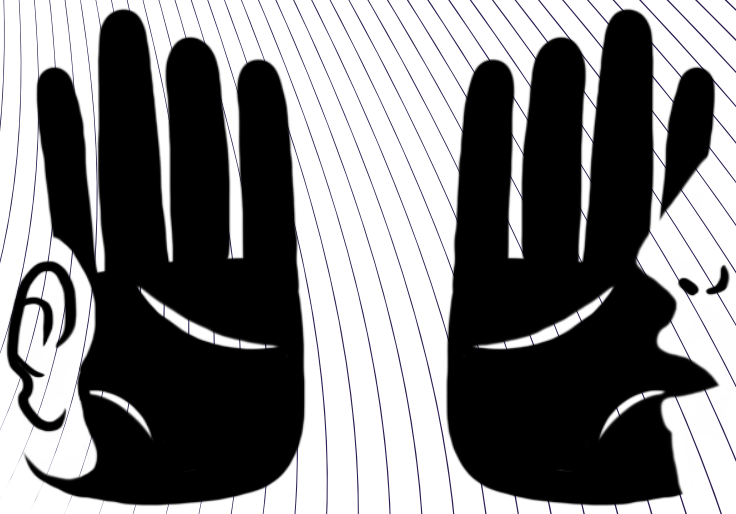


Predictability-Based Objective Evaluation of Sound

Thor Pilgaard Knudsen

Mathematical Engineering, June 2021

Master's Thesis

AALBORG UNIVERSITY
STUDENT REPORT

Mathematical Engineering
Aalborg University
www.en.aau.dk

Title:
Predictability-Based Objective Evaluation of Sound

Theme:
Acoustic Signal Processing

Project period:
September 2020 - June 2021

Participant:
Thor Pilgaard Knudsen

Supervisors:
Jesper Jensen
Jesper Møller

Number of pages:
85

Date of Completion:
4th of June, 2021

Synopsis:

In this work, we explore the potentials of predictability as an objective measurable quantity to identify elements of speech which are most important for intelligibility. Specifically, we propose a measure which is a non-intrusive perceptually relevant novel estimate of the information theoretical quantity *mutual information*. This measure is computed for discrete time frames of speech signals and utilizes deep convolutional neural networks. In a listening test, the proposed measure was compared to two existing methods, namely sound intensity and cochlea-scaled spectral entropy. We found that - in certain conditions - the proposed measure better identified time frames important for speech intelligibility compared to sound intensity and cochlea-scaled spectral entropy. However, in other conditions, the proposed measure failed to identify such frames. The results suggest potentials of predictability as an objective measure, however, alterations should be made to the proposed measure in order to more sensible evaluate the measure.

Danish Summary

På verdensplan rammes mange mennesker af nedsat hørelse. Evnen til at høre er essentiel i de fleste hverdagsaktiviteter, og nedsat hørelse påvirker derfor livskvaliteten. Heldigvis har mennesker med nedsat hørelse mulighed for at tilegne sig et høreapparat. State-of-the-art høreapparater er udstyret med taleforbedringsalgoritmer, som har til formål at afhjælpe høretab ved blandt andet at forbedre taleforståeligheden og dermed forbedre livskvaliteten hos høreapparatbrugeren. For at kunne optimere sådanne taleforbedringsalgoritmer er det nødvendigt at have en viden om, hvilke dele af talesignaler, der er vigtigst for taleforståeligheden.

I denne afhandling undersøges potentialet i at kvantificere vigtigheden af elementer af talesignaler ud fra, hvor forudsigelige nuværende tidsintervaller af talesignaler er givet de forrige tidsintervaller. Ud fra en spektro-temporal opsplitning af talesignaler estimeres specifikt den informationsteoretiske størrelse 'mutual information' mellem nuværende tidsintervaller og de forrige tidsintervaller. Det bærende element i den foreslåede metode er brugen af dybe neurale netværk til at ekstrapolere spektro-temporale værdier af talesignaler.

I afhandlingen præsenteres først de nødvendige akustiske procedurer samt informationsteoretiske størrelser anvendt i projektet. Dernæst undersøges eksisterende metoder til kvantificering af vigtigheden af elementer af talesignaler for taleforståeligheden og efterfølgende, præsenteres den foreslåede metode. Den resterende del af afhandlingen beskæftiger sig med dokumentation af opsætning, resultater samt diskussion af en lyttetest, som er udført for at vurdere den foreslåede metode.

På baggrund af resultaterne fundet i denne afhandling, kan det konkluderes, at den foreslåede metode - under nogle forhold - bedre identificerer dele af talesignaler, som er vigtigst for taleforståeligheden, sammenlignet med de undersøgte eksisterende metoder. Dog er den foreslåede metode - under andre forhold - ude af stand til at identificere disse vigtige dele af tale. På baggrund af dette samt det faktum, at den foreslåede metode har nogle ulemper, foreslås videre udarbejdelse af metoden før mere præcise konklusioner kan drages.

Preface

This Master's Thesis (60 ECTS) is written by Thor Pilgaard Knudsen of the Master program Mathematical Engineering at Aalborg University, Department of Mathematical Sciences in the period September 2020 - June 2021. The thesis is compiled in collaboration with the Department of Mathematical Sciences at Aalborg University, the Department of Electronic Systems at Aalborg University and Oticon A/S.

For notation, vectors and matrices are denoted by bold letters while scalars are denoted by non-bold letters. Indexing of vectors and matrices use brackets, e.g., $A[i, j]$ denotes the scalar corresponding to the ith row and jth column of the matrix $\boldsymbol{A}$. Furthermore, in the context of indexing, we use a colon to refer to either all values, e.g., $\boldsymbol{A}[:, j]$ denotes the jth column of $\boldsymbol{A}$, or a slice, e.g., $\boldsymbol{A}[:, i : i']$ denotes the columns from i to (and inclusive) i' of $\boldsymbol{A}$.

Citations follow the IEEE citation and referencing guide in that they are numbered after order of appearance and optionally specify the location in the source, e.g., [1, p. 2] refers to the second page of the first appearing citation. A bibliography presenting the used sources in order of appearance is presented in the end of the thesis.

Referencing of theorems, definitions, examples, etc. share a common counter, e.g., Theorem 1.2 follows Definition 1.1. Figures, tables and algorithms have individual counters. Furthermore, a black filled square on the right hand side of a page, indicates the conclusion of a proof, while a black filled triangle concludes either a definition, theorem, or example.

All figures presented in this thesis are created by the author.

The author would like to thank the supervisors Jesper Jensen (Oticon A/S, Department of Electronic Systems) and Jesper Møller (Department of Mathematical Sciences) as well as former supervisor Adel Zahedi (former postdoctoral researcher at Oticon A/S) for their guidance during the development of this thesis.

Aalborg University, June 4, 2021

Thor Pilgaard Knudsen
tpkn15@student.aau.dk

Contents

Abbreviations, Nomenclature, Operators and Units

Abbreviations

AED	Average extrapolation distortion.
ANOVA	Analysis of variance.
CDF	Cumulative distribution function.
CNN	Convolutional neural network.
CSE	Cochlea-scaled spectral entropy.
DFT	Discrete Fourier transform.
DNN	Deep neural network.
ERB	Equivalent rectangular bandwidth.
FFT	Fast Fourier transform.
GUI	Graphical user interface.
LMMSE	Linear minimum mean square error.
LSD	Log-spectral distortion.
MMSE	Minimum mean square error.
MSE	Mean square error.
RAU	Rationalized arcsine unit.
ReLU	Rectified linear unit.
RMS	Root mean square.
ROEX	Rounded exponential.
SIS	Speech intelligibility score.
SSN	Speech shaped nose.
STFT	Short-time Fourier transform.
TIMIT	Texas Instruments/Massachusetts Institute of Technology.
VAD	Voice activity detector.

Nomenclature

$\mathbb{R}^N$	N-dimensional real coordinate space.
$\mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	N-dimensional Gaussian distribution.
ρ_S	Spearman's correlation coefficient.
j	Imaginary unit.

Operators

$(\cdot)^T$	Vector/matrix transposition.
$(\cdot)^{-1}$	Matrix inversion.
$\mathrm{Cov}[\cdot,\cdot]$	Covariance.
$\det(\cdot)$	Determinant.
$\mathbb{E}[\cdot]$	Expectation.
$\ln(\cdot)$	Natural logarithm (base e.
$\log_a(\cdot)$	Base a logarithm.
$\overline{(\cdot)}$	Vector mean.
$\mathrm{Card}(\cdot)$	Set cardinality.
$\max\{\cdot\}$	Maximum.
$\mathrm{tr}(\cdot)$	Trace.
$\lvert\cdot\rvert$	Element-wise absolute value.
$\lVert\cdot\rVert_2$	Euclidean norm.

Units

bit	Bit.
dB	Decibel.
Hz	Hertz.
nat	Nat.
Pa	Pascal.
s	Second.

1 Introduction

In 2015, the world's second leading impairment by the number of people affected was hearing loss [1, p. 1551]. The ability to hear is essential to many activities making most ordinary tasks difficult for the hearing impaired individual, ultimately influencing the quality of life [2, Ch. 4], [3], [4]. In order to combat hearing loss, hearing impaired individuals have the option to adopt hearing assistive devices such as hearing aids. Most state-of-the-art hearing assistive devices are equipped with speech processing algorithms which serve the purpose of amplifying sound sources of interest, while suppressing noise sources, hence increasing satisfaction for the hearing impaired end-user in an everyday setting [5, pp. 93-96].

The focus of speech processing algorithms is usually improving one of the two attributes: quality or intelligibility. Speech quality is a measure of how good or bad the listening experience is for the listener and is a composition of many dimensions, e.g., naturalness, clarity, pleasantness, brightness and many more. Speech intelligibility entails only to which degree the listener is able to understand spoken words and the meaning thereof [6, pp. 623-624]. The terms speech quality and seech intelligibility are not equivalent. Consider, for example telecommunications, where, if a large amount of data packets are lost in the transmission, then, at the receiver end, speech will be perceived as interrupted and possibly less intelligible as some words might be gone. However, the perceived speech quality of the remaining words might be unchanged [7, Sec. 10.1]. Hence, we are in need of different methods to assess the two attributes. In this thesis, the sole focus is speech intelligibility.

A necessary phase of development - as well as evaluation - of speech processing algorithms is listening tests. These tests are realized by having a group of typical end-users listen to a set of speech signals processed by the given algorithm. While listening tests are the only means of obtaining ground truth data, this approach is both expensive and time consuming [6, p. 623], [8, p. 1]. This motivates the need for objective speech intelligibility measures. Such an objective measure can - if sufficiently accurate - replace at least some listening tests in the development and evaluation phases of speech processing algorithms [9], [10].

In order to construct algorithms which can accurately predict speech intelligibility, knowledge about which elements of speech that are of greatest importance for the intelligibility is needed.

With the acquisition of such insight, we would be able to tailor speech processing algorithms to prioritize these elements which would hopefully increase satisfaction for the end-user.

Several decades of research in the area of identifying the elements most important for speech intelligibility [7, Sec. 10.0] have led to well-established techniques for objective speech intelligibility prediction, e.g., the (Extended) Short-Time Objective Intelligibility ((E)STOI) [10], [8], the mid-level coherence speech intelligibility index (CSII-MID) [11] and the normalized covariance speech transmission index (NCM) [12]. Typically, these techniques involve decomposing a noisy (or processed) and a clean time-domain speech signal into a spectro-temporal domain and afterwards grouping and weighting the frequency subbands to crudely reflect the inner workings of the human auditory system. Then a fidelity measure is computed between each spectro-temporal tile of the noisy signal and the corresponding tile in the clean signal. Finally, combining the fidelity scores for all tiles results in the final score [7, p. 516], [13, Sec. 2.1]. However, ideally, the objective measure should be able to assess the intelligibility without access to the clean signal [7, p. 479]. Such measure is referred to as a non-intrusive measure, as opposed to intrusive measures which have access to the clean signal. Being non-intrusive, would widen the use-case for the objective measure, e.g., to be adopted in a hearing aid where no clean reference signal is available [13, Sec. 2.3].

