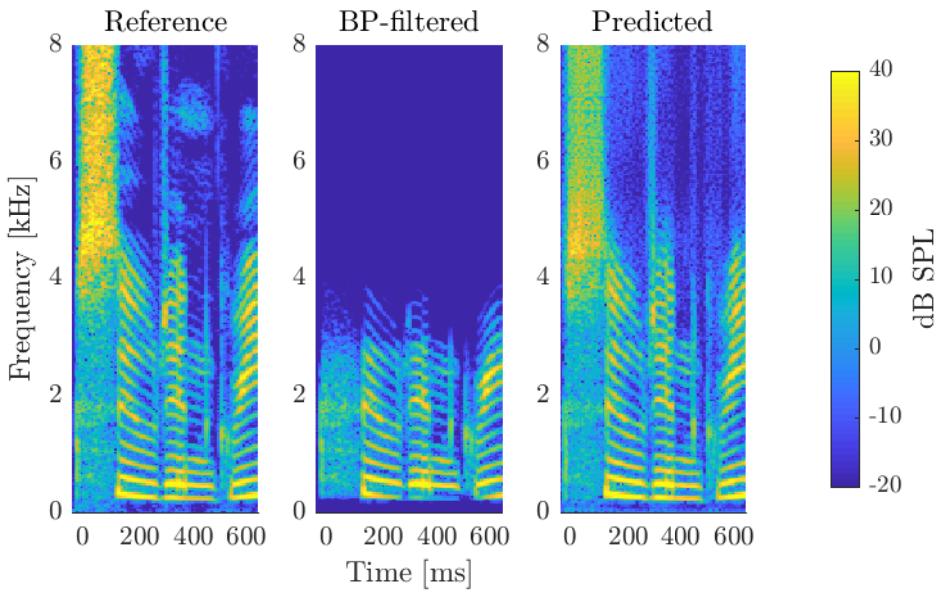


Figure 3.7: Visual inspection of a single validation file. The left spectrogram illustrates the wide-band clean speech signal, the middle spectrogram illustrates the band-limited speech signal and the right spectrogram illustrates the predicted wide-band speech signal. It was observed that the network did perform bandwidth extension.

The evaluation of the initial exploratory network analysis was based on the objective PESQ metric, a visual inspection of the spectrograms and informal subjective listening tests. The PESQ score was calculated from an average of 50 speech files in the validation set and resulted in a relative improvement of *0.17* PESQ points. A visual inspection of the spectrograms showed a significant extension of the bandwidth of the speech. An example is shown in figure 3.7. The visual inspection further showed that the networks did successfully reconstruct harmonic parts of the speech.

This argues that the trained neural network was able to create and furthermore shape the high frequency content which was removed from the input features by the BP-filter. The informal subjective listening inspection further indicated an increase in perceived speech quality.

The exploratory network analysis was only performed as a proof of concept, and no extensive evaluation was therefore performed.

3.4.2 Step Two - Training with Synthetic Data

Based on the positive results from the simplified task in step one a more complex task was defined.
A diverse and representative training data set is an essential part of successfully training deep neural networks and achieving a high performance in real world applications. In many cases it's not easy to obtain such a data set, and it often involves a large amount of time spent on collecting and labeling data. Time which often is costly. To solve the task of this project it would mean recording hours and hours of bone conducted speech, and use this as a viable data set. This was simply not possible in the given time-frame. A way to overcome this issue was to generate a synthetic data set. A synthetic data set that approximates the properties of real data with respect to the degradations imposed by the BCM. The synthetic data set can then be used to train the neural network.
Even though this reduces both cost and time, it also introduces several obstacles which needs to be taken into consideration. First of all does the synthetic data need to reflect the real data. This means that the synthetic input speech signal needs to look and sound as close to the real bone conduction speech signal as possible. Second of all it is crucial to consider the variety of the constructed training set. If the training set is too uniform in form of different types of degradations, it will negatively effect the networks ability to generalize to unseen data.

Extraction of Transfer Function

The generation of the synthetic data set was done by first estimating a set of transfer functions. Transfer functions which relates a recording of speech from a reference ACM to the BCM. Different variations of the transfer function was estimated from a set of real recordings. They were then applied to the speech corpus by the convolution operation. The speech corpus used were once again the TIMIT data set.

The real recordings were obtained in a controlled lab environment, with 22 different

speakers (4 womens/18 male). Each speaker recorded for a total duration of 15 minutes with 3 refitting. Meaning that each recording had a duration of 5 minutes. The specific fitting of the microphone have a significant impact on how the transfer function is composed. By refitting the microphone multiple times a more comprehensive and general representation of the BCM was obtained. In addition to the BCM a reference ACM was placed near the speakers mouth to record the clean reference signals. The total data set of real BCM recordings consists of 66 recordings each lasting 5 minutes.

The MATLAB function *tfestimate* was used to estimate the transfer function between the ACM and the BCM. The MATLAB function is based on Welch's method to estimate transfer functions [Welch, 1967]. The Welch's method estimates the power spectral density at different frequencies and computes a modified periodogram which reduces the noise in the estimated power spectrum, by reducing the frequency resolution. With the ACM signal as input x and the BCM signal as output y the transfer function was estimated as:

$$H_1(f) = \frac{P_{yx}(f)}{P_{xx}(f)}$$

where P_{yx} is the cross power spectral density of x and y, and P_{xx} is the power spectral density of x.

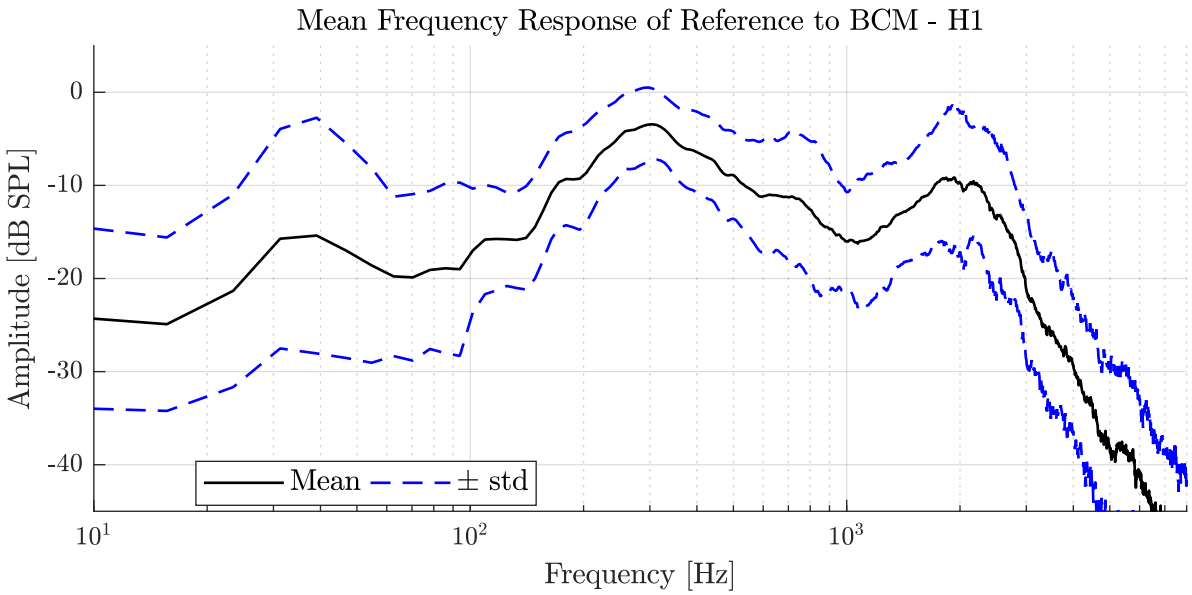


Figure 3.8: Variation in the transfer functions estimated on the basis of the speech corpus recorded through the bone conduction microphone. The transfer function estimate was very speaker depended and changed a lot according to the specific fitting. These large variations are shown by the plotted standard deviations.

Figure 3.8 illustrates the variation in the estimated transfer functions. The variation in the estimated transfer functions are largest for frequencies below 100 Hz. The variation is smaller and more constant for the frequency range above 100 Hz.

From the variation in figure 3.8 it's clear that the transfer function depends strongly on the speaker, and the specific fitting. However the tendencies are pretty consistent. Tendencies as the peak at 300 Hz and around 2 kHz. Furthermore a dip around 1 kHz is pretty consistent.

The synthetic data set should be a more general training set which is a better representation of the real problem.

Noise Estimation

To achieve the best possible performance it is crucial to approximate the real recordings as closely as possible. To be able to do this the bone conduction microphone was inspected. By inspection it was clear that the transfer function extracted above was not the only thing which distinguish and signifies the BCM. From audible inspection a

significant stationary noise component was found. The noise component was similar to white noise and by inspecting the spectrogram the noise was found to be wide-band.
To approximate the real data this noise component was also approximated and added to the synthetic training set. To get the best and the most accurate approximation, the noise component was extracted from real recordings. This means that several silent recordings was recorded. This way the recordings only contained the noise components which were mixed in on top of the synthetic data, at different SNRs. The SNRs was chosen on the basis of audible and visual inspection.

This more complex data set was used in a second training process which also result in a more demanding prediction task. However it's a much better approximation of the real problem and is an essential step towards a high task performance using real bone conduction speech data.

Bias-Variance Trade-off

An important dilemma when considering training data in general is the bias-variance trade-off [Bishop, 2006]. The bias-variance trade-off is of interest when comparing model complexity and amount of training data.
High variance in the model parameters leads to overfitting. In that case the network is too complex to the amount of data present in the training set, which means that the network can't learn the patterns of the data. This issue is often indicated by the training error not correlating with the validation error. The issue of overfitting is further explained in section 3.4.4. To solve the high variance problem it can be necessary to increase the amount of data, not only the size but more importantly the diversity.
A high bias in a model leads to underfitting, and is caused by a model which is too simple or general for a large amount of data, meaning that there isn't enough flexibility in the model to adapt to all the training data. In this case adding more data will not help, and it is necessary to limit the data set. The network can simply miss relevant relations between input and output features.

Hyper-parameter Tuning

Hyper-parameters optimization is an important part of deep neural networks and the large amount of parameters are one of the central reasons why DNNs are difficult to configure. Furthermore individual models are normally slow to train. This means that it is often not time-wise feasible to test all the different parameter combinations. It is therefore essential to choose the most significant hyper-parameters which have the largest impact on the network performance.
Hyper-parameters can be divided into two groups; optimization and model specific parameters. Optimization parameters includes, but are not limited to, learning rate, mini-batch size and total number of training steps. Model specific parameters includes the number of layers, number of units, type of activation function, type of regularization and weights initialization.
To limit the amount of time spent investigating the different hyper-parameters only a few were chosen to be further investigated with a grid-search.
According to [Bengio, 2012] learning rate is the single most important parameter to tune, as it controls how the optimization algorithms moves around the weight-variable space. A too large learning rate can result in overshooting and makes finding the global minimum of the cost function difficult, which makes convergence hard. A too small learning rate can on the other hand result in slow convergence, and the chance of ending up in a local minimum increases.

Based on informal pilot tests the second hyper-parameter chosen to tune was the number of layers. The number of layers are a complicated thing and there aren't any specific rules on the subject. However there is a few guidelines on the subject. One hidden layer is often suffice for simple data sets, and with more complex data sets additional layers may be valuable. The network is able to learn and adapt to more complex data structure when containing multiple layers, which allow the network more flexibility. However a large number of hidden layer also leads to challenges. By increasing the number of layers the neural network gets much more complicated to train [Nielsen, 2017]. This is primarily due to the vanishing gradient problem. The vanishing gradient problem occur from using the gradient descent optimization scheme. The weights are updated backwards in

the back-propagation scheme, by calculating the gradients of the errors with respect to the weights. The specific problem occurs because the calculated gradients get smaller as the scheme moves backward in the network. This result in a slower learning process for the neurons in the first hidden layer compared to the learning process in the later layers.

It's important to notice that a full multi-dimensional grid search of all the hyper-parameters are the most thorough way of optimizing a neural network, but the amount of time and computation power needed to do such an experiment was not doable in this project. The grid-search experiment description and results are found in section 4.1 and in section 5.1.

3.4.3 Step Three - Training with Real Recordings

The third and final step aims to solve the overall goal of extending the bandwidth of real recordings of the BCM. Even though the performance of a neural network trained on synthetic band-limited speech data can be interesting from an academic point of view, it is rather uninteresting if the performance is not reflected in a real world setting. Meaning that if the performance decrease significantly when the network is exposed to unseen real world bone conducted speech signals. Since it is close to impossible to construct a synthetic data set which is indistinguishable from real world data, it is expected that the performance of the pre-traind neural network won't be able to generalize to real world inputs. And informal listing test confirmed that theory.
To solve this issue the obvious choice would be to use real world recordings as training data instead of the synthetic data. But he idea behind generating synthetic data was because there wasn't enough real world data. A solution to the restriction in the amount of training data could be to utilize the means of transfer learning. The best feed-forward and convolutional network configuration from the two grid searches were picked and laid the basis for models covered in this section. And thus, the third and last training step revolved around utilizing transfer learning to improve the network's generalization ability towards real world data.

Transfer Learning

Transfer learning is a method where knowledge gained from solving one problem can be utilized and applied to a related problem [Torrey and Shavlik, 2009]. This means that the neural network which is trained on the synthetic data set can be adapted through a new training process and applied to a real world task. This method allows the network to first learn the coarse-grained features using the large synthetic data set, and afterwards using the small real world data set to learn the fine-grained features which fine-tunes the performance of the network. This type of knowledge transferring can be considered as a form of network weight initialization.

Transfer learning were applied for both the network configurations. The process consists of two steps. First the models are trained on the synthetic data set. The idea behind this step is that the models will learn to extend the bandwidth bandwidth of many different speakers, but the band-limitaion is less varied than in with the real world data. The amount of training is controlled by an early stopping criteria. The stopping criteria was tracking the epoch with the lowest validation error. If the validation error didn't drop below the previous lowest within ten epochs, the training would stop and the weight configuration corresponding to the epoch with lowest validation error was saved. In the second step the models were trained on the data set of real world recordings from the BCM. This data set consists of significantly fewer speakers than the synthetic set, but the variation in degradation and SNRs are larger. The idea is that the models have learned the basic overall patterns to extend the bandwidth of signals, but the variations in the degradations were limited. In order to fit the models to the actual use case the real world data was used to fine-tune the models.

Both models were validated by a subset of the real recordings and the training were also controlled by the same early stopping criteria. To further increase the performance different generalization techniques were used. These are describe in the next section.

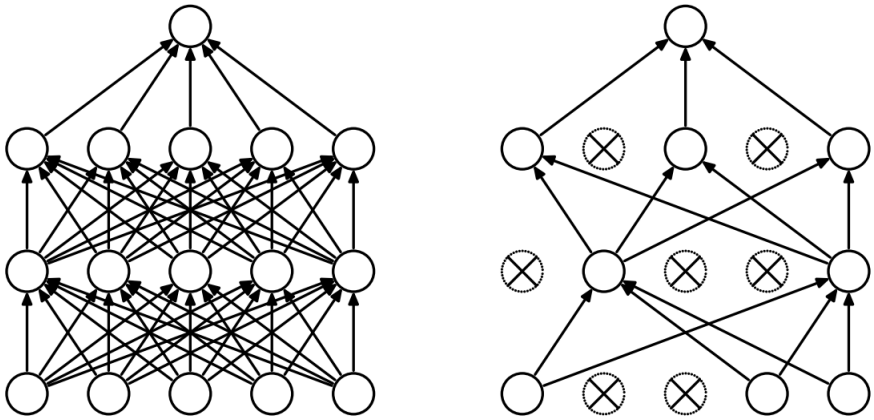
3.4.4 Generalization Ability

A term that has been the focus of the whole project was generalization, which is the ability to perform well on unseen data. This ability can be reduced by a number of things the most important being overfitting. Overfitting occurs when the model is matched too perfectly to the training data and that way learns specific correlations which only occur in the training data. When applying the model to a test set, the model don't perform as expected. Overfitting is an important challenge in deep learning, and has been the focus of countless research papers and is crucial avoid.

The most obvious solution to this problem was to add more data, which were why transfer learning has been utilized. But other methods are also needed to further increase the performance. Some of these methods are known as regularization [Goodfellow et al., 2016].

Dropout

Dropout is another powerful way to avoid overfitting. Dropout has received a lot of attention over the last years and is popular due to its simplicity and efficiency. The key idea is to randomly discard neurons and their corresponding connections temporarily in each training step during the training phase [Srivastava et al., 2014]. Discarding neurons and their corresponding connections prevents the network from forming complicated co-adaptions during training. This restrains the network from only using a subset of the neurons and forces the network to use all the weights. By using all the weights the network is more likely to generalize to new unseen data in the testing phase.



(a) Standard Neural Net (b) After applying dropout.

Figure 3.9: The process of dropout [Srivastava et al., 2014]. During the training process neurons and their corresponding connection are temporarily removed from the network architecture to prevent overfitting and improve generalization.

Figure 3.9 illustrates the process of dropout. Figure (a) illustrate a standard feed-forward neural network with an input layer, two hidden layers and an output layer. Figure (b) illustrates an possible example of a thinned version of the same neural network during a training step. The thinned version is produced by applying dropout to the network. Such a thinned version of the network is randomly produced for every weight update. Many of the thinned versions produced during the training process will contain the same neurons along with the same incoming and outgoing connections, meaning that every weight will with high probability be updated multiple times during the training process. It is hypothesized that by presenting a changing network structure for every weight update, the individual neurons will be forced to learn internal features independent of the rest of the network.

Another way of understanding the idea behind the dropout technique, is to think of dropout as an averaging procedure. Randomly discarding different sets of neurons correspond to training a large amount of different neural networks. The dropout procedure correspond to averaging the effect of a large amount of different neural networks.

Dropout is often not the only regularization technique used to improve generalization and is usually combined with other regularization techniques.

Chapter 4

Experimental Evaluation

In this chapter a series of experiments for evaluating the ABE algorithm is presented. The experiments are split into three main categorizes, the pre-experiment, the objective experiments and the subjective experiments. The objective evaluation includes two experiments and is performed for both the CNN and the FFN network structures and includes the use of PESQ, MCC and STOI. The subjective experiments consists of two listening experiments. One experiment for evaluating the perceptual quality of the output and one for evaluating the speech intelligibility.

To conduct a thorough evaluation of the ABE methods both objective and subjective metrics were used. Objective evaluation is a metric of mathematical comparison of the original and processed signals. It's a way of quantifying speech quality or speech intelligibility improvements via numerical distance between the original and processed speech. Subjective evaluation quantifies the same things but by the means of human listeners [Loizou, 2013]. For a objective evaluation metric to be valid it needs to correlate well with the subjective evaluation results. A way of accomplishing this is by constructing the objective metric to reflect certain aspects of the human auditory system.

4.1 Preliminary Experiment - Network Complexity

When the overall structure of the network has been chosen and the data set has been collected and prepared, the final network configuration needs to be determined. This

process determines the complexity of the network and involves tests and evaluations of several different combinations of network configurations. This is an comprehensive process and the most time-efficient way to do that was to perform a grid search.

4.1.1 Experiment 1 - Hyper-parameter Tuning

The purpose of the grid search was to identify the optimal hyper-parameters combinations for the different neural network configurations. A grid search was performed both for the feed forward and the convolution model structure. The PESQ metric was used as the primary evaluation metric in the grid search. The synthetic TIMIT data set with added noise was used as the training data *(section: 3.4.2)*. The grid search looped through number of layers ranging from 1 to 5 and learning rates $1e^{-3}, 1e^{-4}, 1e^{-5}$ and $1e^{-6}$. The different configurations was controlled by an early stopping criteria evaluated on the validation error. The evaluation was performed on the unseen validation data set.

Table 4.1: Overview of the most important hyper-parameters for the models in the grid search.

Grid Search	**FFN**	**CNN**
Learning Rate	Variable	Variable
Fully Connected Layers	Variable	2
Neurons in FC Layers	2048	2048
Convolutional Layers	N/A	Variable
Filters in Conv. Layers	N/A	32
Dropout	25%	25%
Activation Function	Leaky ReLU	Leaky ReLU
Batch Size	32	32

4.2 Objective Experiments

The objective evaluation consisted of three different metrics. PESQ was used as in the pre-experiment. Furthermore the magnitude correlation coefficient (MCC) and STOI were utilized to get an objective measure of speech intelligibility.

A combination of these three objective metrics evaluates the different aspects of speech enhancement algorithms, and are a comprehensive way to objectively quantifying the performance of the artificial bandwidth extension algorithm.

4.2.1 Experiment 2 - Objective Evaluation - PESQ and STOI

To evaluate the optimal network type and the different aiding methods, the objective evaluation metrics were used on a test set. The test data set was a subset of the real BCM recordings. It contained 20 percent of the total data set and included 12 recordings, spoken by 4 speakers. Each speaker recorded 3 sentences, where the bone conduction microphone was refitted to obtain a new variation of the transfer function. The test data set was never seen by the trained model. This is one of the fundamental rules for a test data set in order to be credible.
The objective experiments tested a series of different network configurations. First of all the two main network types (FFN and CNN) were tested and compared against the unprocessed BCM recordings. And for each of the two main network types four different algorithm configurations of aiding techniques, with and without transfer learning and variance scaling (expansion). The different network configuration are illustrated in table 5.1 along with the results of the objective metrics.

4.2.2 Experiment 3 - Bandwidth Extension - MCC

To evaluate and examine the actual bandwidth extension more comprehensively the MCC metric was used. The MCC metric is a simple spectral correlation measure and is valuable in the process of examining the similarity between the spectrum of two signals. The metric gives the correlation coefficient at the different frequency bands which means that the improvement for isolated frequency components can also be assessed. A modification to the algorithm was implemented in order to get a make a more detailed analysis. In the original implementation the time-frequency (T-F)-representation is decompositioned into critical bands. This decomposition was omitted in order to get a higher frequency resolution and enable a closer inspection of the bandwidth extension.
The experiment investigated the correlation between the test data set under different

network configurations and the corresponding reference speech signal. The test data set was the same as experiment 2.
The network configurations under test consisted of two main network types, the FFN and the CNN along with the expansion technique. The different network configurations were compared to the unprocessed BCM speech signal. The correlation coefficients of theses different configurations were based on the mean values across the complete test set.