In written English, consonants convey more information than vowels [14], [15]. This is demonstrated by the following two sentences where asterisks replace vowels and consonants, respectively: 1) Th∗s s∗nt∗nc∗ ∗s l∗g∗bl∗ ∗v∗n w∗th th∗ v∗w∗ls r∗m∗v∗d.[1] 2) ∗∗i∗ ∗e∗∗e∗∗e i∗ o∗ ∗∗e o∗∗e∗ ∗a∗∗ ∗o∗ ∗e∗i∗∗e a∗ ∗o∗∗o∗a∗∗∗ a∗e ∗e∗o∗e∗.[2] The former sentence is legible while the latter is not [16, p. 22]. One explanatory factor is the fact that the number of consonants by far outweighs the number of vowels in the English alphabet, raising larger uncertainty about missing consonants [17], meaning that it is more probable to correctly guess a missing vowel than a missing consonant. Similar experiments have been conducted for the spoken English language. However, speech is often decomposed into phonemes which are the smallest units of sound differentiating one word from another [18, Sec. 9.2A]. Phonemes are then often grouped into the categories of vowel and consonant sounds - amongst others. Literature [19], [20] suggest that - in general - vowel and consonant sounds do not carry the same amount of information in the English language, specifically, vowel sounds are more important for speech intelligibility - contrary to the written English language. The authors of [19], [20] found that by replacing either all vowel sounds or all consonant sounds in sentences with noise, about twice as many words were intelligible when vowel sounds were retained.

Despite results in [19], [20], the decomposition of speech into vowel and consonant sounds might not fully isolate the most prominent factors contributing to speech intelligibility. In [21], the authors report that the human perception system responds primarily to change and not simply linguistic constructs. This is in line with Shannon information theory, which states that

[1] This sentence is legible even with the vowels removed.

[2] This sentence is on the other hand not legible as consonants are removed.

there is no new information in events that do not change or are predictable [17]. In fact, in [22], the authors found that a time-varying tone with increasing and decreasing frequency glides was perceived audible when short time durations were replaced with wideband noise, suggesting that the human brain and auditory system are able to extrapolate missing elements of sounds that are predictable. Based on these observations, we - in this thesis - explore the potentials of predictability as an objective measurable quantity to identify elements of speech which are most important for intelligibility. Specifically, in a spectro-temporal decomposition of a signal, if a frame (the frequency content at a given time interval) is predictable from the frequency content of the previous frames (referred to as previous context), then it does not carry any new information, making it redundant, and hence, irrelevant for speech intelligibility. By quantifying to which extend frames are predictable, we can rank the importance of frames in a signal, in terms of speech intelligibility. If a frame which is predictable from the previous context, is corrupted by acoustic noise, e.g., a noise burst from a passing car, then no speech intelligibility is lost, as any corrupted information can simply be extrapolated. Hence, a predictable frame can be interpreted as being protected from acoustic noise.

The use of predictability or uncertainty for objective evaluation of sound is not unexplored and is often quantified in terms of information theoretical quantities. In [17], the measure cochlea-scaled spectral entropy (CSE) is introduced as a simple measure of difference between successive spectral frames. In [9] and [23], estimates of Shannon mutual information between a processed and corresponding clean signal are used to predict speech intelligibility. Other types of information theoretical measure have been used. Specifically, originally presented for vision in [24] and adopted for audio in [25], [26], the Kullback-Leiber divergence is used for identifying salient events.

There are some potential drawbacks with the aforementioned methods. In [17], CSE does not bare resemblance to the mathematical definition of entropy, raising questions to the credibility. Furthermore, with the definition of CSE in [17], a spectro-temporal tile is assumed to be dependent only on the tile at the previous time frame and the same frequency subband, which is a rather crude assumption. In [9], [23], the proposed methods are intrusive, which as earlier described limits the use-case. In [25], [26], the proposed method is used for salient event detection of sounds and not specifically speech. Though, not necessarily a direct drawback, this raises uncertainty of whether salient events of sound translate to speech intelligibility. Furthermore, these methods are all computed based on a linear frequency scale, whereas the use of a logarithmic scale such as the decibel scale, is arguably more perceptually relevant as is it more relatable to the human auditory system [27], [28, Sec. 1.9].

In this thesis, we propose a non-intrusive perceptually relevant novel estimate of mutual information for discrete time frames of speech signals - inspired by that of [9]. Specifically, our proposed measure aims to identify time frames that are most important for speech intelligibility by estimating the mutual information between each time frame of the signal and the previous context corresponding to approximately 112 [ms].

As mentioned earlier, a predictable frame can be considered protected from acoustic noise. We - in this thesis - examine if elements of speech are protected from acoustic noise by more than one means. Vowel sounds are generally characterized by large sound intensity and a quasi-periodic behavior of the waveform in the time domain, whereas consonant sound are generally characterized by lower sound intensity and aperiodicity of the waveform in the time domain [16, Sec. 2.3], [7, Sec. 3.1]. A quasi-periodic signal must - by all means - be easier predictable than an aperiodic signal. We examine if frames characterized by high sound intensity - which according to [27] are of great importance for speech intelligibility - are more easily predictable than frames characterized by low sound intensity. Hence, we hypothesize that important speech frames are double protected from acoustic noise, by being characterized by large sound intensity and high predictability.

To summarize, in the spoken English language different stimuli conveys different amounts of information. We examine if the degree to which the stimuli is predictable is a contributing factor for speech intelligibility, and thus, seek a way to objectively quantify predictability in speech. To quantify predictability, we consider the information theoretical quantity mutual information, as this quantity has proven useful for intrusive speech intelligibility prediction. In this thesis, we therefore seek to answer:

Research Question:

How can a non-intrusive objective measure of predictability of speech be operationalized in terms of mutual information in order to locate elements of speech, which contribute most to speech intelligibility?

Sub-Questions:

1. In terms of speech intelligibility, is importance of speech frames governed by how predictable they are from past context?

2. Are the speech frames most important for speech intelligibility double protected from acoustic noise, by being characterized by both high sound intensity and high predictability?

Delimitations:

Listening tests are usually conducted in laboratories with control over the acoustic environment. Due to the COVID-19 pandemic, laboratories at Aalborg University have not been accessible at the time of need. Hence, the listening test conducted in this thesis took place in an as-quiet-as-possible location. This location was generally quiet, however, from time to time noise sources such as passing cars and bypassers were audible, introducing a source of error in the listening test.

Outline:

The remainder of this thesis is structured as follows: A display of acoustic processing steps used in this thesis is presented in Chapter 2 followed by information theoretical results presented in Chapter 3. In Chapter 4, objective measures - including our proposed measure - are introduced. In Chapter 5, we present specification for the deep neural network which is embedded in our proposed measure. Chapter 6 first introduces specifications for the aforementioned measures and afterwards introduces the listening test and results thereof. Chapter 7 presents a discussion of the experimental framework and results of the listening test as well as the proposed measure as a whole. The conclusion is given in Chapter 8, and in Chapter 9, we present thoughts on aspects of the proposed measure and the experimental framework for further development.

Acoustic Preprocessing 2

Whether the focus is analysis of speech, noise reduction or intelligibility prediction, speech signals are processed to a state in which desired information is exposed. The degree to which a signal is processed is dependent on the application. However, many applications rely on the same basic processing. This chapter aims to present acoustic processing techniques commonly used in speech processing algorithms.

2.1 Normalization

When processing speech signals, the first step is often to normalize the time domain signal. The operation of normalization is to apply a gain factor to a signal, in order to bring the amplitude of the signal to a desired level. Since normalization amounts to a constant scaling of the entire signal, this operation does note change the signal-to-noise ratio or the dynamics of the signal [29, p. 10]. Two types of normalization are peak normalization and loudness normalization.

2.1.1 Peak Normalization

In peak normalization, the amplitude of a signal is adjusted based on the largest magnitude present in the signal. This is operationalized by dividing each sample of the signal by the largest magnitude, thus bringing the amplitude of the signal in the range $[-1, 1]$. This leads to the following definition.

Definition 2.1 (Peak Normalized Signal)
Let $\boldsymbol{x} \in \mathbb{R}^N$. Then, the peak normalized version, $\boldsymbol{x}'$, of $\boldsymbol{x}$ is given as

$$\boldsymbol{x}' = \frac{\boldsymbol{x}}{\max\{|\boldsymbol{x}|\}},$$

where $\max\{\boldsymbol{x}\}$ returns the largest element of $\boldsymbol{x}$ and $|\boldsymbol{x}|$ denotes element-wise magnitude of $\boldsymbol{x}$. ▲

2.1.2 Loudness Normalization

In loudness normalization, the amplitude of a signal is adjusted, such that the average amplitude of the signal meets a desired level. Often, this average is computed as the root mean square (RMS), which is defined as follows.

Definition 2.2 (Root Mean Square)
Let $\boldsymbol{x} \in \mathbb{R}^N$. Then, the root mean square of $\boldsymbol{x}$ is given as [30, p. 12], [7, (11.42)]

$$\mathrm{RMS}(\boldsymbol{x}) = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]}. \qquad \blacktriangle$$

Before, introducing RMS normalization, the decibel scale is introduced.

Definition 2.3 (Decibel Scale)
Let p denote the sound pressure on the linear scale. The equivalent sound pressure value on the decibel scale is given as [28, Sec 1.9]

$$L_p = 20\log_{10}\left(\frac{p}{p_0}\right)\,[\mathrm{dB}],$$

where p_0 is the reference sound pressure. Likewise, a sound pressure on the decibel scale can be converted to the linear scale according to

$$p = p_0 10^{(L_p/20)}. \tag{2.1}$$

$\blacktriangle$

Definition 2.4 (Root Mean Square Normalized Signal)
Let $\boldsymbol{x} \in \mathbb{R}^N$. Then, the root mean square normalized version, $\boldsymbol{x}'$, of $\boldsymbol{x}$ is given as

$$\boldsymbol{x}' = \sqrt{\frac{NR^2}{\sum_{n=0}^{N-1} x^2[n]}}\boldsymbol{x},$$

where $R = 10^{R_{\mathrm{dB}}/20}$ is the desired linear RMS level, with R_{dB} denoting the desired RMS level in decibels. $\blacktriangle$

Note, that in contrast to peak normalization, loudness normalization does not restrict the amplitude to a given interval.

2.2 Signal Representations

Consider the time domain signal $\boldsymbol{x} \in \mathbb{R}^N$. Most speech processing algorithms rely on frequency analysis, hence, after normalization, a frequency transformation is often applied to $\boldsymbol{x}$. This frequency transformation is often the discrete Fourier transform (DFT). The use of the DFT does, however, have a drawback, in that the temporal structure of the signal is disregarded. To circumvent this, the short-time Fourier transform (STFT) is used.

The STFT partitions the time domain signal into (perhaps) overlapping segments and then applies the DFT to each segments. Thus, the STFT allows for representation of the temporal change in frequency content of $\boldsymbol{x}$.

Definition 2.5 (Short-Time Fourier Transform)
Let $\boldsymbol{x} \in \mathbb{R}^N$. The short-time Fourier transform of $\boldsymbol{x}$ is given as [31, pp. 854-855]

$$X[k,m] = \sum_{p=0}^{P-1} x[mR+p]\, w[p]\, e^{-j2\pi \frac{kp}{P}},$$

where $k = 0, 1, \ldots, P/2$ denotes frequency bin index, $m \geq 0$ denotes temporal frame index, w is an analysis window sequence of length P, and R denotes a shift in samples between successive windows, such that two adjacent windows overlap $P - R$ samples. ▲

After obtaining the short-time spectrum, it is often convenient to convert it to the short-time magnitude spectrum by taking the element-wise absolute value of the complex DFT coefficients. Also, the magnitude spectrum is often converted from the linear scale to the logarithmic scale, as this it more relatable to the human auditory system [28, Sec. 1.9].

2.3 Auditory Filters

One of the fundamental ideas when examining speech intelligibility is to structure the frequency content of a signal in a manner that resembles the human auditory system. How humans perceive frequencies in the cochlea can be modeled as a bank of filters covering the range of frequencies present in speech. Literature [16, p. 78], suggests that the frequency resolution in the human auditory system decreases with frequency. Therefore, in order to model this behavior, auditory filter banks are constructed with non-uniform center frequency spacing as well as increasing bandwidth with increasing frequency. In this section multiple filter banks - crudely reflecting the frequency segmentation of the human perception system - are presented.

A filter bank is applied to each frame of the STFT magnitude spectrum in the following manner. For the ith filter, the magnitude spectrum at the current frame is element-wise multiplied by the magnitude response of the filter and the sum of the products is taken as the filter output, thus resulting in the ith bin in the resulting spectrum. This procedure is repeated for all L filters in the filter bank, resulting in a spectrum with with L frequency bins. Thus, after the filter bank is applied, the spectro-temporal representation retains the temporal indexing, while the resolution over the frequency axis is reduced to the number of filters in the filter bank, see Figure 2.1.

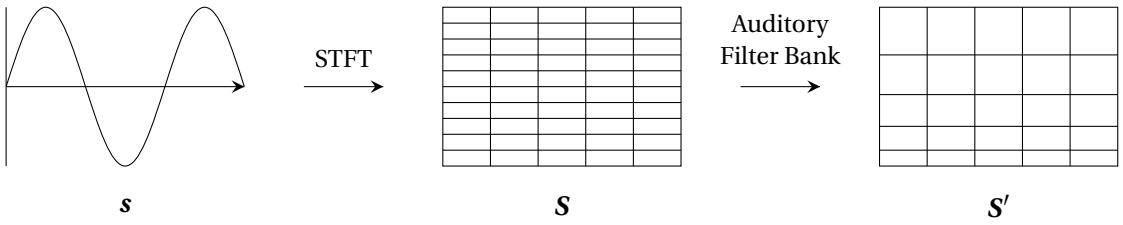


Figure 2.1. The auditory filter bank process. First the time domain signal $\boldsymbol{s}$ is transformed to the short-time magnitude spectrum domain $\boldsymbol{S}$ via the STFT and afterwards each frame is filtered by the auditory filter bank to obtain $\boldsymbol{S}'$.

2.3.1 One-Third Octave Filter Bank

A simple yet widely used filter bank for audio analysis is the one-third octave filter bank. An octave filter spans - as the name suggests - one octave, which is defined as when the upper band frequency is twice that of the lower band frequency. As there are few octaves in the frequency range spoken by humans, this would result in low resolution of frequencies, therefore, the one-third octave filters are used, in which the upper band frequency is equal to the lower band frequency multiplied by a factor $\sqrt[3]{2}$ [28, Sec. 1.10.1].

The filters used in a one-third octave filter bank are rectangular filters centered around a set of center frequencies. These center frequencies can be computed according to

$$f_{c,k} = 2^{k/3} f_{\min} \text{ [Hz]}, \quad k = 0, 1, \ldots L-1,$$

where $f_{\min}$ denotes the lowest desired center frequency and $L \in \mathbb{N}$ denotes the total number of filters. Furthermore, the upper band frequencies and the lower band frequencies of the L filters can be computed as [28, (1.86)]

$$f_{u,k} = 2^{1/6} f_{c,k} \text{ [Hz]}, \quad k = 0, 1, \ldots L-1,$$
$$f_{l,k} = \frac{f_{c,k}}{2^{1/6}} \text{ [Hz]}, \quad k = 0, 1, \ldots L-1,$$

respectively.

An example of an one-third octave filter bank with 23 filters is depicted in Figure 2.2. The filters are 10th order Butterworth filters.

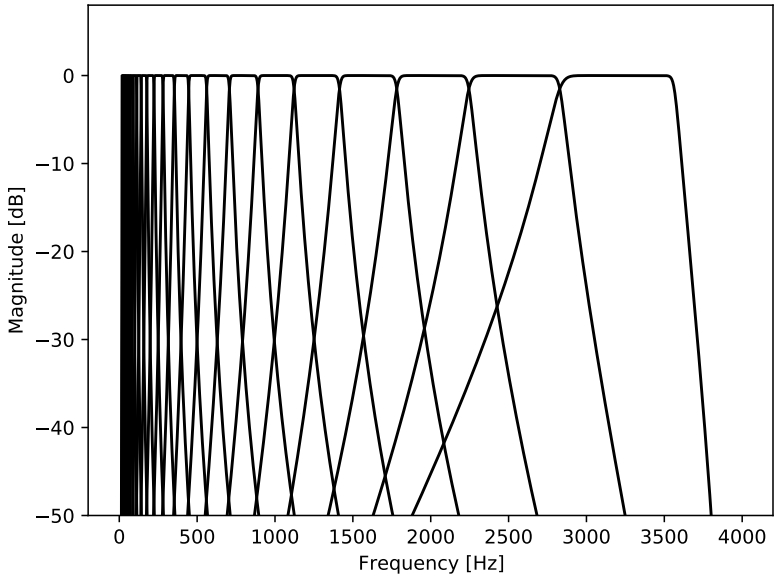


Figure 2.2. One-third octave filter bank of 23 10th order Butterworth filters.

2.3.2 ROEX and Gammatone Filter Banks

Psychoacoustic studies have revealed estimates for the bandwidths of the auditory filters in the human perception system, which is believed to be more accurate than those used in the one-third octave filter bank. The estimate of the bandwidth at frequency f [Hz] is named equivalent rectangular bandwidth (ERB) and - even though multiple approximations exists - is given as [30, p. 76]

$$\mathrm{ERB}(f) = 24.7\left(\frac{4.37}{1000}f + 1\right) \text{ [Hz]}. \tag{2.2}$$

The ERB gives an estimate of the bandwidth at a given frequency, but to create a filter bank a set of center frequencies is needed, which can be constructed with the ERB-scale. By creating a set of equidistant values on the ERB-scale and converting them to frequencies via [30, p. 76]

$$f = \frac{1000}{4.37}\left(10^{\frac{\mathrm{ERBS}}{21.4}} - 1\right) \text{ [Hz]},$$

where ERBS is the value on the ERB-scale, a set of non-uniform center frequencies is obtained. Note, that it is common practice to space the center frequencies 1 ERB apart, e.g. ERBS = $1, 2, \ldots$.

Next, it is a matter of designing the shapes of the filter. Two widely used filters are the rounded exponential (ROEX) and the gammatone [32, Sec. IV].

ROEX Filter Bank

The magnitude response for the ROEX filter is given as [18, (4)]

$$W_{\mathrm{ROEX}}(f) = \left(1 + pg\right)e^{-pg}, \tag{2.3}$$

where p determines both the bandwidth and the shape of the skirts and $g = \frac{|f-f_c|}{f_c}$ is the normalized distance to the center frequency f_c. According to [18, p. 211], the ERB for the ROEX filter is given as

$$\mathrm{ERB}_{\mathrm{ROEX}} = \frac{4f_c}{p} \text{ [Hz]}. \tag{2.4}$$

Substituting (2.2) and (2.4) into (2.3), yields

$$\begin{aligned}
W_{\mathrm{ROEX}}(f) &= \left(1 + p\frac{|f-f_c|}{f_c}\right)e^{-p\frac{|f-f_c|}{f_c}} \\
&= \left(1 + \frac{p}{4f_c}4|f-f_c|\right)e^{-\frac{p}{4f_c}4|f-f_c|} \\
&= \left(1 + (\mathrm{ERB}_{\mathrm{ROEX}})^{-1}4|f-f_c|\right)e^{-(\mathrm{ERB}_{\mathrm{ROEX}})^{-1}4|f-f_c|} \\
&= \left(1 + \frac{4|f-f_c|}{24.7\left(\frac{4.37}{1000}f+1\right)}\right)e^{-\frac{4|f-f_c|}{24.7\left(\frac{4.37}{1000}f+1\right)}}.
\end{aligned}$$

Gammatone Filter Bank

In order to find the magnitude response of the gammatone filter, the impulse response is first considered. The gammatone impulse response is a product of a scaled density of a gamma

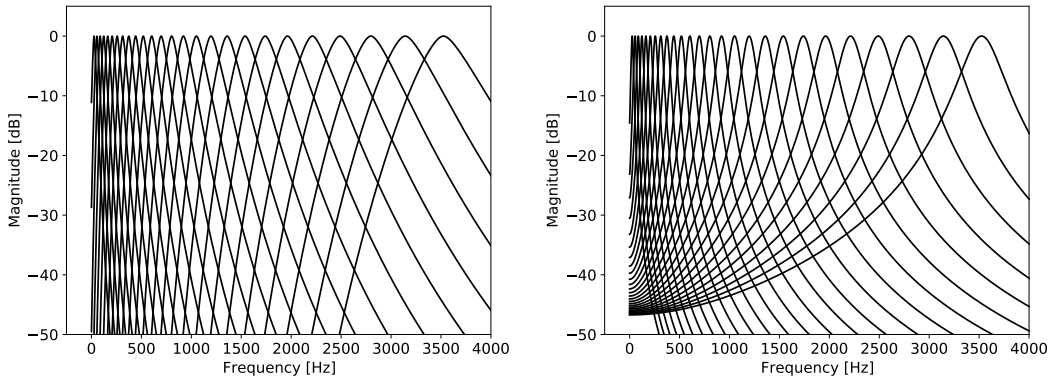
distribution and a sinusoidal wave (the "tone") and is defined as [33, (1)]

$$g(t) = At^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \phi), \quad t \geq 0,$$

where A denotes the amplitude parameter often chosen to make the peak gain equal to unity, t denotes time, n denotes the filter order, b denotes the bandwidth and ϕ denotes the initial phase. Typically, a filter order of $n = 4$ is chosen, as it has been shown that the gammatone filter then provides a good fit to the filters in human auditory system [33, p. 7].

With the center frequencies obtained with the ERB-scale and the bandwidths obtained with (2.2), the gammatone magnitude response can be obtained as the magnitude of the DFT of the gammatone impulse response and afterwards scaled to have peak gain equal to unity.

An example of ROEX and gammatone filter banks are depicted in Figure 2.3. Both filter banks consist of 26 filters with center frequencies spaced 1 ERB apart, resulting in center frequencies in the range $[26-3525]$ [Hz].



(a) ROEX filter bank.

(b) Gammatone filter bank.

Figure 2.3. Magnitude response of ROEX and gammatone filter banks consisting of 26 filters with center frequencies spaced 1 ERB apart, resulting in center frequencies in the range $[26-3525]$ [Hz], with the bandwidths obtained from (2.2).

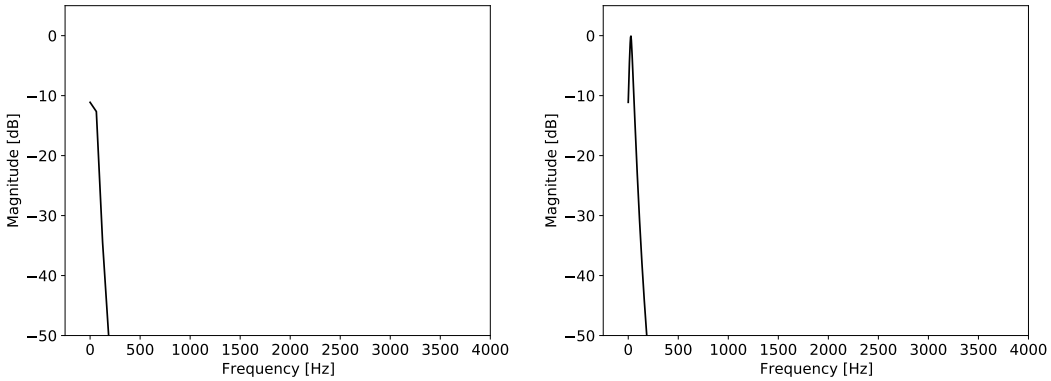
2.3.3 Implementation Aspects

In this section, we shortly present some considerations in terms of the application of the filter banks.