4.3 Subjective Experiments

It is important to emphasize that the need of subjective evaluation is essential as a final evaluation. If the quality/intelligibility improvements are not seen with listening subjects the actual effect of the ABE algorithm would have to be re-evaluated. The subjective evaluation was performed on the basis of two different metrics.
The first subjective evaluation experiment was a preference test and evaluated the speech quality of the processed and unprocessed speech signals. The second subjective experiment was a speech intelligibility listening test and evaluated the speech intelligibility of the processed and unprocessed speech signals.

4.3.1 Experiment 4 - Speech Quality - MUSHRA

To evaluate the perceptual sound quality of the ABE algorithms the subjective MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) method was used. The MUSHRA is a method for subjective assessment of intermediate audio quality and is described in the ITU-R BS.1534-1 Recommendation. The MUSHRA test uses an unprocessed recording as the global reference, hidden reference, an anchor which is a low-passed version (cutoff at 3.5kHz) of the reference and degraded signals that are under test. The advantage of MUSHRA over the commonly used MOS test is that the subjects are presented with all of the conditions at a time. This allows the user to make direct comparisons of the conditions and it's lowers the demand of remembering the full range of degradatios from excerpt to excerpt.
In the MUSHRA test the subjects are asked to rate all the speech excerpts on a continuous scale from 0 to 100 where 0 is bad and 100 is excellent. The subjects are listening

for the *Basic audio quality*, which includes all degradation's and artifacts present in the signal. This gives a single value that describes the difference between the conditions under test and the reference.

Experiment

This experiment was constructed in order to assess the difference in perceptual sound quality of the processed speech signals. The null hypothesis (H_0) was that there were no difference, on average, between the perceptual quality of the BCM signals and the processed signals. The alternative hypothesis (H_a) was that there were a difference, on average positive or negative, between the perceptual quality of the BCM signals and the processed signals. In order to reject H_0 in favor of H_a the significance level α was set to 0.05.

The experiment was two-fold. One experiment was conducted under controlled conditions using DTU's facilities. The other experiment was an online version of the MUSHRA test. Identical in every way, but it wasn't controlled. A website-link was sent to every subject providing an online MUSHRA test along with guidelines for the participation in the experiment.

Listeners

The subjects were split into two groups. The first group included 8 listeners, screened by a professional audiologist and found to be normal hearing. Every subject were native danish speakers and every subject signed an informed consent document, and were reimbursed for their efforts.
The second group were a select group of experienced listeners from the Hearing Systems group at DTU and employees from Invisio Communication A/s.

Stimuli

The stimuli consisted of 11 selected sentences from the BCM recorded validation set and one training sentence from the TIMIT database. Each sentence were truncated to a maximum duration of 10 seconds and adjusted to 60 dB SPL. The unprocessed BCM

signal was processed by the respective ABE algorithms. No noise was added, so only the self-noise of the BCM was present. The ABE algorithms under test were the CNN with and without expansion, and the FFN with and without expansion.
Based on an initial pilot experiment of 4 persons, the traditional low-pass anchor, was omitted. The anchor normally has the role of being the lowest quality signal presented. This gives the subject a range between the best and the worst quality. But in this case the unprocessed BCM signal was significantly worse in quality than the anchor, which resulted in a lower dynamic range of effects. The unprocessed BCM was therefore in this experiment used as the lower bound of the quality range.

Experimental Design

In the MUSHRA test the subjects were asked to first listen to the reference signal, then listen to each of the presented conditions and rate each of the condition according to the perceived quality.

Figure 4.1: Graphical user interface of the online MUSHRA test. The interface allowed the subject to freely switch between the different conditions, listen and rate the quality of the sentences under each condition.

The sound excerpt was looping and the subject could freely switch between the different conditions. When all conditions were rated the subject moved on to the next sentence. At any time the subject could go back to previous sentences and alter their response and if necessary take a break. At the end of the experiment the subjects were asked to

state their nationality and gender.

Procedure and Apparatus

All signals were processed in Python 3.6 with a sampling rate of 16 kHz. For group one the playback was over a Sennheiser HD 650 headset on a HP Notebook PC with integrated sound card. For group two the equipment couldn't be reported as this was part of the strategy to get as many listeners to participate as possible. The participants were asked to use good quality headphones but since it was not possible to oversee this we can't be sure.

4.3.2 Experiment 5 - Speech Intelligibility - DANOK

The DAnsk Nonsens Ords Korpus (DANOK) is a type of corpus used to evaluate the subjective speech intelligibility by consonant recognition. We refer to this experiment as DANOK.

Experiment

The experiment was constructed in order to test if the speech intelligibility was better or worse after the speech was processed by the different networks. The null hypothesis (H_0) was that there were no difference. The alternative hypothesis (H_a) was that there were a difference, positive or negative. In order to reject H_0 in favor of H_a the significance level α was set to 0.05. The DAnsk Nonsens Ords Korpus (DANOK) speech corpus were re-record with the BCM. The consonant recognition test was performed on the re-recorded speech corpus and the speech intelligibility was evaluated on the unprocessed BCM speech and on the processed ABE configurations.

Listeners

The same eight native, normal hearing, danish listeners from experiment 4 also participated in this experiment. They signed an informed consent document, and were reimbursed for their efforts.

Stimuli

The corpus consisted of 15 consonants followed by the vowel /i/ spoken by a female non-professional speaker. The sound levels of the 15 CV combinations were equalized using the VUSOFT method developed by Lobdell and Allen (2007), which was also used in [Dau, 2017]. This technique ensures that the sound level of the vowels stay close to constant, over different CVs, but the sound levels of the consonant parts change, this is close to the dynamics of natural speech.
Recordings of a helicopter recorded through the BCM were used as masking noise. Different SNR scenarios (25 dB, 20 dB and 15 dB) were created by fixing the level of the speech to 60 dB sound pressure level (SPL) and adjusting the noise to the desired level.

The re-recording of the corpus was conducted according to the method described in [Christiansen, 2011] with a few modifications. The corpus consisted of 6 lists with 11 CV combinations. All CV combinations were followed by the nonsense word *tu*, in order to make the pronunciation more natural.
Each list was initiated by a prompting sentence "Nu bliver der sagt" followed by three repetitions of the first CV:tu, then four seconds of silence and then the next three CV:tu's and so on. After a list was recorded a small break was taken where the speaker was able to drink some water. After recording the hole corpus each recording of the triplets (three repetitions of CV:tu combination) were screened for inappropriate noises such as mouth sounds, scratches, sound from moving etc., and re-recorded if the noises were too inappropriate.

Experimental Design

The experiment consisted of four trials; One training and three tests. Each test at a different SNR (25 dB, 20 dB and 15 dB). In a single trial the 15 CV combinations were repeated six times for the three different conditions (BCM, CNN and CNN with expansion) resulting in a total of 270 CV combinations. 1080 CV presentations in total during the entire experiment. After each trial a break was taken.

Procedure and Apparatus

The experiments was conducted in an audio visual immersion lab (AVIL) Psy booth at DTU. In the booth a Sennheiser HD 650 headset and a computer terminal showing the graphical user interface was present. The presentation sound levels were adjusted to a comfortable level for each subject. The subjects would listen to the stimuli and immediately after chose the consonant corresponding the the perceived stimuli. After each trial the break was taken in order for the subject to recover and clear their mind.

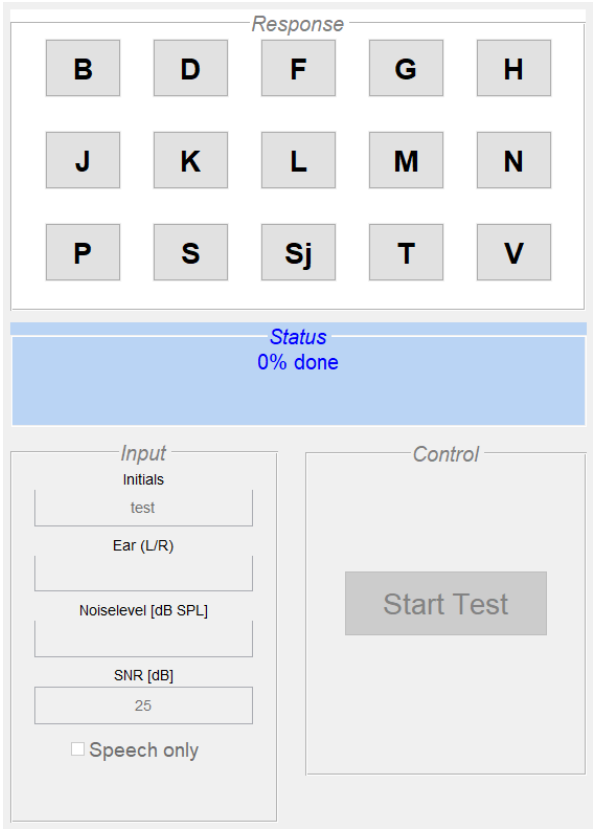


Figure 4.2: Graphical user interface of the DANOK experiment. The interface was provided by Johannes Zaar and were similar to the setup used in [Dau, 2017].

Chapter 5

Results

In this chapter the results from the evaluation experiments is presented. Like the previous chapter, the experiments are split into three main categorizes, the preliminary experiment, the objective experiments and the subjective experiments.

5.1 Preliminary Experiment - Network Complexity

5.1.1 Experiment 1 - Hyper-parameter Tuning

The hyper-parameters are determined and evaluated on the basis of the performance of the models trained on synthetic data set, developed in the second training step. The different model configurations were analyzed and evaluated in a grid search. There are infinitely many different model configuration combinations to choose from. This project focused on learning rate and number of layers. The results from the grid search (preliminary-experiment) are presented in figure 5.1 for the FFN model and in figure 5.2 for the CNN model.

It is clearly illustrated in the figures that the learning rate have a large impact on the performance of the network. The learning rate is the factor which determines how much the weights of the network are adjusted with respect to the gradient of the cost function in each training step.

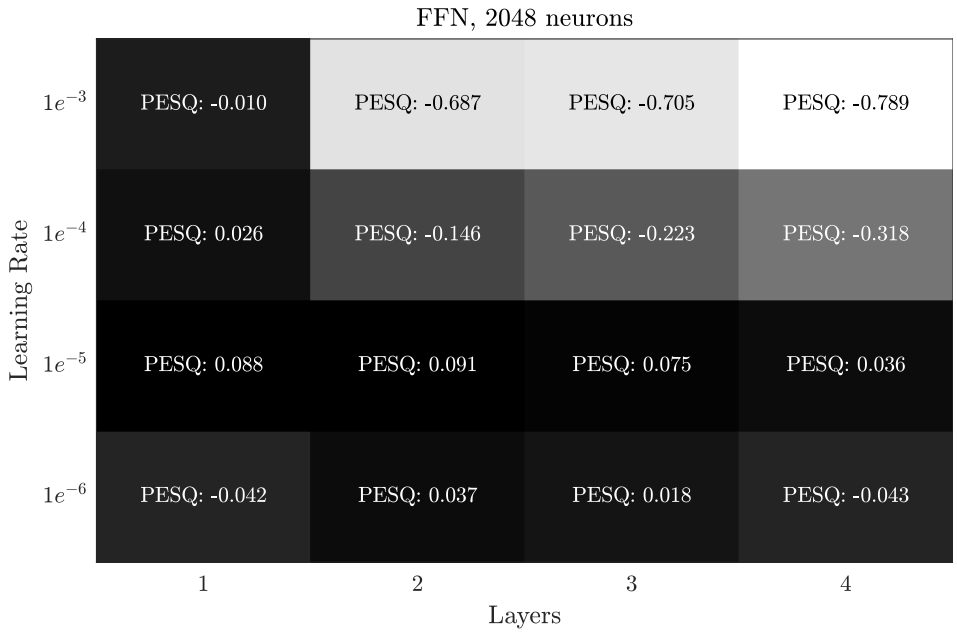


Figure 5.1: FFN grid search results from the preliminary experiment. The focus was on learning rate and the number of layers. Every model was trained with 2048 neurons in each layer.