The resolution in frequency of the STFT spectrum is determined by the number of time samples in a frame, as the DFT of a real signal of length N, results in a frequency resolution of $\frac{N}{2}+1$ [31, p. 668]. Since the bandwidth of low frequency auditory filters is small, using frames of short length may result in undersampled filters, as the filter resolution must equal the frequency resolution. This phenomena is depicted in Figure 2.4a, in which a DFT length of 128 samples is used to represent the lowest frequency filter from Figure 2.3a.

To deal with the issue of undersampled low frequency filters, the time domain frames can be zero-padded. This results in a larger frequency resolution and the effect is depicted in Fig-

ure 2.4. From Figure 2.4, it is clear that, if an insufficient DFT-length is used, the filter is not accurately represented, as the peak at 0 [dB], in Figure 2.4a, is not captured. Thus, frames might need to be zero-padded, depending on the frame length and sampling frequency of the signal in question.



(a) 128-length DFT.

(b) 2048-length DFT.

Figure 2.4. The Lowest frequency ROEX filter from Figure 2.3a at different resolutions.

Information Theory 3

As described in Chapter 1, the aim of this thesis is to explore the potential of using predictability to identity elements of speech which contribute most to intelligibility. Predictability - or rather unpredictability - can be quantified mathematically as entropy. This chapter aims to introduce some fundamentals of information theory, useful to quantify information.

The logarithmic operator with no specified base is used to indicate that an arbitrary base can be used. Then, the choice of base simply determines in which unit the result is expressed. In information theoretical work, two bases are frequently used: The natural base e resulting in the unit nat and the base 2 resulting in the unit bit. We use an arbitrary base, as we can simply convert the results from one base to another according to

$$\log_b(x) = \frac{\log_a(x)}{\log_a(b)}. \tag{3.1}$$

For example, if a result is expressed in nats and the desired unit is bits, (3.1) becomes

$$\log_2(x) = \frac{\ln(x)}{\ln(2)},$$

meaning that 1 [nat] corresponds to $\frac{1}{\ln(2)} \approx 1.44$ [bits]. Note that in definitions related to entropy, the conventions that $0\log\left(\frac{0}{0}\right) = 0$, $0\log\left(\frac{0}{q}\right) = 0$ and $p\log\left(\frac{p}{0}\right) = \infty$ is used [34, p. 14]. Furthermore, for notational convenience we use the convention

$$\int_{\boldsymbol{X}} \mathrm{d}\boldsymbol{x} \triangleq \int_{X_{N-1}} \int_{X_{N-2}} \cdots \int_{X_0} \mathrm{d}x_0 \cdots \mathrm{d}x_{N-2} \mathrm{d}x_{N-1},$$

where subscript on integrals denote integration over the support of the indicated variable.

3.1 Differential Entropy

We start by introducing the most fundamental measure of uncertainty, which is entropy or more specifically differential entropy as we consider continuous random variables. The measure of differential entropy captures the uncertainty associated with continuous random variables.

Definition 3.1 (Differential Entropy)
Let $\boldsymbol{X} \in \mathbb{R}^N$ be a continuous random vector with density $f_{\boldsymbol{X}}$. Then, the differential entropy is defined as [34, (8.31)]

$$h(\boldsymbol{X}) = -\int_{\boldsymbol{X}} f_{\boldsymbol{X}}(\boldsymbol{x}) \log\left(f_{\boldsymbol{X}}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}. \tag{3.2}$$

▲

In the field of speech processing, signals are often considered as realizations of a Gaussian stochastic process. Hence, an expression for the differential entropy of a Gaussian random vector is desired.

Theorem 3.2 (Differential Entropy of Multivariate Gaussian Distribution)
Let $\boldsymbol{X} \sim \mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, the differential entropy of $\boldsymbol{X}$ is given as [34, Th. 8.4.1]

$$h(\boldsymbol{X}) = \frac{1}{2} \log\left((2\pi e)^N \det\left(\boldsymbol{\Sigma}\right)\right).$$

▲

Proof
Recall that the density for a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is [35, Def. 3.22]

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{3.3}$$

Inserting (3.3) into (3.2), yields

$$\begin{aligned} h(\boldsymbol{X}) &= -\int_{\boldsymbol{X}} f_{\boldsymbol{X}}(\boldsymbol{x}) \log\left(f_{\boldsymbol{X}}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \\ &= -\mathbb{E}_{\boldsymbol{X}}\left[\log\left(f_{\boldsymbol{X}}(\boldsymbol{x})\right)\right] \\ &= -\mathbb{E}_{\boldsymbol{X}}\left[\log\left(\frac{1}{\left(2\pi\right)^{\frac{N}{2}} \det\left(\boldsymbol{\Sigma}\right)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)\right)\right], \end{aligned} \tag{3.4}$$

where $\mathbb{E}[\cdot]$ denotes expectation. Using of the logarithmic product identity $\log(a) + \log(b) = \log(ab)$ and the logarithmic power identity $b\log(a) = \log(a^b)$, (3.4) becomes

$$\begin{aligned} h(\boldsymbol{X}) &= -\mathbb{E}_{\boldsymbol{X}}\left[-\log\left(\left(2\pi\right)^{\frac{N}{2}}\right) - \log\left(\det\left(\boldsymbol{\Sigma}\right)^{\frac{1}{2}}\right) + \left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)\log(e)\right] \\ &= \frac{1}{2}\log\left(\left(2\pi\right)^N\right) + \frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right) + \frac{1}{2}\log(e)\mathbb{E}_{\boldsymbol{X}}\left[(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]. \end{aligned} \tag{3.5}$$

Next, consider the trace operator denoted as $\mathrm{tr}(\cdot)$. Utilizing that the trace of a scalar is equal to the scalar itself, that trace is invariant under cyclic permutations, and the linearity of the trace operator, (3.5) becomes

$$\begin{aligned} h(\boldsymbol{X}) &= \frac{1}{2}\log\left(\left(2\pi\right)^N\right) + \frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right) + \frac{1}{2}\log(e)\mathbb{E}_{\boldsymbol{X}}\left[\mathrm{tr}\left((\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)\right] \\ &= \frac{1}{2}\log\left(\left(2\pi\right)^N\right) + \frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right) + \frac{1}{2}\log(e)\mathbb{E}_{\boldsymbol{X}}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right)\right] \\ &= \frac{1}{2}\log\left(\left(2\pi\right)^N\right) + \frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right) + \frac{1}{2}\log(e)\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{X}}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right]\right) \\ &= \frac{1}{2}\log\left(\left(2\pi\right)^N\right) + \frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right) + \frac{1}{2}\log(e)\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) \\ &= \frac{1}{2}\log\left(\left(2\pi\right)^N\right) + \frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right) + \frac{1}{2}\log(e)\mathrm{tr}\left(\boldsymbol{I}_N\right), \end{aligned} \tag{3.6}$$

where $\boldsymbol{I}_N$ denotes the N-dimensional identity matrix. As the trace is defined as the sum of the diagonal elements, the trace of the N-dimensional identity matrix equals N, thus, (3.6) becomes

$$h(\boldsymbol{X}) = \frac{1}{2}\log\Big((2\pi)^N\Big) + \frac{1}{2}\log\big(\det\big(\boldsymbol{\Sigma}\big)\big) + \frac{1}{2}\log(e)N.$$

Finally, using the logarithmic product and power identities once again, yields

$$\begin{aligned} h(\boldsymbol{X}) &= \frac{1}{2}\log\Big((2\pi)^N \det\big(\boldsymbol{\Sigma}\big)e^N\Big) \\ &= \frac{1}{2}\log\Big((2\pi e)^N \det\big(\boldsymbol{\Sigma}\big)\Big), \end{aligned}$$

thus, completing the proof.

■

The uncertainty associated with a random vector is captured by the differential entropy. Given the fact that another random vector is known might decrease the uncertain about the first vector. This situation is described by the conditional differential entropy.

Definition 3.3 (Conditional Differential Entropy)
Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^N$ be continuous random vectors. Let $f_{\boldsymbol{X},\boldsymbol{Y}}$ and $f_{\boldsymbol{X}|\boldsymbol{Y}}$ denote joint and conditional densities, respectively. Then, the conditional differential entropy is defined as [34, (8.32)]

$$h(\boldsymbol{X}|\boldsymbol{Y}) = -\int_{\boldsymbol{Y}}\int_{\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\log\big(f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})\big)\,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}. \tag{3.7}$$

▲

3.2 Mutual Information and Kullback-Leiber Divergence

With the differential entropy in mind, two related concepts are introduced in this section. First, the measure of mutual information is introduce, followed by the Kullback-Leiber divergence. The latter of which is often referred to, in literature, as relative entropy.

Mutual information is a measure of the amount of information one set of random variables contains about another set of random variables. In other words, mutual information expresses the reduction in the uncertainty of a set of random variables given another set of random variables and is defined as follows [34, pp. 19].

Definition 3.4 (Mutual Information)
Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^N$ be continuous random vectors with marginal densities $f_{\boldsymbol{X}}$ and $f_{\boldsymbol{Y}}$, respectively. Let $f_{\boldsymbol{X},\boldsymbol{Y}}$ and $f_{\boldsymbol{X}|\boldsymbol{Y}}$ denote joint and conditional densities, respectively. Then, the mutual information is defined as [34, (8.47)]

$$I(\boldsymbol{X};\boldsymbol{Y}) = \int_{\boldsymbol{Y}}\int_{\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\log\left(\frac{f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x})f_{\boldsymbol{Y}}(\boldsymbol{y})}\right)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}. \tag{3.8}$$

▲

It may often be advantageous to expressed mutual information in terms of differential entropy and conditional differential entropy, hence, the following theorem is presented.

Theorem 3.5 (Mutual Information in Terms of Entropy)
Let $\boldsymbol{X}, \boldsymbol{Y} \in$ be continuous random vectors. The mutual information given in (3.8) may be expressed in terms of differential entropy and conditional differential entropy as [34, (8.48)]

$$I(\boldsymbol{X};\boldsymbol{Y}) = h(\boldsymbol{X}) - h(\boldsymbol{X}|\boldsymbol{Y}). \qquad \blacktriangle$$

Proof
Consider the mutual information in (3.8):

$$I(\boldsymbol{X};\boldsymbol{Y}) = \int_{\boldsymbol{Y}}\int_{\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}}(\boldsymbol{y})}\right) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}. \tag{3.9}$$

Using that $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) = f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y})$, (3.9) becomes

$$I(\boldsymbol{X};\boldsymbol{Y}) = \int_{\boldsymbol{Y}}\int_{\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

Using the logarithmic quotient identity $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$ and splitting the integral, yields

$$I(\boldsymbol{X};\boldsymbol{Y}) = -\int_{\boldsymbol{Y}}\int_{\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \log\big(f_{\boldsymbol{X}}(\boldsymbol{x})\big) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} + \int_{\boldsymbol{Y}}\int_{\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \log\big(f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})\big) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}. \tag{3.10}$$

Next, using Definition 3.3 and interchanging the order of integration, (3.10) becomes

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= -\int_{\boldsymbol{Y}}\int_{\boldsymbol{X}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \log\big(f_{\boldsymbol{X}}(\boldsymbol{x})\big) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} - h(\boldsymbol{X}|\boldsymbol{Y}) \\ &= -\int_{\boldsymbol{X}}\int_{\boldsymbol{Y}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \log\big(f_{\boldsymbol{X}}(\boldsymbol{x})\big) \mathrm{d}\boldsymbol{x} - h(\boldsymbol{X}|\boldsymbol{Y}). \end{aligned} \tag{3.11}$$

Integration of the joint density over $\boldsymbol{Y}$, results in the marginal density for $\boldsymbol{X}$, hence, (3.11) becomes

$$I(\boldsymbol{X};\boldsymbol{Y}) = -\int_{\boldsymbol{X}} f_{\boldsymbol{X}}(\boldsymbol{x}) \log\big(f_{\boldsymbol{X}}(\boldsymbol{x})\big) \mathrm{d}\boldsymbol{x} - h(\boldsymbol{X}|\boldsymbol{Y}). \tag{3.12}$$

Finally, using Definition 3.1, yields

$$I(\boldsymbol{X};\boldsymbol{Y}) = h(\boldsymbol{X}) - h(\boldsymbol{X}|\boldsymbol{Y}),$$

thus, completing the proof.