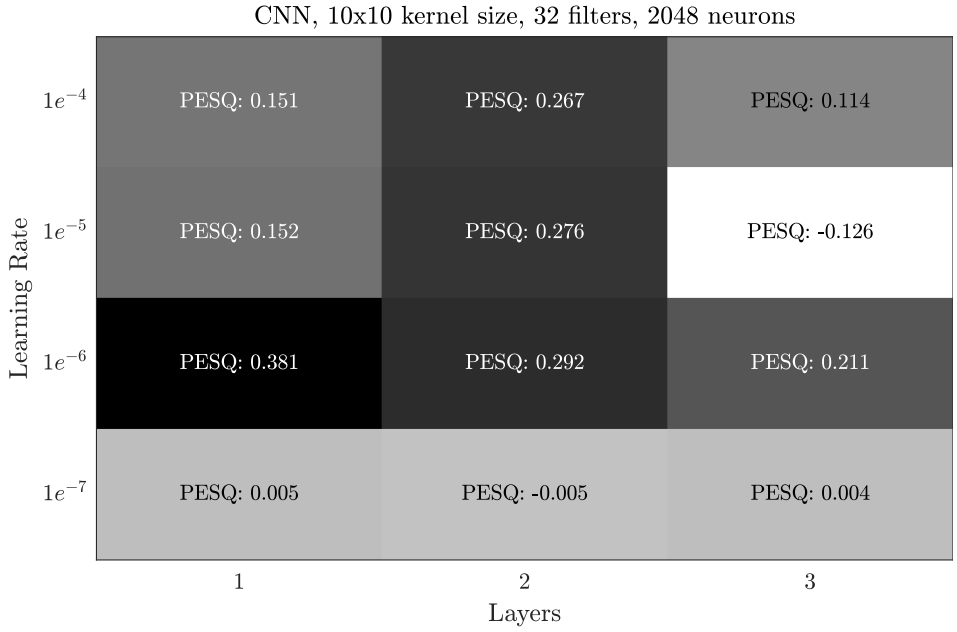


Figure 5.2: CNN grid search results from the preliminary experiment. The focus was on learning rate and the number of convolutional layers. Every model was trained with a kernel size on 10x10, 32 filters in each convolutional layer and with 2048 neurons in each fully connected feed-forward layer.

The grid search showed that for the FFN model, the best configuration was with a learning rate of $1e^{-5}$ and two layers and for the CNN model the best performing model had a learning rate of $1e^{-6}$ and one convolutional layer.
For the FFN model the general trend was that the performance was lower for configurations with many layers and low learning rate. The trends were harder to point out for the CNN and the grid search showed larger variations between the configurations.

5.2 Objective Results

5.2.1 Experiment 2 - Objective Evaluation - PESQ and STOI

The results of the objective experiments are presented in table 5.1. The mean value across the test set of each metric is presented and the best performing configuration is highlighted in bold. The first row in table 5.1 correspond to the performance of the unprocessed BCM recordings, the baseline.

Table 5.1: Results for the objective experiment performed on the the different neural network configurations. Variance scaling is noted as Expansion. *Note: Results are presented as the mean value across speakers and sentences of each metric.*

Model	Transfer Learning	Expansion	PESQ	MCC	STOI
BCM *(baseline)*	N/A	N/A	1.49	0.40	0.78
FFN	yes	yes	1.92	0.68	0.83
		no	1.71	0.67	0.83
	no	yes	1.93	0.66	0.82
		no	1.72	0.67	0.83
CNN	yes	yes	**2.30**	0.70	0.83
		no	2.17	**0.72**	**0.85**
	no	yes	2.01	0.70	0.83
		no	1.85	0.70	0.84

Studying the PESQ results of the objective experiments all FFN model configurations and all CNN model configurations significantly improved the PESQ score. The FFN configuration without any expansion scored $1.71 - 1.72$ which is an improvement of

$0.22 - 0.23$, and when applying the expansion method the improvement increased to $0.44 - 0.45$. An interesting thing regarding the PESQ improvement for the FFN configuration is the little effect which influence of prior knowledge have. There was no significant difference between the model with pre-trained weights and biases and the model without.
The CNN model configurations out-performed the FFN configurations significantly, and pre-training the weights and biases by the use of transfer learning had opposite to the FNN an significantly positive effect. For the model without any pre-trained weights and biases the improvement from the baseline was 0.36 without expansion and 0.52 with expansion. When pre-training the model on synthetic data the PESQ improvement rose to 0.68 without expansion and 0.81 with expansion. These are significant improvements which were audible.

The results from the MCC measure presented a some what similar story. There were a significant improvements from the unprocessed baseline to the processed model configurations, for both FFN and CNN. The MCC improvements varied between $0.26 - 0.32$ and the different model configurations did not effect the MCC performance significantly. As the MCC metric was primarily used to measure how well the frequency content of the processed/unprocessed speech signals correlates with the clean reference speech signal, the mean value improvement was not that interesting. Instead the MCC results for the different model configurations will be presented on a more frequency specific level in section 5.2.2.

The results from the STOI metric did not show any significant improvement. The largest improvement over the BCM was 0.07 which was well within the variations of the observations. Among the different network configurations the largest difference was only 0.02.

5.2.2 Experiment 3 - Bandwidth Extension - MCC

The correlation between the spectrum of the different processed speech signals and the reference speech signal has been evaluated in experiment three by the MCC metric.

Figure 5.3 presents a comparison of the correlation results obtained for the unprocessed BCM baseline and for four different model configurations for frequencies ranging from 50 Hz to 8 kHz. There was a high correlation between the unprocessed baseline and the reference speech signal for frequencies between 300 Hz and 550 Hz. The correlation drops between 600 Hz and 1.5 kHz. Around 5 kHz the correlation drops close to zero.

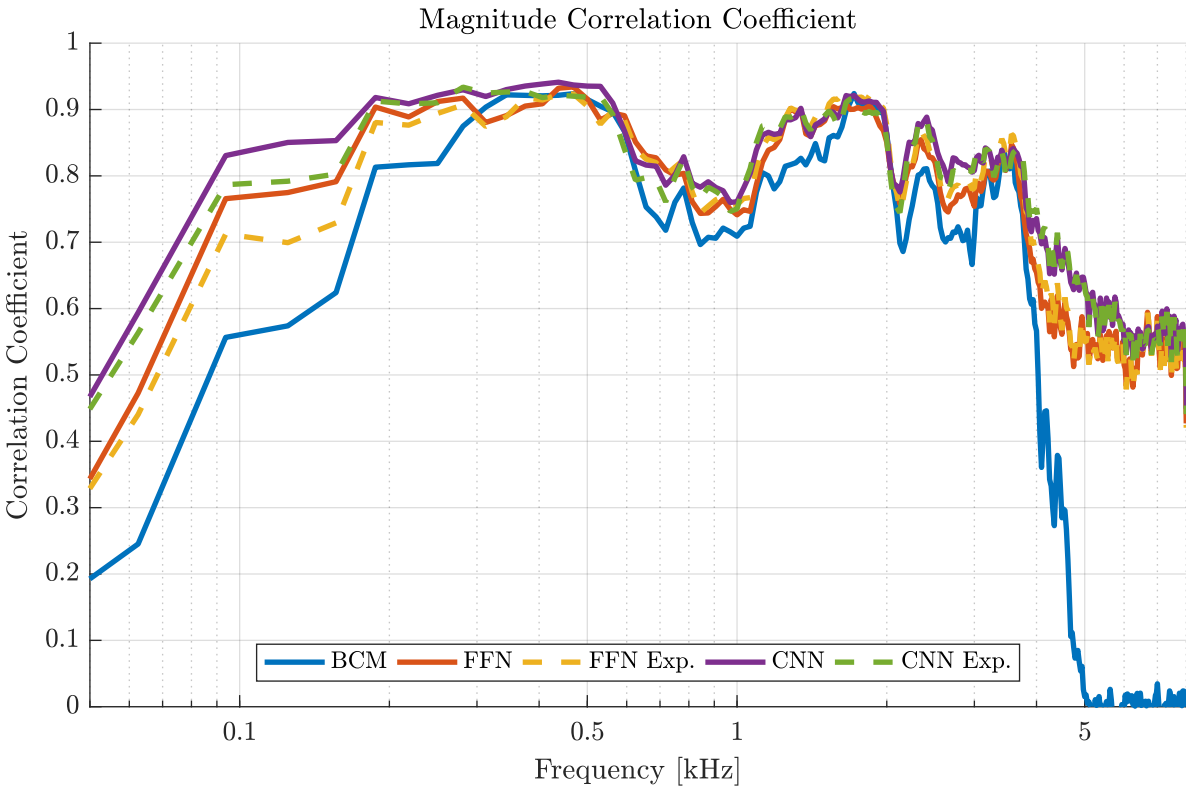


Figure 5.3: Comparison between the different magnitude correlations versus frequency for the different network configurations and the unprocessed BCM baseline. All network configurations showed higher correlation with the reference signal than the unprocessed BCM baseline.

The correlation between the processed speech signals *(FFN, FFN Exp., CNN and CNN Exp.)* and the reference speech signals showed very similar behavior as the unprocessed baseline for the frequency range between 300 Hz and 4 kHz. However, the results from the processed speech showed a small improvement, particularly in the dips.

The correlation for the processed speech signals compared to the unprocessed baseline presented a significant improvement for frequencies below 300 Hz. The correlation was

likewise significantly improved for frequencies above 4 kHz.
The correlation for the CNN and CNN Exp configurations were nearly identical to the correlation for the FFN and FFN Exp. with a small positive offset. Meaning the correlation value for the CNN and CNN Exp. was a bit higher
The significant improvements presented for frequencies above 4 kHz, could indicate a partial reconstruction of high frequency content.

5.3 Subjective Results

5.3.1 Experiment 4 - Speech Quality - MUSHRA

The results from group one of the MUSHRA test are presented in figure 5.4. The mean results across all sentences are reported along with the 95% confidence interval.
The reference condition was as expected clearly the most preferred and scored close to 100 with very little variance. The results from the speech processed by the different models were more varied. Three out of four networks scored a higher preference rating than the unprocessed BCM speech condition. The three networks were the FFN with expansion (Exp.), the CNN, and the CNN with expansion.
The FFN model without expansion was the only configuration which scored lower preference rating than the unprocessed speech condition and was also the model with the lowest score.