■

Next, the Kullback-Leiber divergence is introduced. The Kullback-Leiber divergence is a measure of distance between two distributions.

Definition 3.6 (Kullback-Leiber Divergence)
Let p_X and q_X denote two density functions. Then, the Kullback-Leiber divergence is given as [34, (2.27)]

$$D(p_X||q_X) = \mathbb{E}_{p_X}\left[\log\left(\frac{p_X(X)}{q_X(X)}\right)\right]. \tag{3.13}$$

▲

The Kullback-Leiber divergence is interpreted as the distance from the true distribution p_X to the assumed distribution q_X, however, note that the Kullback-Leiber divergence is asymmetric, and hence, not strictly speaking a distance.

3.3 Jensen's Inequality

This sections serves the purpose of introducing Jensen's inequality. This result is of crucial importance in the field of information theory, amongst others.

Theorem 3.7 (Jensen's Inequality)
Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function and let $X \in \mathbb{R}$ be a random variable. Then, [34, Th. 2.6.2]

$$\mathbb{E}\big[f(X)\big] \geq f\big(\mathbb{E}[X]\big). \tag{3.14}$$

Moreover, if f is a concave function, then

$$\mathbb{E}\big[f(X)\big] \leq f\big(\mathbb{E}[X]\big). \tag{3.15}$$

▲

Proof
This proof is inspired by [36, pp. 243-244]. First, consider the convex case, i.e., (3.14). Assume that f is a convex function. Then, consider an arbitrary point, and without loss of generality assume that this point is $\mathbb{E}[X]$ and construct the tangent of f at the point $\mathbb{E}[X]$. Next, since f is convex, the value of f is greater than or equal to the value of the tangent for all X, i.e.,

$$f(X) \geq f\big(\mathbb{E}[X]\big) + a\big(X - \mathbb{E}[X]\big), \quad \forall X, \tag{3.16}$$

where a is the slope of the tangent. These geometrical arguments are illustrated in Figure 3.1. Taking the expectation of (3.16), yields

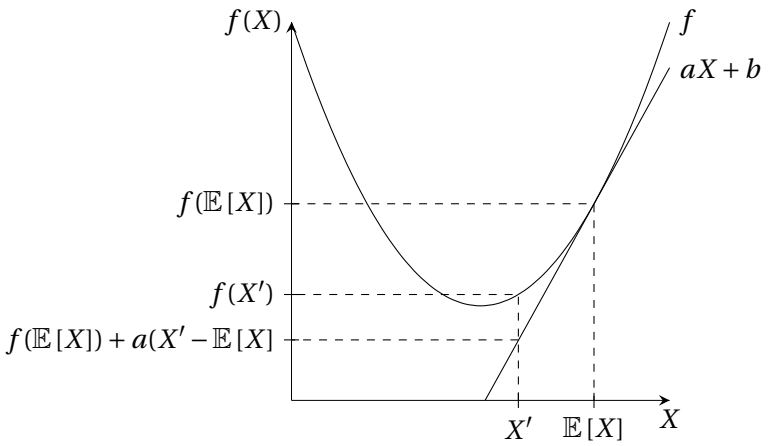


Figure 3.1. Illustration of geometric arguments for the proof of Jensen's inequality in Theorem 3.7. This illustration depicts the convex function f as well as the tangent denoted $aX + b$, of f at the point $\mathbb{E}[X]$. Furthermore, the point X' is depicted to illustrate the inequality in (3.16) enforced by the convexity of f.

$$\begin{aligned}
\mathbb{E}\big[f(X)\big] &\geq \mathbb{E}\big[f\big(\mathbb{E}[X]\big) + a\big(X - \mathbb{E}[X]\big)\big] \\
&= \mathbb{E}\big[f\big(\mathbb{E}[X]\big)\big] + \mathbb{E}\big[a\big(X - \mathbb{E}[X]\big)\big] \\
&= f\big(\mathbb{E}[X]\big) + a\big(\mathbb{E}[X] - \mathbb{E}[X]\big) \\
&= f\big(\mathbb{E}[X]\big),
\end{aligned} \tag{3.17}$$

which concludes the convex case.

For the concave case, i.e., (3.15), consider (3.17). Multiplying both sides by -1, yields

$$\mathbb{E}\left[-f(X)\right] \leq -f\big(\mathbb{E}\left[X\right]\big).$$

Since, the function f is convex, the function $-f$ is concave [37, Sec. 3.1.1], concluding the proof. ■

Predictability as Objective Measure 4

In Chapter 1, we briefly described the concept of utilizing predictability to identify speech frames important for speech intelligibility. In this chapter, we delve deeper into this subject. In Section 4.1, we present the measure of cochlea-scaled spectral entropy and in Section 4.1.1, we introduce the measure of sound intensity. Finally, in Section 4.2, we propose a formulation of mutual information to quantify predictability of speech frames.

4.1 Cochlea-Scaled Spectral Entropy

In this section, the measure of cochlea-scaled spectral entropy (CSE) - originally presented in [17] - is described. The idea of this measure has its foundation in information theory and is based on predictability, hence, the name entropy. The CSE is a simple manner of operationalizing unpredictability in acoustic signals. The measure of CSE aims to capture unpredictability in speech. Since unpredictable elements of speech convey more information than predictable elements, these unpredictable elements are hypothesized to be important for speech intelligibility [17].

The measure of CSE is computed by first RMS normalizing the time domain signal according to Definition 2.4 with $R_{\text{dB}} = 0$. After normalization, the signal is partitioned into non-overlapping frames of 16 [ms]. Next, the magnitude spectrum for each frame of 16 [ms] is obtained by applying the DFT and taking the absolute value of the complex DFT coefficients. Note that this amounts to a STFT with a non-overlapping rectangular analysis window.

The short-time magnitude spectrum is filtered by an auditory filter bank (see Section 2.3) with 33 ROEX filters. The center frequencies for the filters are spaced one ERB apart corresponding to the frequency range $[26 - 7743]$ [Hz]. This range is chosen, as the Texas Instruments/Massachusetts Institute of Technology (TIMIT) speech corpus [38] is used for experimental studies in [17], thus, covering the frequency up to half the sampling frequency of 16 [kHz].

At last, the Euclidean norm is computed between the filter output of adjacent frames. The CSE value is then computed as the sum of the previous five or seven Euclidean distances [17, p.

12391]. The choice of summing over either five (corresponding to 80 [ms]) or seven (corresponding to 112 [ms]) frames is to contain the length of consonant and vowel sounds, respectively [17, p. 12388].

The computation of CSE is summarized in Algorithm 1. Note that in the use of the STFT, as in line 3 in Algorithm 1, the fast Fourier transform (FFT) replaces the DFT, as this is simply a matter of decreasing the computational complexity.

Algorithm 1 Cochlea-Scaled Spectral Entropy (CSE) [17]

Input: $\boldsymbol{s}$: time domain signal
Output: $\boldsymbol{c}$: vector containing CSE values for all frames
1: Set $\boldsymbol{c} = \mathbf{0}$
2: RMS normalize $\boldsymbol{s}$ to 0 [dB]
3: Apply STFT to $\boldsymbol{s}$ - with a non-overlapping rectangular analysis window with length corresponding to 16 [ms] - to obtain $\boldsymbol{S}$
4: Apply auditory filter bank consisting of 33 ROEX filters spaced 1 ERB apart to $|\boldsymbol{S}|$ to obtain $\boldsymbol{S}'$
5: **for all** m **do**
6: $\quad c[m] = \sum_{j=0}^{J-1} \left\| \boldsymbol{S}'[:, m-j] - \boldsymbol{S}'[:, m-j-1] \right\|_2, \quad J \in \{5, 7\}$
7: **end for**

Figure 4.1 exemplifies the process of computing CSE. From top to bottom, the panes depict: RMS normalized time domain representation $\boldsymbol{s}$, short-time magnitude spectrum $|\boldsymbol{S}|$, short-time magnitude spectrum filtered with a 33 ROEX auditory filter bank $\boldsymbol{S}'$ and at the bottom CSE. When short-time magnitude spectra are displayed, they are usually represented on the decibel scale. The computation of CSE does not apply this transformation, making the middle panes difficult to interpret for the human observer.

In [17], [39], the authors report that the measure of CSE better predicts speech intelligibility than traditional distinction between vowel and consonant sounds. Specifically, they found that when replacing segments characterized by high CSE with noise, speech intelligibility was more degraded than when replacing segments characterized by low CSE. The authors claim that the close relation between CSE and speech intelligibility is due to the fact that CSE is a measure of potential information, which in turn greatly contributes to the human perception system. Despite the seemingly great potential of CSE, concerns have been raised about the measure in [27]. We address these concerns in the following section.

4.1.1 CSE vs. Intensity

In [27], the authors raise concern with some aspects of the CSE computation. They question the fact that a linear scale is used to represent the magnitude spectra - and thus the filter outputs - in contrast to a logarithmic scale such as the decibel scale. This is because a logarithmic scale is more commonly used in auditory application, it is more relatable to the human auditory system and the use of the linear scale might result in CSE being dominated by large sound intensity [28, Sec. 1.9], [27, Sec. 1]. The issue of using a linear amplitude scale for CSE is exem-

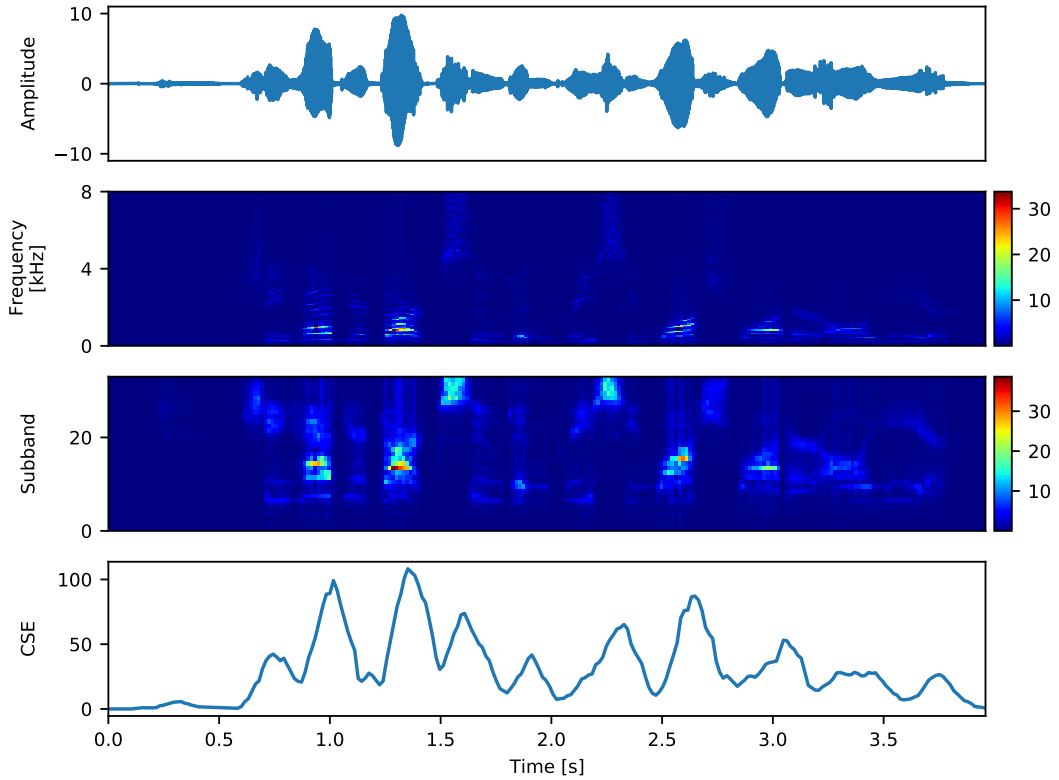


Figure 4.1. One speech sentence from the TIMIT speech corpus. From the top, the panes depict: 1) RMS normalized time domain representation $\boldsymbol{s}$, 2) Short-time magnitude spectrum $|\boldsymbol{S}|$, 3) ROEX-filtered short-time magnitude spectrum $\boldsymbol{S}'$, and 4) CSE with $J = 7$.

plified in Example 4.2. Before proceeding to the example, we introduce the measure of sound intensity.

Definition 4.1 (Sound Intensity)

Let $\boldsymbol{x} \in \mathbb{R}^N$. Then, the sound intensity of $\boldsymbol{x}$ is defined as

$$i = \sum_{i=0}^{N-1} x^2[i].$$

▲

Example 4.2

This example is based on [27, Sec. 1]. Recall the relation between the decibel and linear scale for sound pressure in (2.1). Consider the change in amplitude from 70 [dB] to 71 [dB] and from 30 [dB] to 31 [dB] with arbitrary reference level p_0, i.e.,

$$\frac{p_0 \cdot 10^{(71/20)} - p_0 \cdot 10^{(70/20)}}{p_0 \cdot 10^{(31/20)} - p_0 \cdot 10^{(30/20)}} \approx 100.$$

This illustrates that a change in amplitude of 1 [dB] - which is perceived similar across the decibel scale - results in different CSE values depending on the level of the sound pressure. Specifically, the change from 70 [dB] to 71 [dB] results in a CSE value which is 100 times larger than that of a change from 30 [dB] to 31 [dB], exemplifying the fact that CSE might be dominated by high amplitude sounds. ▲

In [27], CSE is compared to two measures. The first measure is obtained by applying the windowing process of the STFT to the time domain signal, resulting in the waveform being divided into smaller frames. For each frame, the sound intensity is computed according to Definition 4.1. This is done in order to examine if CSE really is dominated by high amplitude sounds. We shall refer to sound intensity measured on a frame basis as INT.

The second measure which is compared to CSE in [27], is CSE computed on the decibel scale (termed dB-CSE), which is arguably more perceptually relevant as it takes into account the logarithmic behavior of the human auditory system. The measure dB-CSE is computed like CSE with the only difference being that the auditory filter outputs are transformed to the decibel scale, i.e., line 6 in Algorithm 1 becomes

$$c_{\mathrm{dB}}[m] = \sum_{j=0}^{J-1} \left\| 20\log_{10}\left(\boldsymbol{S}'[:, m-j]\right) - 20\log_{10}\left(\boldsymbol{S}'[:, m-j-1]\right) \right\|_2, \quad J \in \{5, 7\}.$$

Note that, due to the logarithmic quotient identity $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$, dB-CSE describes a ratio between linear filter outputs as opposed to CSE, which describes a difference.

In [27], two experiments were conducted. In the first experiment, correlation between CSE and INT as well as correlation between CSE and dB-CSE was examined. In this experiment, the speech sentences were first RMS normalized to the same level and afterwards the first and last 80 [ms] were discarded. The resulting sentences were partitioned into 112 [ms] non-overlapping segments, in contrast to the "sliding" evaluation in the original CSE algorithm. For CSE and dB-CSE, each 112 [ms] segment was further divided into seven non-overlapping 16 [ms] frames. These frames underwent the frequency transformation and filter bank process associated with the CSE. Then, the Euclidean distance between successive filter outputs (on the decibel scale for dB-CSE) were computed and the sum of these six Euclidean distances was taken as the value for the given 112 [ms] segment. For INT, the value for the 112 [ms] segment was computed by application of Definition 4.1 to the waveform in this segment. Hence, for each sentence, vectors $\boldsymbol{i}$, $\boldsymbol{c}$, and $\boldsymbol{c}_{\mathrm{dB}}$ containing the values of INT, CSE, and dB-CSE for each segment, respectively, were obtained. In [27], correlation was measured using the Spearman's correlation coefficient, which is given in the following definition.

Definition 4.3 (Spearman's Correlation Coefficient)
Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$. Then, the Spearman's correlation coefficient is defined as [35, Def. 7.7]

$$\rho_S = \frac{\sum_{n=0}^{N-1}\left(r_{\boldsymbol{x}}[n] - \overline{r_{\boldsymbol{x}}}\right)\left(r_{\boldsymbol{y}}[n] - \overline{r_{\boldsymbol{y}}}\right)}{\sqrt{\sum_{n=0}^{N-1}\left(r_{\boldsymbol{x}}[n] - \overline{r_{\boldsymbol{x}}}\right)^2 \sum_{n'=0}^{N-1}\left(r_{\boldsymbol{y}}[n'] - \overline{r_{\boldsymbol{y}}}\right)^2}},$$

where $\boldsymbol{r}_{\boldsymbol{x}}$ and $\boldsymbol{r}_{\boldsymbol{y}}$ denote the ranks of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and $\overline{(\cdot)}$ denotes vector mean.

Furthermore, if no ties occur in the ranking of the data, the Spearman's correlation coefficient is defined as [35, Prop. 7.20]

$$\rho_S = 1 - \frac{6}{N(N^2-1)} \sum_{n=0}^{N-1} \left(r_{\boldsymbol{x}}[n] - r_{\boldsymbol{y}}[n]\right)^2. \qquad \blacktriangle$$

In [27], a Spearman's correlation coefficient between CSE and INT was computed by Definition 4.3 with $\boldsymbol{c}, \boldsymbol{i}$ as $\boldsymbol{x}, \boldsymbol{y}$ for each sentence. Likewise, a Spearman's correlation coefficient between CSE and dB-CSE was obtained with $\boldsymbol{c}, \boldsymbol{c}_{\text{dB}}$ as $\boldsymbol{x}, \boldsymbol{y}$ for each sentence. In [27], results were obtained for the AZBio speech corpus [40]. The average Spearman's correlation coefficient, across all sentences, between CSE and INT was found to be $\rho_S = 0.926$, while the average Spearman's correlation coefficient between CSE and dB-CSE was found to be $\rho_S = -0.168$.

We recreate the correlation experiment of [27], however, we compute the Spearman's correlation coefficients for the sentences of the TIMIT speech corpus. The results of our experiment are presented in Figure 4.2, where each dot represents a correlation coefficient. On the left hand side of the figure, correlation coefficients between CSE and INT are depicted while correlation coefficients between CSE and dB-CSE are depicted on the right hand side. Note that the horizontal dispersion between dots within each of the two groups is solely for visualization purposes. The results presented in Figure 4.2 are somewhat similar to those of [27], in that

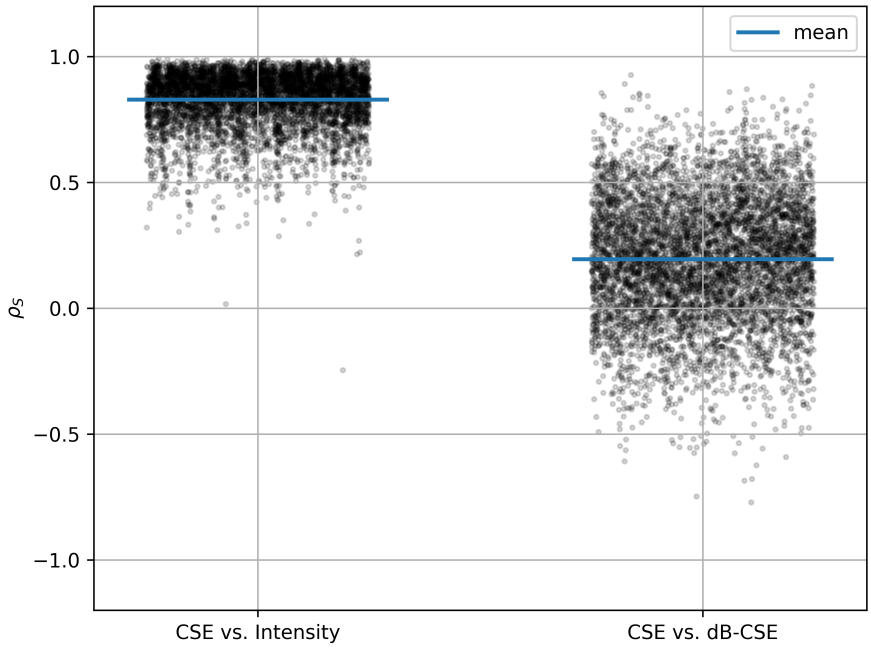


Figure 4.2. Spearman's correlation coefficients between CSE and INT and between CSE and dB-CSE. One correlation coefficient is computed for each speech signal (sentence) in the TIMIT speech corpus and represented as a dot. Means are represented as blue lines with values of $\rho_S = 0.829$ for CSE vs. intensity and $\rho_S = 0.195$ for CSE vs. dB-CSE. Horizontal dispersion between dots within each of the two groups is solely for visualization purposes.

they highlight the same two findings. Firstly, a relatively large correlation is found between CSE and INT, supporting the hypothesis that CSE is dominated by sound intensity. Secondly, low correlation is found between CSE and dB-CSE.

In the second experiment presented in [27], segments of speech signals were replaced with noise. The replaced segments were determined by the value of CSE, INT, or dB-CSE in these segments. The resulting sentences containing segments of noise were then presented for a set

of listeners. The percentage of correctly identified words was taken as the speech intelligibility score. In this experiment, the authors found that speech intelligibility scores were degraded more when segments characterized by high CSE or high INT were replaced, than when segments characterized by low CSE or low INT were replaced. This is inline with the findings of [17] for CSE. Also, this was found that no such tendency was present for the measure of dB-CSE.

From their experiments, the authors of [27] concluded that CSE is closely related to sound intensity, and that it is not possible to conclude whether predictability measured as CSE, as proposed in [17], or simply intensity best predicts speech intelligibility. Furthermore, the authors of [27] concluded that dB-CSE fails to predict speech intelligibility, although it is arguably more perceptually relevant than CSE.

In [41], experiments like those of [17], [27] is reported, in which they replace segments characterized by high and low levels of CSE and INT with noise. They found - like in [27] - that the replacement of high CSE and high INT segments results in larger degradation of speech intelligibility, when compared to the replacement of low CSE and low INT segments. They did, however, find that speech intelligibility is degraded more with the replacement of high CSE segments than with the replacement of high INT segments and they pointed out that this was also the case for the experiment conducted in [27]. This finding suggests that CSE captures some contributing factor to speech intelligibility that INT does not.

Since CSE seems to capture some aspect of speech intelligibility that INT does not, there might be some truth to the hypothesis that (un)predictability is important for speech intelligibility. However, CSE does not bare resemblance to the mathematical definition of entropy, and hence, it might not accurately approximate potential information as desired. Therefore, in this thesis, we assess the potentials of predictability as an objective measurable quantity to identify elements of speech which are most important for speech intelligibility.

4.2 Proposed Predictive Measure

In this thesis, we aim to develop a predictive measure which is able to identity frames that are most important for speech intelligibility. Predictability can be formulated in a multitude of ways. We do, however, restrict our field of view to the field of information theory. The measure CSE - though providing promising results - does, as mentioned, not coincide with the mathematical definition of entropy. This motivates the introduction of a predictive measure which has a more rigorous connection to mathematical definitions of information theoretical quantifies.

In the field of information theory, there are three primary measures, namely, entropy (see Definition 3.1), mutual information (see Definition 3.4), and the Kullback-Leiber divergence (see Definition 3.6). We are aiming to construct a measure which measures uncertainty by taking into account previous spectro-temporal frames. We will refer to these previous spectro-temporal frames as previous context. Constructing such a measure, naturally points towards

the latter two measures, as these are functions of two arguments.