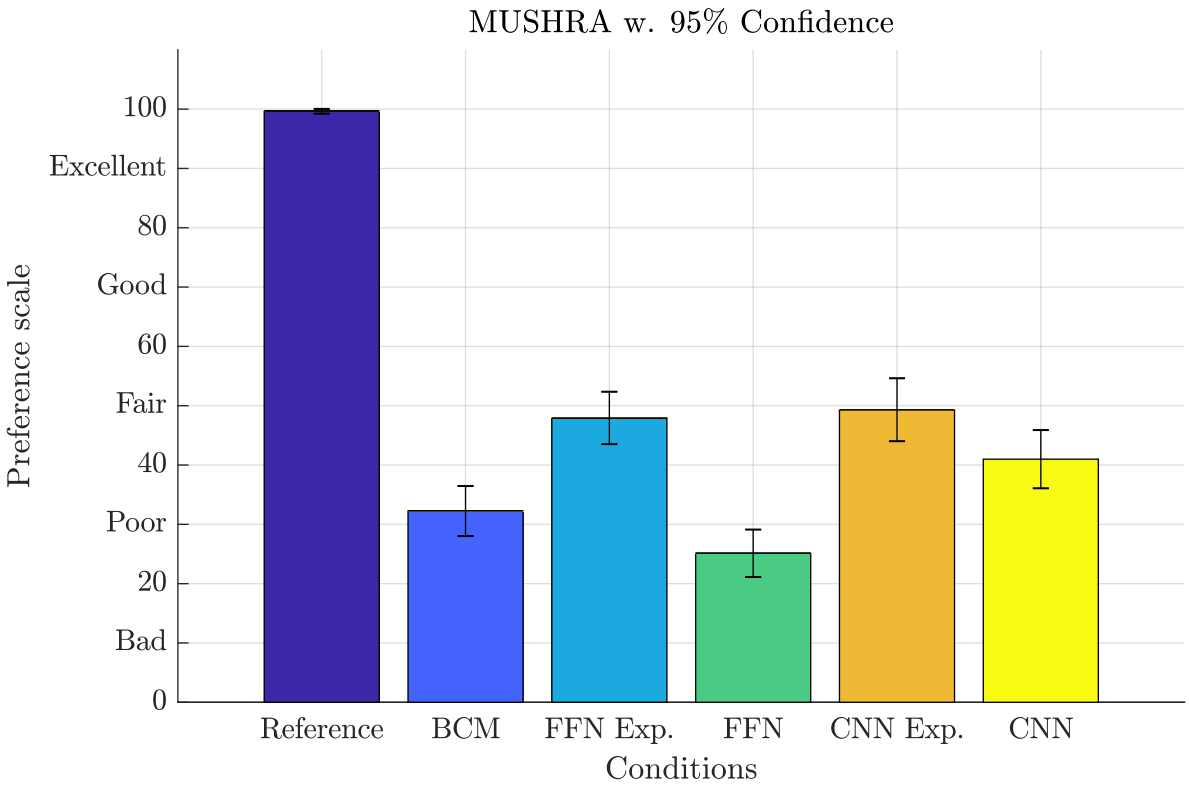


Figure 5.4: The average results from the *controlled* MUSHRA preference test across all test sentences. The error bars mark the 95% confidence interval. The reference was clearly preferred. The three network types FFN Exp., CNN Exp. and CNN were significantly preferred over the unprocessed BCM signal.

In order to draw a conclusion to either accept or reject the null hypothesis a statistical analysis was performed. The test was based on a two-sample t-test. The largest improvement over the BCM was observed for the CNN network with expansion at around 20 percentage points. From the statistical analysis seen in table 5.2 it's observed that the three network conditions that scores higher than the BCM were all statistic significantly higher. And the FFN configuration was significantly worse.

Between the three networks that scored higher than the unprocessed speech signals, none of them were significantly different from the others.

The results from the online MUSHRA test showed a slightly different picture than the controlled test *(fig: 5.5)*. The results indicated that the quality of the speech processed

Table 5.2: Statistical analysis of the controlled (Booth) and the uncontrolled (Online) MUSHRA experiments.

	Booth		**Online**	
	H_0 **rejected**	**p-value**	H_0 **rejected**	**p-value**
FFN Exp.	true	$1.5 \cdot 10^{-06}$	true	0.014
FFN	true	0.018	false	0.451
CCN Exp	true	$2.4 \cdot 10^{-06}$	true	0.036
CNN	true	0.009	true	$1.5 \cdot 10^{-05}$

by the same three network configurations (CNN Exp., FFN Exp. and FFN) scores significantly higher than the unprocessed BCM speech. However the significance level was lower compared to the controlled experiment.

A surprising result was that the FFN without expansion received a higher speech quality rating than the unprocessed BCM speech. Even though it ain't significant it was noticeably different from the controlled experiment result, which scored lower than the BCM.

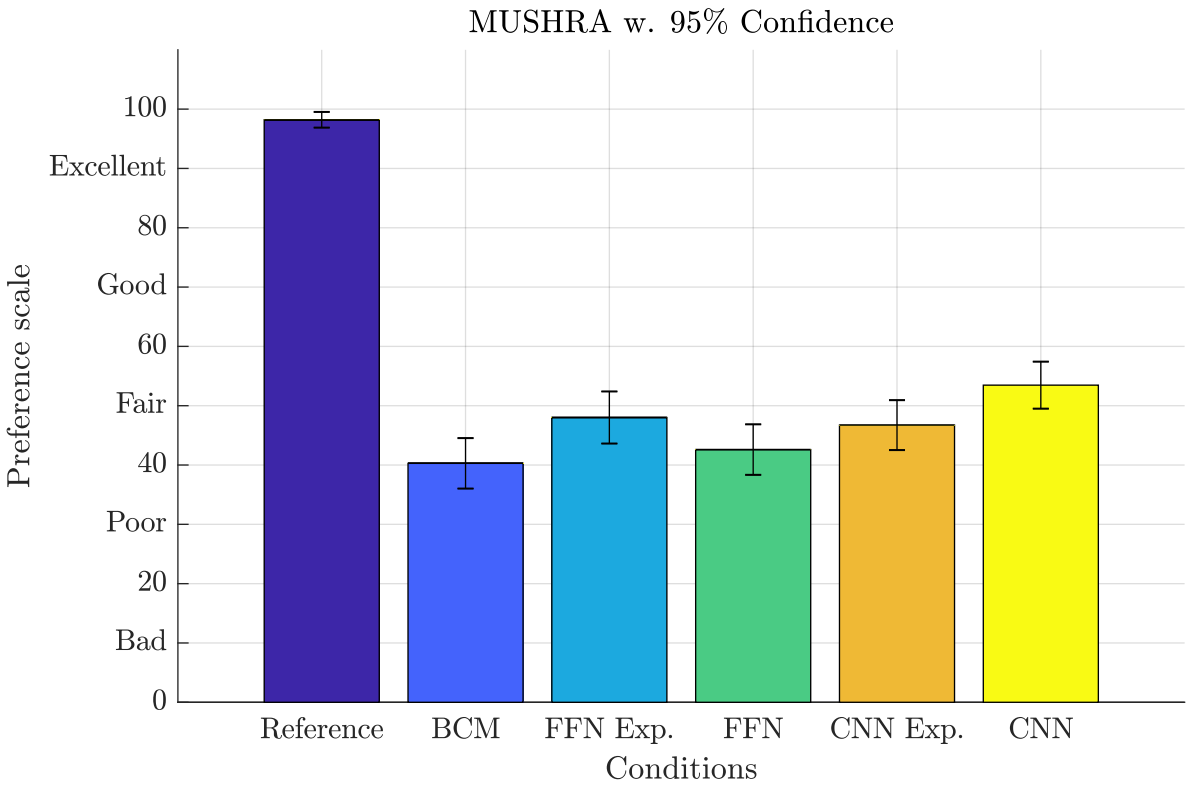


Figure 5.5: Mean results from the *uncontrolled* MUSHRA preference test. The error bars mark the 95% confidence interval. The reference was clearly preferred. The three network types FFN Exp., CNN Exp. and CNN are significantly preferred over the unprocessed BCM signal.

The dynamic range of responses was lower than for the controlled booth measurement and the largest improvement was also lower for the online test.

5.3.2 Experiment 5 - Speech Intelligibility - DANOK

The results of experiment 5 are presented in figure 5.6. The results are presented as the mean across all CV combinations for a given network configuration and a given SNR. The results are illustrated as the accuracy in percentage of consonant recognition for the different SNR levels in dB, and were presented along with the 95% confidence interval.

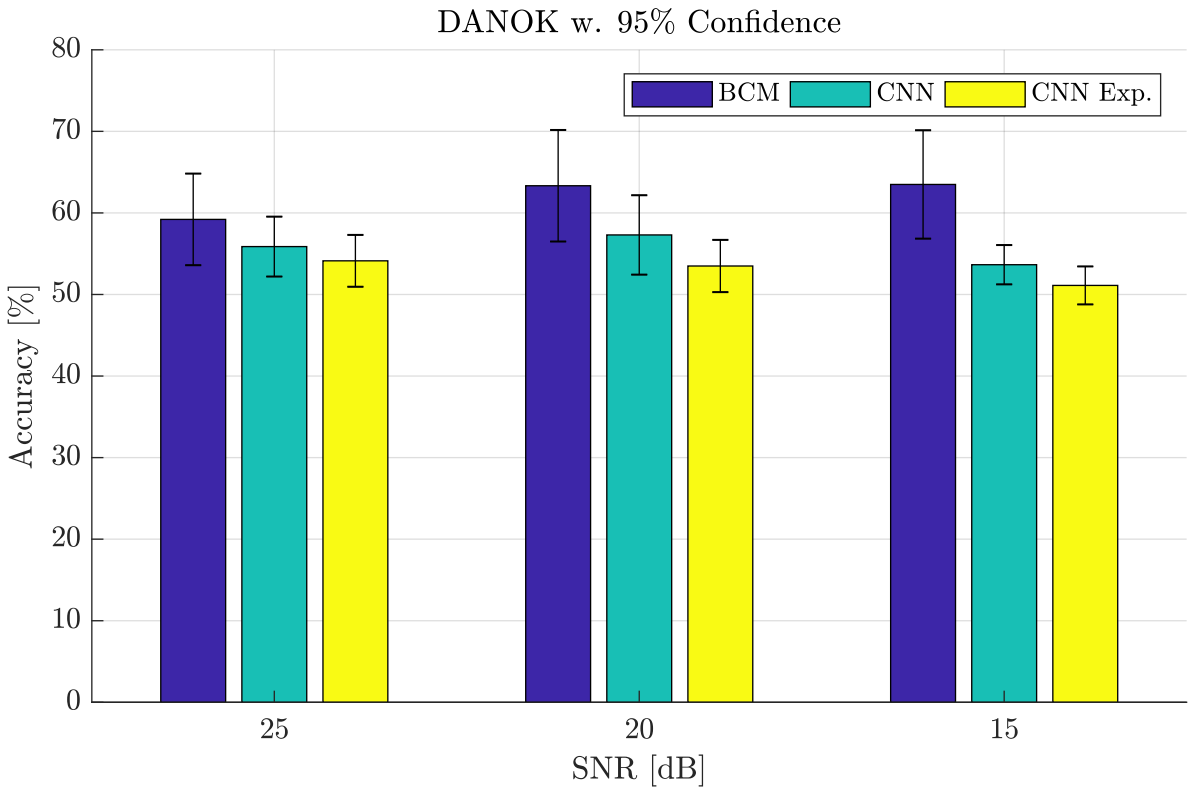


Figure 5.6: The average results from the DANOK experiment with 95% confidence interval across all CV combinations for the different SNR conditions. The results showed very little variation across the different SNR's. A summary of the test of significance can be seen in table 5.3.