As mentioned in Chapter 1, mutual information has proven useful in predicting speech intelligibility when measured between a clean and a noisy/processed signal [9], [23]. For the Kullback-Leiber divergence, a method termed Bayesian surprise - originally presented for vision in [24] and adopted for audio in [25], [26] - has proven useful in detection of salient visual and acoustic events. This method measures the Kullback-Leiber divergence between the prior and posterior distribution of the stimuli. Salient events of sound does not necessarily translate to speech intelligibility, hence, in this thesis, we consider mutual information, as this quantity has proven useful in speech processing.

To obtain an expression for mutual information, the general approach of [9] is adopted. In [9], a lower bound on mutual information between clean and processed speech is considered, making the measure intrusive, in order to predict the effect of the processing on speech intelligibility. In this adaptation, the mutual information between successive magnitude spectrum frames of the same signal is considered, resulting in a non-intrusive measure.

The acoustic preprocessing of [9] is adopted with slight alterations and will be elaborated upon in the following. We use capital letter to represent stochastic processes and variables, while lower-case letters are used to represent corresponding realizations. To this end, let $\boldsymbol{S}_t$ denote the time domain representation of a stochastic process modeling the speech signal of interest. The STFT (see Definition 2.5) is applied to $\boldsymbol{S}_t$ to obtain a spectro-temporal representation $\tilde{\boldsymbol{S}}$. Next, an one-third octave auditory filter bank (see Section 2.3.1) is applied as

$$S[i,m] = 20\log_{10}\left(\sqrt{\sum_{k\in C_i}\left|\tilde{S}[k,m]\right|^2}\right)\ [\text{dB}], \quad i = 0,1,\ldots,L-1, \tag{4.1}$$

where C_i denotes the frequency index set corresponding to the ith one-third octave filter. This grouping of frequencies corresponds to the one-third octave filter bank process. In contrast to [9], we consider the filter outputs on the decibel scale, as this scale is arguably more relevant to the human auditory system [27], as mentioned in Section 4.1.1.

Our aim is to measure the amount of information in a given frame which is captured by the previous context. Let this previous context be defined as

$$\boldsymbol{X}_m \triangleq \begin{bmatrix}\boldsymbol{S}[:,m-M]^T & \boldsymbol{S}[:,m-(M-1)]^T & \cdots & \boldsymbol{S}[:,m-1]^T\end{bmatrix}^T, \tag{4.2}$$

where $M > 0$ denotes the number of previous frames to be included in the context. Thus, at frame m, the mutual information between $\boldsymbol{S}[:,m]$ and $\boldsymbol{X}_m$ is considered. According to Theorem 3.5, this may be expressed as

$$I\big(\boldsymbol{S}[:,m];\boldsymbol{X}[m]\big) = h\big(\boldsymbol{S}[:,m]\big) - h\big(\boldsymbol{S}[:,m]|\boldsymbol{X}[m]\big). \tag{4.3}$$

For notational convenience, the frequency subband and frame indices are omitted where possible. To obtain the mutual information in (4.3), expressions for the differential entropy $h(\boldsymbol{S})$

and the conditional differential entropy $h(\boldsymbol{S}|\boldsymbol{X})$ is needed. As we make no prior assumptions about the underlying distribution of the data, exact expressions will not be obtained. Instead, a lower bound on the mutual information is obtained by finding an upper bound on $h(\boldsymbol{S}|\boldsymbol{X})$ and a lower bound on $h(\boldsymbol{S})$.

4.2.1 Upper Bound on Conditional Differential Entropy

As described in Section 4.2, in order to obtain a lower bound on the mutual information, an upper bound on the conditional differential entropy is needed. Considering the conditional differential entropy in Definition 3.3, we obtain

$$\begin{aligned} h(\boldsymbol{S}|\boldsymbol{X}) &= -\int_{\boldsymbol{X}}\int_{\boldsymbol{S}} f_{\boldsymbol{S},\boldsymbol{X}}(\boldsymbol{s},\boldsymbol{x}) \log\left(f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x})\right) \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{x} \\ &\overset{(a)}{=} -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\boldsymbol{S},\boldsymbol{X}}(\boldsymbol{s},\boldsymbol{x}) \log\left(f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x})\right) \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{x} \\ &\overset{(b)}{=} -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x}) \log\left(f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x})\right) \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{x} \\ &= -\int_{-\infty}^{\infty} f_{\boldsymbol{X}}(\boldsymbol{x}) \int_{-\infty}^{\infty} f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x}) \log\left(f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x})\right) \mathrm{d}\boldsymbol{s}\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{4.4}$$

where (a) comes from the fact that each element of $\boldsymbol{S}$ and $\boldsymbol{X}$ are represented on the decibel scale, and hence, supported on the real line, and where (b) is obtained by recalling that a joint density is the product of the conditional and marginal density.

Let $\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}}$ denote the covariance matrix of the random vector distributed according to the density $f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x})$. Then, (4.4) may be upper bounded by considering the maximum entropy distribution, which is the multivariate Gaussian distribution [34, Th. 8.6.5]. By using the differential entropy of a multivariate Gaussian random vector, as given in Theorem 3.2, (4.4) becomes

$$\begin{aligned} h(\boldsymbol{S}|\boldsymbol{X}) &\leq \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(\boldsymbol{x}) \left(\frac{1}{2}\log\left((2\pi e)^L \det\left(\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}}\right)\right)\right) \mathrm{d}\boldsymbol{x} \\ &= \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{2}\log\left((2\pi e)^L \det\left(\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}}\right)\right)\right] \\ &\leq \frac{1}{2}\log\left((2\pi e)^L \det\left(\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}}\right]\right)\right), \end{aligned} \tag{4.5}$$

where the last inequality in (4.5) is obtained by the use of Theorem 3.7, as the logarithm of the determinant of a symmetric positive definite matrix is concave [37, p. 73]. As $\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}}$ is a covariance matrix, it is symmetric and positive semi-definite. Furthermore, assuming that no element in $\boldsymbol{S}|\boldsymbol{X}$ is a linear combination of the other elements, $\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}}$ becomes positive definite.

Now consider $\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}}$ in (4.5) and let

$$\boldsymbol{\mu}_{\boldsymbol{S}|\boldsymbol{x}} = \int_{-\infty}^{\infty} \boldsymbol{y} f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x}) \mathrm{d}\boldsymbol{y}$$

denote the mean of the random vector distributed according to the density $f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x})$. Then, the covariance is expressed as

$$\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}} = \int_{-\infty}^{\infty} \left(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{S}|\boldsymbol{x}}\right)\left(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{S}|\boldsymbol{x}}\right)^T f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x}) \mathrm{d}\boldsymbol{y}. \tag{4.6}$$

Considering then the mean $\boldsymbol{\mu}_{\boldsymbol{S}|\boldsymbol{x}}$. This conditional expectation is equal to the minimum mean square error (MMSE) estimator of $\boldsymbol{S}$ given $\boldsymbol{x}$ (see Theorem A.1). Thus, by letting $\hat{\boldsymbol{s}}_{\text{MMSE}}(\boldsymbol{x})$ denote the MMSE estimator of $\boldsymbol{S}$ given $\boldsymbol{x}$, the covariance in (4.6) may be expressed as

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}} &= \int_{-\infty}^{\infty} \big(\boldsymbol{y} - \hat{\boldsymbol{s}}_{\text{MMSE}}(\boldsymbol{x})\big)\big(\boldsymbol{y} - \hat{\boldsymbol{s}}_{\text{MMSE}}(\boldsymbol{x})\big)^T f_{\boldsymbol{S}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x})\mathrm{d}\boldsymbol{y} \\ &\triangleq D_{\text{MMSE}}(\boldsymbol{x}). \end{aligned} \tag{4.7}$$

Furthermore, let D_{MMSE} denote (4.7) averaged across all realizations of $\boldsymbol{X}$, i.e.,

$$D_{\text{MMSE}} \triangleq \int_{-\infty}^{\infty} D_{\text{MMSE}}(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{4.8}$$

Now, by returning to the conditional differential entropy in (4.5), we obtain

$$\begin{aligned} h(\boldsymbol{S}|\boldsymbol{X}) &\leq \frac{1}{2}\log\big((2\pi e)^L \det\big(\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{\Sigma}_{\boldsymbol{S}|\boldsymbol{x}}\right]\big)\big) \\ &\overset{(a)}{=} \frac{1}{2}\log\big((2\pi e)^L \det\big(\mathbb{E}_{\boldsymbol{X}}\left[D_{\text{MMSE}}(\boldsymbol{x})\right]\big)\big) \\ &\overset{(b)}{=} \frac{1}{2}\log\big((2\pi e)^L \det\big(D_{\text{MMSE}}\big)\big) \\ &\triangleq h_{\text{MMSE}}(\boldsymbol{S}|\boldsymbol{X}), \end{aligned} \tag{4.9}$$

where (a) and (b) comes from (4.7) and (4.8), respectively.

We have now introduced an expression for an upper bound on the condition differential entropy $h(\boldsymbol{S}|\boldsymbol{X})$. In order to evaluate this quantity, the MMSE estimator $\hat{\boldsymbol{s}}_{\text{MMSE}}(\boldsymbol{x})$ of $\boldsymbol{S}$ given $\boldsymbol{x}$ must be formed. We propose to realize this MMSE estimator via a deep neural network (DNN), the development of which, is described in Chapter 5. With the expression for the upper bound on conditional differential entropy in place, we turn our attention to the lower bound on the differential entropy.

4.2.2 Lower Bound on Differential Entropy

In this section, we present a lower bound on differential entropy, $h(\boldsymbol{S})$, developed in [42] which will be adopted in this work.

In [42, Th. 4], the authors present a result stating that if the density of the underlying distribution of $\boldsymbol{S}$ is symmetric and logarithmically concave, then, the differential entropy is bounded from below as

$$h(\boldsymbol{S}) \geq \frac{L}{2}\log\left(\frac{4\sqrt{2}(L+2)\det\big(\boldsymbol{\Sigma}_{\boldsymbol{S}}\big)^{\frac{1}{L}}}{e^2 L^2}\right), \tag{4.10}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{S}}$ is the covariance matrix of $\boldsymbol{S}$. Note that a function is logarithmically concave if the logarithm of the function is concave.

Auditory filter outputs for speech are often modeled as being Gaussian distributed. The multivariate Gaussian distribution is logarithmically concave [37, Ex. 3.39] as well as symmetric. Furthermore, preliminary experiments suggested that assuming logarithmic concavity and

symmetry of $\boldsymbol{S}$ is reasonable. For these reasons, we use the lower bound in (4.10) and define

$$h_{\mathrm{LB}}(\boldsymbol{S}) \triangleq \frac{L}{2}\log\left(\frac{4\sqrt{2}(L+2)\det\left(\boldsymbol{\Sigma}_{\boldsymbol{S}}\right)^{\frac{1}{L}}}{e^2L^2}\right) \tag{4.11}$$

to be the lower bound on the differential entropy.