Studying the speech intelligibility results, a minor decrease in intelligibility was observed for all network configurations, at all SNR conditions. A statistical analysis was done with a two-sample t-test in order to test the significance of the results. The two-sample t-test analyses the difference of the mean values in the results and evaluates if these differences can be characterized as significant. The result from the statistical test is presented in table 5.3.

Table 5.3: Statistical analysis of the DANOK speech intelligibility results for the BCM, CNN and CNN with expansion in figure 5.6. The analysis showed if there was a statistical significant difference in speech intelligibility between conditions. The confidence value was set to $\alpha = 0.05$.

SNR	**Expansion**	H_0 **rejected**	**p-value**
25	yes	false	0.35
	no	false	0.15
20	yes	false	0.18
	no	true	0.025
15	yes	true	0.018
	no	true	0.005

From the statistical analysis in table 5.3 it was observed that there wasn't any evidence of an increase or decrease in speech intelligibility for the 25 dB SNR conditions. For the 20 dB SNR condition only the CNN with expansion showed a significant decrease. Both the network types shows a significant decrease in performance for the 15 dB SNR condition.

Although a slight indication of an increasing intelligibility for the unprocessed BCM with decreasing SNR was observed, the variation in the observations was too large to denote the increase as statistical significant. There wasn't either any significant difference between the different networks when compared across SNR levels.

In order to obtain a more in-depth view of the the results were analyzed by studying the specific consonant confusions. Figure 5.7 presents a confusion matrix of the actual pronounced consonant versus the subject reported consonant. The results are pooled across SNRs to give a clearer view of the different consonant confusions. The confusion matrix is a good way of investigating any patterns in consonant confusions and that way identify if certain groups of phonemes are harder to understand than others.

The size of the marks indicates the consonant recognition rate. A larger size meaning a higher percentage. Marks in the diagonal indicates correct answers and marks outside of the diagonal indicates incorrect answers.

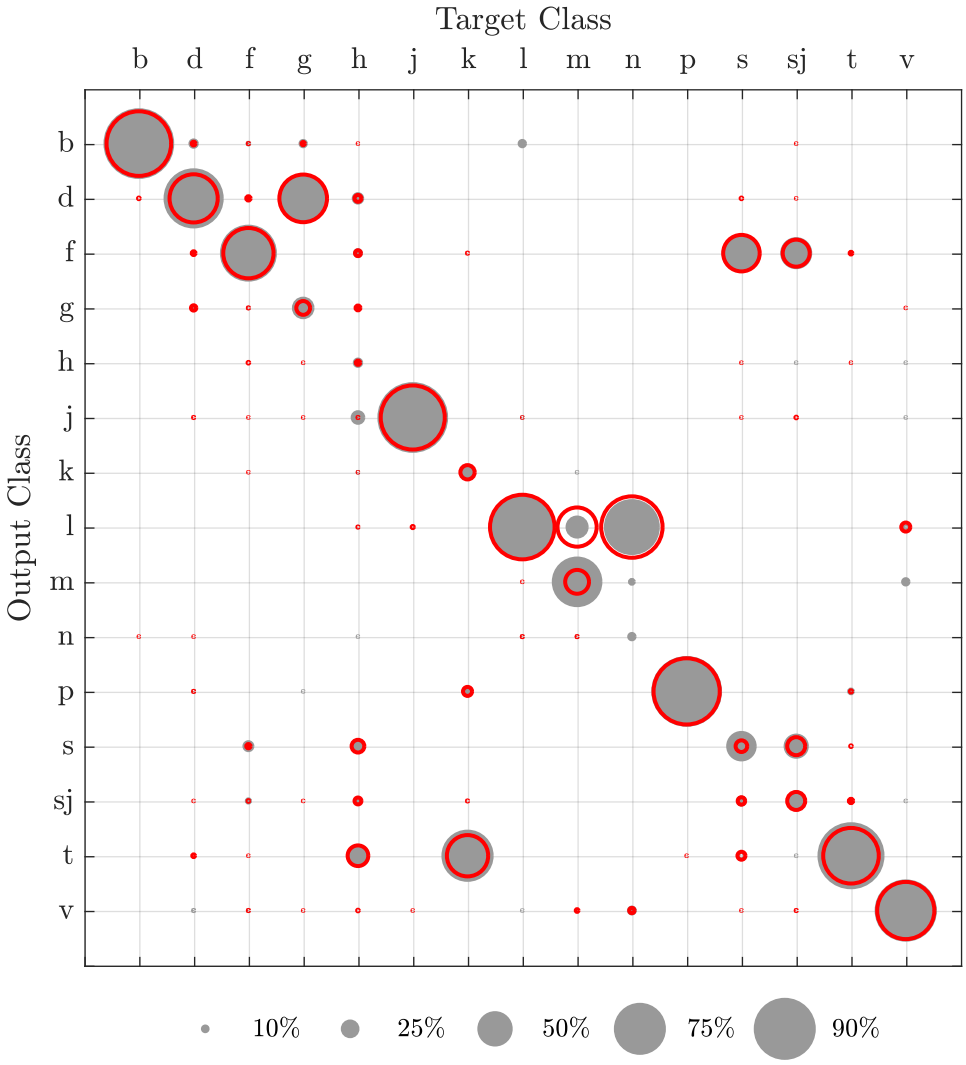


Figure 5.7: Across listener confusion matrix for the unprocessed BCM (gray) and the CNN without expansion (red). The responses were averaged across SNR levels. The size of the circles indicates the proportions of responses for the individual response alternative. Circles in the diagonal indicates correct consonant recognition's.

From figure 5.7 it's observed that some consonants are clearly harder to understand than others. Some consonant almost never get classified correctly while some gets misclassified almost every time. An example is the */g/* and */d/* consonants. */g/* was often classified as a */d/* than as */g/*. Where as */d/* was very seldomly classified as a *g*. The two consonants are in the same phoneme category of *stops without aspiration*, which could indicate that they are similar in the time-frequency domain. An inspection of this hypothesis is shown in figure 5.8.

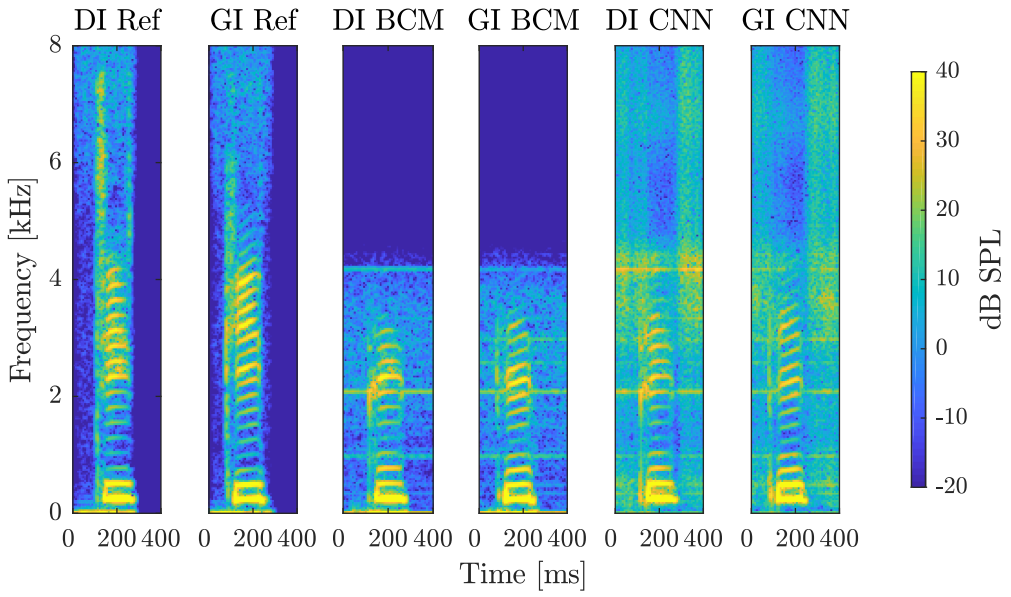


Figure 5.8: Spectrogram representation of the consonants /d/ and /g/ under three given conditions. The two spectrograms to the left presents the clean reference version, the middle spectrograms presents the unprocessed BCM and the spectrograms to the right present the CNN processed.

The spectrograms in figure 5.8 of /d/ and /g/ are presented for three conditions, the clean reference, bone conducted and enhanced versions. Even in the reference condition the consonants looks very similar. Both consonants are initialized by a short burst of broadband energy followed by a harmonic part. The /d/ has a bit more high frequency energy around 6-7.5 kHz and the /g/ has a slight positive slope towards higher frequencies in the harmonic part.
When looking at the BCM condition the high frequency part, is as expected, totally attenuated. The initial burst is almost gone as well. The only part to differentiate them is the harmonic part. When looking at the enhanced spectrograms it's clear to see that some of the high frequency components (3.5-4.5 kHz) of the harmonic part has been enhanced, but the initial burst has not. As this is the part which distinguish the two consonants, one could argue that this is the reason for the confusion.

Another interesting observation from figure 5.7 is that out of the six consonants that almost never gets predicted correct, three of them are fricatives. The sound of the fricative consonants gets produced by air being pressed through the lips or the teeth,

and can be characterized as turbulent noise, and noise don't gets easily picked up by the BCM.

Both observations support the claim that phonemes from the same category gets more easily confused under hard listening conditions, which is in accordance to [Poulsen, 2008].

Chapter 6

Discussion

This chapter contains a discussion of the results from the experiments described in the preceding chapter. The main discoveries are highlighted and evaluated for each experiment. The last section contains additional findings which are considered of substantial value.

6.1 Objective Findings

6.1.1 Experiment 2 - PESQ and STOI

The PESQ results observed in the experiment showed a significant increase of speech quality. An increase achieved by all of the DNN configurations, with the best having a relative improvement of 53%. This was a noticeable improvement, however before concluding anything on the basis of this result, it is crucial to return to the theory behind the PESQ metric. The PESQ metric was developed and standardized by the telecommunication industry, and is used to evaluate the effect which a telephone transmission systems has on speech quality. It can't be stressed enough that the PESQ metric does not evaluate the specific ABE of a given speech signal and multiple studies show that the PESQ metric in many cases is insufficient when comparing different ABE approaches to subjective listening tests [Bauer et al., 2014; Abel et al., 2016].

The PESQ results from the objective analysis showed that the effect of pre-training the network with a synthetic data only increased the performance of the convolutional networks. The performance for the feed forward networks didn't increase. This could

indicate that the convolutional network benefits from pre-training due to it's more complex network structure, with more weights which can needs more information to learn. The FFN on the other hand has fewer weights and may not be able to capture as much information from the pre-training data set, information which can be transferred to the final training phase. An interesting experiment could be to further do a grid-search on the pre-trained networks with the final training set, to see if a more complex FFN network would benefit more from the pre-training.

The BCM training set contains a noticeable amount of constant wide-band background noise originating from the transducer design of the BCM. Given the inherent limitations of the BCM training set it's difficult to pre-define what the network should learn. This means that the network automatically also learns to remove the background noise, because that factor also contributes to reducing the MSE cost function. This is a factor that needs to be accounted for when evaluating the explicit bandwidth extension ability of the algorithm. The large improvements in PESQ may partial reflect precisely this noise reduction. Even though the noise reduction was a positive side effect of the network, it makes it harder to evaluate and analyze the specific bandwidth extension of the speech signal.
This issue indicates that PESQ may not be the optimal choice, and does definitely not present the complete picture of the performance. However the PESQ metric is probably the best standardized objective metric to predict speech quality, which is also the reason why it's so widely used. To account for the limitations of PESQ, both STOI and MCC were used to consolidate and produce a more comprehensive analysis.
The STOI results didn't present any noticeable improvement. Which will be tried to explained with the results obtained from the MCC metric, which will be discussed in the section below.

The test set used to assess the performance of the final neural networks consisted of the same 12 sentences which were used for the MUSHRA experiment. The test set should not only assess the performance, but also the generalization ability. In order to analyze the generalization ability a more varied test set could be of interest. Even though the test set contained real recordings from different speakers, they have all been

recorded in a controlled lab environment. A test set containing recordings from real world situations would provide a more realistic test.

6.1.2 Experiment 3 - Bandwidth Extension - MCC

The limitations imposed by the PESQ metric regarding the explicit evaluation of bandwidth extension motivated the second experiment. The MCC metric is better designed for the purpose of actually examining bandwidth extension. It is a simple but very effective metric which evaluates the correlation in each frequency bin. This means that it's possible to observe the correlation variation specifically for the reconstructed frequencies. The MCC results shown in figure 5.3 presents a comparison between the MCCs obtained for the unprocessed BCM baseline and for the different model configurations across frequencies ranging from 50 to 8 kHz. The results showed a correlation improvement of around 0.5 for the high frequencies above 5 kHz. As the correlation was calculated between the clean reference speech signal and the processed speech signal one could argue that a large part of the high frequency components has been reconstructed. It's paramount to keep the underlying theory behind the MCC calculation in mind. The MCC calculation is a simple energy correlation measure, meaning it does not care where the energy comes from. The improved correlation can therefore be due to introduced noise components, which would result in wrong interpretations of the performance.

The effect of bandwidth degradation on speech has been studied in several publications. In the book [Erik Larsen, 2005] the impact of limited bandwidth on speech quality has been compared to the impact on speech intelligibility. The study showed that high frequency components (> 3 kHz) of speech only has little effect on speech intelligibility. Removing the high frequency components have a much more critical impact on the speech quality. This effect was also reflected in our objective experiments, where it was found that an extension of the high frequencies significantly improved PESQ scores, whereas the speech intelligibility *(STOI)* already scored high for the unprocessed speech and did not improve significantly.
This indicates the significant PESQ improvement can be linked to the extension of the high frequencies. Indicating that the method actually extends the bandwidth and reconstruct the high frequency content, which causes the quality of the perceived speech

to increase.

6.2 Subjective Findings

6.2.1 Experiment 4 - Speech Quality

Two MUSHRA experiments were conducted. An experiment conducted at DTU in controlled settings with selected test subjects, and an uncontrolled online experiment. Both MUSHRA experiments were conducted to evaluate the difference in perceptual sound quality between the unprocessed BCM speech signal and the processed speech signals. The experiments evaluated four different network configurations.
The controlled experiment showed an improvement of 20 percentage points for the speech signal processed by both the FFN and CNN network configuration with expansion. The difference between the FFN and CNN network configurations with expansion was not significant. This is interesting since the two network types vary considerably in complexity and since the PESQ results from the objective experiment showed a significant difference in favor of the CNN configuration. This difference indicates that the PESQ metric does not correlate perfectly with a subjective speech quality evaluation.
The memory usage of the FFN and CNN model configurations differs with a factor 20, meaning that the CNN model configuration is 20 times more memory demanding. As efficiency, power consumption and such aspects of implementation considerations isn't part of the scope for this project, it hasn't been investigated further. But it is definitely worth noticing and it illustrates that network complexity does not always correlate with the final network performance.

In the ITU recommendations [BS.1534-2, 2014] it is stated that the test subjects participating should be experienced and critical listeners. In this controlled MUSHRA experiment the subjects nationality and their hearing ability was taken into consideration. The subject were furthermore partly characterized as experienced listeners, since everybody previously had participated in listening experiments.
However the subjects ability to judge a sound signal critically was not part of the selection criteria and because of this all the results were screened for outliers. The acceptance requirements of the screening was that the reference should, on a mean basis,

not be rated lower than 95. All subjects met the criteria and was included in the results.

The online MUSHRA experiments showed a slightly different result than the controlled experiments. The resultin ratings appear more compressed, meaning the dynamic range between the quality ratings are lower. This could be due to inadequate listening equipment or that the experiments were conducted in noisy acoustic environments.
It's hard to conclude why exactly the results are more compressed and it would require a more comprehensive experimental setup to determine what factors affects the outcome of the experiment. The comprehensive experiment setup would have to contain a more detailed description of the equipment used, and the surrounding acoustic environment which the experiment is conducted in. Such detailed information could help investigate how different parameters affect the experiment outcome.
Online audio experiments have in a long time been impossible due to inadequate web standards for audio processing. But with recent advancements in the Web Audio API such online experiments have become available. The online MUSHRA experiment is very interesting because of it's efficiency and it's abilities to reach a wide ranging pool of listeners. However given these results it could indicate that it's necessary to find a way for subject to carry out the experiment while at the same time being compliant to the ITU recommendations [BS.1534-2, 2014].

6.2.2 Experiment 5 - Speech Intelligibility

The results from experiment 5 presented the speech intelligibility scores for all CV combinations given different CNN model configurations at different SNR levels. The results indicates that there is a small decrease in speech intelligibility for the CNN network configurations compared to the unprocessed BCM baseline. It should be stated that not all the results are statistical significant.
The initial goal of extending the bandwidth of the degraded speech signal was to improve the speech quality while maintaining or improving the speech intelligibility in noisy conditions. However the results from the DANOK experiment indicates that this goal has not been accomplished.
There are multiple reasons why this is. First of all predicting and improving the band-

width for the different CV combinations under the chosen SNR conditions is an incredible difficult task for the network. The combination of unseen noise and low SNR levels challenges the network in terms of it's robustness and it's ability to generalize to unseen acoustic environments. Taking this into consideration along with the results which showed that the speech intelligibility doesn't drop significantly at 25 dB SNR, one could argue that the robustness of the network is good. As the SNR drops to 20 dB and 15 dB the difference between the unprocessed BCM baseline and the processed CV tokens increases, which makes sense since the prediction task gets even harder.

The DANOK results also indicated a small increase in speech intelligibility for the unprocessed BCM baseline with a decreasing SNR. Again the result didn't present a statistical significant increase. However it is an interesting result and it could indicate that the effect of pre-training affect the results. A factor which is important to consider when evaluating the implications of the DANOK experiment. For future experiments it could be considered to expand the preliminary training phase of the experiment. This would allow the subject to gain experience with the experiment method which hopefully would lower the potential effect of training. This training issue could also be a substantial cause for the increasing difference between the processed and unprocessed results going from high to low SNR level.
It is suspected that the training effect isn't seen to the same extent for the processed CV tokens since the noise is more unpredictable and non-stationary.

Finally it could be argued that the DANOK experiment does not represent a realistic listing setting, and that complete sentences would be more realistic than different CV combinations. The DANOK experiment does provide us with a informative macroscopic evaluation and it provides a way to evaluate and study specific consonant confusions.

6.3 Summary of Main Findings

6.3.1 Data and Feature Selection

One of the fundamental requirements for successfully utilizing the deep learning methodology is to have a data set that reflects the problem at hand. One of the major challenges

in deep learning and especially in this project was the data collection process. As getting the right data set for solving the problem is paramount for obtaining good performance, it has been a fundamental part of this project. The process of collecting and labeling large amount of real world data was not an option in the time scope of this project. This inspired the investigation of possible solutions to this issue. The solution became to analyze a small subset of real world data and from that data create a synthetic data set. A data set which approximated the characteristics of the real world.
The small subset of real world data was afterwards used to train and fine-tune the model. A process which increased the generalization ability of the model.
Another way to include the synthetic data could be to complement it with the real world data, and thus creating a combined larger data set, and only training the network once. The reason this wasn't tried was due to memory limitations of the computer used for training the network. But it could be a way to force the network to generalize better to unseen data.

The correct feature selection is a vital part of designing a deep neural network and is crucial for achieving a high performance. Even though a diverse data set is crucial, the process of interpreting and preparing the feature representation is just as important. Selecting and extracting the right features helps build a solid foundation for the actual learning process. In this project the main focus has been on using frequency magnitude features. Given the right resolution this provide a neural network with a lot of useful information.
A frequency representation consist of two things, the magnitude part and the phase part. This means that by only feeding the magnitude features compressed on a logarithmic scale to the neural network does not provide the full picture. Studies show that to obtain a natural sounding speech signal and thereby improve the speech quality it is crucial to both enhance the magnitude and the phase response of the degraded signal [Williamson and Wang, 2017; Moon et al., 2010]. So an improved result could possibly be obtained by jointly enhancing both the magnitude and phase parts.
Enhancing the phase is tricky thing and after a couple of preliminary experiments we chose not pursue this issue any further, primarily because it's clear that a successful result can be obtained by only enhancing the magnitude features. The phase issue is

worth keeping in mind for future research, since it definitely have an impact on the speech quality.

6.3.2 Variance Scaling - Expansion

Variance scaling (Expansion) was utilized as a post-processing step and is a method of expanding the variance of the magnitude distribution of the output. The purpose of this step was to further improve the speech quality and the generalization ability of the networks.

The results from the objective speech quality experiment showed a significant positive effect from applying this post-processing step. To find the reason for the increased PESQ value inspected the calculations for which the PESQ metric is determined on. The PESQ metric is calculated on the basis of the absolute difference between the clean reference and degraded speech signals. The final PESQ score is a non-linear computation over time and frequency and is calculated by two averages, one symmetric and one asymmetric. The asymmetric weighted factor is only measuring additive distortions, meaning that if the degraded signal has energy present in the frequency bins, where the reference doesn't, it will have a larger negative impact on the final PESQ result, than if the degraded signal is missing energy in the frequency bins where the reference have energy.

The non-linear process of variance scaling attenuates a portion of the unwanted noise components which are introduced by the neural network, this increases the similarity between the processed and clean reference signals. This increase in similarity improves the final PESQ result. However the method also attenuates some of the speech components. This were clearly visualized in the MCC evaluation which presented a lower correlation for signals with expansion.

From an informal listening test it was concluded that the listening experience was more pleasant, and that less artificial noise artifacts are audible. Because the expansion scheme attenuated a portion of the unpleasant noise artifacts introduced by the network, the listening experience was more smooth. However it was also noticed that some of the wanted speech components were attenuated. This were also illustrated in the

subjective evaluation which presented an improvement of speech quality, most significantly for the FFN.

To limit the attenuation of reconstructed speech components, it would be interesting to look into other ways of applying the variance scaling. Based on the findings in [Wiinberg et al., 2018] it could be interesting to look into a T-F-frame wise SNR-dependant enhancement scheme. By only expanding the frequency frames that met a certain SNR-criteria the scheme could distinguish between frames containing speech and frames containing noise and that way limit the attenuation of speech components. Another interesting thing could be to investigate which components of the signal are attenuated and which components are amplified. The variance scaling blindly amplified all the T-F-bins that, in the normalized log-domain, had positive values, and attenuated all the bins that had negative values. If it were possible to identify the T-F-bins below zero which contained speech information, and shift them to the positive region, the problem of suppressing reconstructed speech components could be minimized.

6.3.3 Quality vs. Intelligibility

An interesting observation which was drawn from the results of the subjective experiments were that speech quality and intelligibility didn't necessarily correlate. An improvement in quality was found for three of the network configurations, while no significant difference in intelligibility was found for high SNR levels. The same observation could be made for the objective measures from experiment 2. Here the PESQ results showed significant improvements, while the STOI didn't.
The task of increasing speech intelligibility is hard and have been the focus of countless studies. So when both the objective and subjective results show an increase in quality while maintaining the intelligibility the algorithm can still be regarded as successful. It could be interesting to evaluate the speech intelligibility further by lowing the SNR level in the subjective DANOK experiment. By lowering the SNR, the consonant recognizing rate should theoretically drop. This would hopefully produce a more righteous evaluation.
It should be noted that results actually already showed a slight drop in speech intelligibility for lower SNR levels, which were properly due to the networks inadequate ability

to generalize to unseen acoustic environments
A higher speech quality often lead to a lower listening effort and when considering the stress-full and critical environment where the ABE algorithm is meant to be used, this could prove very valuable.

6.3.4 Objective vs. Subjective

When comparing the results obtained from the objective and subjective speech quality experiments, two clear discrepancies were observed. First the CNN configuration with expansion scored almost 0.4 PESQ points higher than the FFN configuration with expansion, but when looking at the subjective results the difference was negligible. Second the FFN configuration without expansion scored higher in PESQ but lower on subjective speech quality measure than the unprocessed BCM.
These discrepancies reveals one of the limitations with PESQ. The method does not distinguish between the different subcategories of speech quality. Meaning PESQ predicts an overall value for speech quality and does not evaluate the specific aspects of speech quality like speech bandwidth, noise, artificial sounding noise components.
The objective measure of speech intelligibility appears to correlate well with the subjective measurement. Here both metrics show no significant improvement or decline in intelligibility. But it should be noted that a direct comparison can't happen since the two experiments are conducted using two different speech data sets.

Chapter 7

Conclusion

The main objective of this thesis was to implement and evaluate an artificial bandwidth extension framework to increase speech quality while not degrading speech intelligibility for bone conducted, band-limited speech. The main focus was to investigate the capabilities of using deep neural networks to solve the challenges imposed by bone conduction of speech.

In **chapter 3** a framework for training and utilizing neural networks were developed. Two neural network structures were implemented; a feed-forward and a convolutional. To increase the amount of training data and to enhance the generalization abilities of the networks, a synthetic data set was developed. The synthetic data set was based on the TIMIT corpus and approximated the characteristics of bone conducted speech. The data set was created by applying a set of estimated transfer functions based on a small set of matched BCM and ACM recordings to the TIMIT corpus. To further increase the similarity to real world data, recordings of bone conducted noise were added to the corpus at varying SNR levels.

In **chapter 4** five experiments were constructed in order to analyze the effects of the bandwidth extension performed by the neural networks. The experiments consisted of both objective and subjective experiments, as this were a crucial factor for obtaining a comprehensive evaluation.

Chapter 5 showed the results of the objective and subjective experiments. Both showed that the speech quality was significantly improved while maintaining the speech intelligibility for high SNR levels. The results showed a slight drop in speech intelligibility for the lowest SNR levels, which were due to the networks inadequate ability to generalize to unseen acoustic environments. A technique called variance scaling was applied to expand the output magnitude distribution of the network. The technique presented a significant positive impact on both the objective and subjective speech quality experiments.

On the basis of this project, the use of deep neural networks proves as an attractive solution for extending band-limited speech, and reconstruction lost speech components. The capabilities reach beyond bone conducted speech and could be implemented to enhance other kinds of degraded signals.

7.1 Future Works

The algorithm described and evaluated in this project is not ready to be implemented and utilized in a real-time audio application. The size and the complexity of the neural networks are too large and complex to meet the requirements of a real-time system and further research should be carried out to mature the system and meet the requirements. A future step for INVISIO Communication would be to further investigate a feasibility of a implementation of a bandwidth extension algorithm on embedded systems.

The experimental results presented a robustness issue at low SNR levels. A issue which relates to the networks ability to adapt to unseen noise environments. A further investigation into optimizing the robustness of the algorithm towards unseen noise environment could be interesting. Such an investigation should be connected with a further analysis of the speech intelligibility linked with bone conducted speech.

Bibliography

J. Abel and T. Fingscheidt. A dnn regression approach to speech enhancement by artificial bandwidth extension. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 219–223, 2017.

J. Abel, M. Kaniewska, C. Guillaumé, W. Tirry, and T. Fingscheidt. An instrumental quality measure for artificially bandwidth-extended speech signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP, 12 2016.

F. L. A. T. A. M. J. B. Allen. A psychoacoustic method for studying the necessary and sufficient perceptual cues of american english fricative consonants in noise. *The Journal of the Acoustical Society of America*, 132, 2012.

P. Bauer, C. Guillaumé, W. Tirry, and T. Fingscheidt. On speech quality assessment of artificial bandwidth extension. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6082–6086, 2014. ISSN 1520-6149.

Y. Bengio. Practical recommendations for gradient-based training of deep architectures. 2012.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

I.-R. BS.1534-2. Method for the subjective assessment of intermediate quality level of audio systems. *Electronic Publication*, 2014.

B. L. J. T. Z. W. Y. L. D. Bukhari. A novel method of artificial bandwidth extension using deep architecture. *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, pages 2598–2602, 2015.

R. H. C. H. Taal, R. C. Hendriks and J. Jensen. An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech. *Acoustical Society of America*, 130(5):3013–3027, 2011.

M. A. T. T. C. Yagli and E. Erzin. Artificial bandwidth extension of spectral envelope along a viterbi path. *Speech Communication*, 55(1):111–118, 2013.

H. Carl and U. Heute. Bandwidth enhancement of narrow-band speech signals. *Signal Processing Vii, Theories and Applications*, 2:1178–81, 1994.

R. H. Cees H. Taal, Richard C. Hendriks and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. *ICASSP*, 2010.

P. J. H. T. U. Christiansen. Objective evaluation of consonant-vowel pairs produced by native speakers of danish. *ResearchGate*, 2011.

J. Z. T. Dau. Predicting consonant recognition and confusions in normal-hearing listeners. *Journal of the Acoustical Society of America*, 141:1051—-1064, 2017.

R. M. A. Erik Larsen. *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. John Wiley & Sons, Ltd, 2005.

H. S. S. H. G. R. T. Fingscheidi. Survey of speech enhancement supported by a bone conduction microphone. *Proceedings of 10th Itg Symposium on Speech Communication*, 2012.

J. A. M. S. T. Fingscheidt. Artificial bandwidth extension using deep neural networks for wideband spectral envelope estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):71–83, 2018.

J. Flanagan. *Speech Analysis, Synthesis and Perception*. Springer-Verlag, 2nd edition, 1978.

e. G. E. Hinton. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

B. Hagerman. Measurement of speech reception threshold. *Scandinavian Audiology*, 11:191–3, 1982.

D. M. I. Katsir and I. Cohen. Evaluation of a speech bandwidth extension algorithm based on vocal tract shape estimation. *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pages 1–4, 2012.

ITU-T. Pesq: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. recommendation p.862 (02/01). `https://www.itu.int/rec/T-REC-P.862-200102-I/en`, 2011a. Accessed: 2018-05-11.

ITU-T. Mapping function for transforming p.862 raw result scores to mos-lqo recommendation p.862.1 (11/03). `https://www.itu.int/rec/T-REC-P.862.1-200311-I/en`, 2011b. Accessed: 2018-05-11.

P. Jax and P. Vary. Wideband extension of telephone speech using a hidden markov model. *2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium*, pages 133–135, 2000.

K. Li and C. H. Lee. A deep neural network approach to speech bandwidth expansion. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4395–4399, 2015.

R. P. Lippmann. An introduction to computing with neural nets. *Ieee Assp Magazine*, 4:4–22, 1987.

P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013. ISBN 1466504218, 9781466504219.

S. Möller, E. Kelaidi, F. Koster, N. Cote, P. Bauer, T. Fingscheidt, T. Schlien, H. Pulakka, and P. Alku. Speech quality prediction for artificial bandwidth extension algorithms. pages 3439–3443, 01 2013.

S. H. Moon, B. Kim, and I. S. Lee. Importance of phase information in speech enhancement. *International Conference on Complex, Intelligent and Software Intensive Systems*, pages 770–773, 2010.

M. Nielsen. Neural networks and deep learning, 2017. URL `http://neuralnetworksanddeeplearning.com/`.

A. H. Nour-Eldin and P. Kabal. Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech. 2011.

T. Poulsen. Acoustic communication, hearing and speech. *Lecture notes, ver. 2.1.1*, 2008.

H. Pulakka, V. Myllylä, A. Rämö, and P. Alku. Speech quality evaluation of artificial bandwidth extension: Comparing subjective judgments and instrumental predictions. 09 2015.

L. Rabiner and R. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, 1978.

J. Sadasivan, S. Mukherjee, and C. S. Seelamantula. Joint dictionary training for bandwidth extension of speech signals. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5925–5929, 2016.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x.

H. S. Shin, H. G. Kang, and T. Fingscheidt. Survey of speech enhancement supported by a bone conduction microphone. *Speech Communication; 10. ITG Symposium*, 2012.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 06 2014.

L. Torrey and J. Shavlik. Transfer learning, appears in the handbook of research on machine learning applications. *Pre-print; University of Wisconsin, Madison WI, USA*, 2009.

P. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, Jun 1967.

A. Wiinberg, J. Zaar, and T. Dau. Effects of expanding envelope fluctuations on consonant perception in hearing-impaired listeners. *Trends in Hearing (online)*, 22, 2018.

D. S. Williamson and D. Wang. Time-frequency masking in the complex domain for speech dereverberation and denoising. *Ieee-acm Transactions on Audio Speech and Language Processing*, 25(7):1492–1501, 2017.

A. B. X. Glorot and Y. Bengio. Deep sparse rectifier neural networks. *Journal of Machine Learning Research - Workshop and Conference Proceedings*, 15:315–323, 2011.

Y. Xu, J. Du, L. R. Dai, and C. H. Lee. Global variance equalization for improving deep neural network based speech enhancement. pages 71–75, July 2014. doi: 10.1109/ChinaSIP.2014.6889204.

Z.-H. L. Y. Gu and L.-R. Dai. Speech bandwidth extension using bottleneck features and deep recurrent neural networks. *Interspeech*, 2016.

Appendix A

INVISIO - Background

INVISIO is a global market leader within advanced communication and hearing protection systems. The company develops and sells advanced systems that enable professionals in noisy and mission critical environments to communicate and work effectively, while protecting their hearing.
INVISIO's systems give operational advantages and increased security for military and security personnel.

Figure A.1 illustrate the INVISIO X5 headset, which were used to record the bone conducted speech corpus.

Figure A.1: The INVISIO X5 headset, used for recording the bone conducted speech corpus. Picture obtained from INVISIO's website