4.2.3 Estimate of Mutual Information

With the upper bound on the conditional differential entropy $h(\boldsymbol{S}|\boldsymbol{X})$ and the lower bound on the differential entropy $h(\boldsymbol{S})$ in place, we can now obtain an expression for a lower bound on the mutual information $I(\boldsymbol{S};\boldsymbol{X})$. The mutual information is a non-negative measure, however, the introduction of the upper bounding on the condition differential entropy $h(\boldsymbol{S}|\boldsymbol{X})$ might result in negative values of the mutual information. To circumvent this issue, negative values are simply set to zero, hence, we formulate the estimate of the mutual information as

$$I_{\mathrm{LB}}(\boldsymbol{S};\boldsymbol{X}) \triangleq \max\left\{h_{\mathrm{LB}}(\boldsymbol{S}) - h_{\mathrm{MMSE}}(\boldsymbol{S}|\boldsymbol{X}), 0\right\}. \tag{4.12}$$

Inserting (4.9) and (4.11) into (4.12), yields

$$\begin{aligned}
I_{\mathrm{LB}}(\boldsymbol{S};\boldsymbol{X}) &= \max\left\{\frac{L}{2}\log\left(\frac{4\sqrt{2}(L+2)\det\left(\boldsymbol{\Sigma}_{\boldsymbol{S}}\right)^{\frac{1}{L}}}{e^2L^2}\right) - \frac{1}{2}\log\left((2\pi e)^L\det\left(D_{\mathrm{MMSE}}\right)\right), 0\right\} \\
&= \max\left\{\frac{1}{2}\log\left(\det\left(\boldsymbol{\Sigma}_{\boldsymbol{S}}\right)\right) + \frac{L}{2}\log\left(\frac{4\sqrt{2}(L+2)}{e^2L^2}\right) - \frac{L}{2}\log\left(2\pi e\right) - \frac{1}{2}\log\left(\det\left(D_{\mathrm{MMSE}}\right)\right), 0\right\} \\
&= \max\left\{\frac{1}{2}\log\left(\frac{\det\left(\boldsymbol{\Sigma}_{\boldsymbol{S}}\right)}{\det\left(D_{\mathrm{MMSE}}\right)}\right) + \frac{L}{2}\log\left(\frac{2\sqrt{2}(L+2)}{\pi e^3L^2}\right), 0\right\}.
\end{aligned} \tag{4.13}$$

With the expression in (4.13), we have obtained a lower bound on the mutual information between an auditory filter output $\boldsymbol{S}$ and the previous context across time and frequency $\boldsymbol{X}$.

Through experimentation, we found issues with the lower bound on the mutual information in (4.13), in that the lower bound on the differential entropy in (4.11) is relaxed to an extend where, due to the number of frequency subbands L, the second logarithmic term in (4.13) dominates $I_{\mathrm{LB}}(\boldsymbol{S};\boldsymbol{X})$. When measured on a small set of randomly chosen sentences of the TIMIT speech corpus, the value of $I_{\mathrm{LB}}(\boldsymbol{S};\boldsymbol{X})$ will only rarely be non-zero.

To bypass this issue, we estimate $I_{\mathrm{LB}}(\boldsymbol{S};\boldsymbol{X})$ by considering the mutual information between each element of $\boldsymbol{S}$ and the complete context $\boldsymbol{X}$. In the same manner as in Section 4.2.1, the conditional differential entropy $h(S|\boldsymbol{X})$ is for each subband upper bounded as

$$h(S|\boldsymbol{X}) \leq \frac{1}{2}\log\left(2\pi e\mathbb{E}\left[\sigma^2_{S|\boldsymbol{X}}\right]\right), \tag{4.14}$$

where $\sigma^2_{S|\boldsymbol{x}}$ is the variance of the random variable distributed according to the density $f_{S|\boldsymbol{x}}(s|\boldsymbol{x})$. Again, the conditional mean $\mu_{S|\boldsymbol{x}}$ associated with the density $f_{S|\boldsymbol{x}}(s|\boldsymbol{x})$ is equivalent to the MMSE

estimator $\hat{s}_{\mathrm{MMSE}}(\boldsymbol{x})$ of S given $\boldsymbol{x}$. Hence, the variance may be expressed as

$$\sigma^2_{S|\boldsymbol{x}} = \int_{-\infty}^{\infty} \left(y - \hat{s}_{\mathrm{MMSE}}(\boldsymbol{x})\right)^2 f_{S|\boldsymbol{x}}(y|\boldsymbol{x}) \mathrm{d}y \tag{4.15}$$

$$\triangleq D_{\mathrm{MMSE}}(\boldsymbol{x}). \tag{4.16}$$

Note that the quantity $D_{\mathrm{MMSE}}(\boldsymbol{x})$ has been redefined in (4.16) as the definition in (4.7) will no longer be used. The quantities $h_{\mathrm{MMSE}}(S|\boldsymbol{X})$, $h_{\mathrm{LB}}(S|\boldsymbol{X})$, and $I_{\mathrm{LB}}(S;\boldsymbol{X})$ will likewise be redefined in the following.

As in (4.8), (4.16) is the average over all realizations of $\boldsymbol{X}$, i.e.,

$$D_{\mathrm{MMSE}} = \int_{-\infty}^{\infty} D_{\mathrm{MMSE}}(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{4.17}$$

and hence, (4.14) becomes

$$\begin{aligned} h(S|\boldsymbol{X}) &\leq \frac{1}{2}\log\left(2\pi e \mathbb{E}_{\boldsymbol{X}}\left[\sigma^2_{S|\boldsymbol{x}}\right]\right) \\ &\overset{(a)}{=} \frac{1}{2}\log\left(2\pi e \mathbb{E}_{\boldsymbol{X}}\left[D_{\mathrm{MMSE}}(\boldsymbol{x})\right]\right) \\ &\overset{(b)}{=} \frac{1}{2}\log\left(2\pi e D_{\mathrm{MMSE}}\right) \\ &\triangleq h_{\mathrm{MMSE}}(S|\boldsymbol{X}), \end{aligned} \tag{4.18}$$

where (a) and (b) comes from (4.16) and (4.17), respectively.

According to [42, Th. 3], if the density of the underlying distribution of S is logarithmically concave, then, the differential entropy is bounded from below as

$$\begin{aligned} h(S) &\geq \frac{1}{2}\log\left(4\sigma^2_S\right) \\ &\triangleq h_{\mathrm{LB}}(S), \end{aligned} \tag{4.19}$$

where σ^2_S is the variance of S. As for (4.10), we rely on the logarithmic concavity of the Gaussian distribution.

Now, we redefine (4.12) in terms of (4.18) and (4.19), i.e.,

$$\begin{aligned} I_{\mathrm{LB}}(S;\boldsymbol{X}) &\triangleq \max\left\{h_{\mathrm{LB}}(S) - h_{\mathrm{MMSE}}(S|\boldsymbol{X}), 0\right\} \\ &= \max\left\{\frac{1}{2}\log\left(4\sigma^2_S\right) - \frac{1}{2}\log\left(2\pi e D_{\mathrm{MMSE}}\right), 0\right\} \\ &= \max\left\{\frac{1}{2}\log\left(4\right) + \frac{1}{2}\log\left(\sigma^2_S\right) - \frac{1}{2}\log\left(2\pi e\right) - \frac{1}{2}\left(D_{\mathrm{MMSE}}\right), 0\right\} \\ &= \max\left\{\frac{1}{2}\log\left(\frac{\sigma^2_S}{D_{\mathrm{MMSE}}}\right) - \frac{1}{2}\log\left(\frac{\pi e}{2}\right), 0\right\}. \end{aligned} \tag{4.20}$$

With the expression in (4.20), we have obtained a lower bound on the mutual information between an auditory filter output S and the previous context across time and frequency $\boldsymbol{X}$. The

expression in (4.20) is for one subband, and hence, in order to obtain a per-frame measure, the expression in (4.20) is summed across all L subbands. Thus, at frame index m, we define

$$\widehat{I}_m \triangleq \sum_{l=0}^{L-1} I_{\mathrm{LB}}\big(S[l,m]\,;\boldsymbol{X}_m\big). \tag{4.21}$$

An advantage of using (4.21), as opposed to (4.13), is that in numerical computations, determinants might behave erratically.

A frame characterized by low mutual information means large uncertainty of the frame given the context, whereas larger mutual information means reduced uncertainty of the frame. This means that frames characterized by low mutual information are less predictable, hence, we expect these frames to be more important for speech intelligibility compared to frames characterized by high mutual information. Worth noting is that this is opposite to the measures INT and CSE where high value frames are taken to be most important for speech intelligibility.

In order to compute the per-frame mutual information based measure in (4.21), which from this point on is referred to as MI, the quantities σ_S^2 and D_{MMSE} are needed for each frequency subband. The computation of σ_S^2 and D_{MMSE} will be elaborated upon in Section 4.2.4 and Chapter 5, respectively.

4.2.4 Temporal Variance Estimation

As mentioned in Section 4.2.3, the variance component σ_S^2 is needed for each of the L frequency subbands in order to evaluate (4.21). As we do not have prior knowledge about the underlying distribution of S, we are forced to estimate σ_S^2.

Generally, in order to obtain an estimate of σ_S^2, we would average over multiple realizations, which is known as ensemble averaging. As only one realization is available, we turn to temporal averaging to estimate σ_S^2. Assuming that the stochastic process $\boldsymbol{S}$ is stationary - at least in the wide sense - temporal averaging tends to ensemble averaging [43, Sec. 17.5]. This suggests that a temporal average should be taken in wide sense stationary regions. A widely accepted assumption in speech processing applications is that speech is wide sense stationary in regions of $[10-30]$ [ms] [7, p. 45]. Hence, σ_S^2 should be estimated based on data captured in such duration. However, using such a small amount of data for an estimate introduces high variance in the estimate.

To combat the issue of high variance in the estimation of σ_S^2, averaging over a period longer than $[10-30]$ [ms] may be done. When increasing the estimation time period, the validity of the wide sense stationarity assumption decreases, however, the variance of the estimate is reduced. This method has proven useful in [9] and [10], in which signal statistics were computed in regions of 250 [ms] and roughly 400 [ms], respectively. Therefore, we consider temporal estimates computed upon durations exceeding 30 [ms].

At frame m, the variance σ_S^2 in a given subband is estimated as

$$\hat{\mu}_{S[k,m]} = \frac{1}{P}\sum_{n=0}^{P-1} s[k, m-n] \tag{4.22}$$

$$\hat{\sigma}^2_{S[k,m]} = \frac{1}{P-1}\sum_{n=0}^{P-1} \left(s[k, m-n] - \hat{\mu}_{S[k,m]}\right)^2, \tag{4.23}$$

where P is the number of frames over which the average is computed and $\hat{\mu}_{S[k,m]}$ is the sample mean. Note that a temporal estimate rely on realizations, hence, the lower-case letters.

The estimate $\hat{\sigma}^2_{S[k,m]}$ must be computed for each time-frequency tile. The fact that successive estimates mainly use the same data, motivates a recursive estimation scheme, in which the estimate at the previous frame is used for the current estimate. The recursive estimation scheme for (4.22) and (4.23) is given as [9, Sec. IV], [44, p. 101]

$$\hat{\mu}_{S[k,m]} = \beta\hat{\mu}_{S[k,m-1]} + (1-\beta)s[k,m] \tag{4.24}$$

$$\hat{\sigma}^2_{S[k,m]} = \beta\hat{\sigma}^2_{S[k,m-1]} + (1-\beta)\left(s[k,m] - \hat{\mu}_{S[k,m]}\right)^2, \tag{4.25}$$

where $\beta \in [0,1]$. In (4.22) and (4.23), all P terms are weighted equally, while in (4.24) and (4.25) the weighting of previous terms decays exponentially with time. For this reason, this update scheme is known as an exponentially weighted moving average [44, p. 101]. The use of such recursive estimation scheme reduces the computational complexity of (4.23) while the most recent terms are considered the most important. The value of β will be discussed in Section 6.1.3.

4.2.5 Complete Mutual Information Computation

In this section, we briefly summarize the computation of the measure MI. For each subband and each frame, $D_{\mathrm{MMSE}}(\boldsymbol{x}_m)_{k,m}$ is - according to (4.16) - the mean square error (MSE) of using $\hat{s}_{\mathrm{MMSE}}(\boldsymbol{x}_m)[k,m]$ instead of the true value $s[k,m]$. Next, according to (4.17), $D_{\mathrm{MMSE},k,m}$ is formed by averaging $D_{\mathrm{MMSE}}(\boldsymbol{x}_m)_{k,m}$ over all realizations of $\boldsymbol{X}$. As for $\hat{\sigma}^2_{S[k,m]}$, $D_{\mathrm{MMSE},k,m}$ is temporally estimated with the recursive update scheme presented in Section 4.2.4.

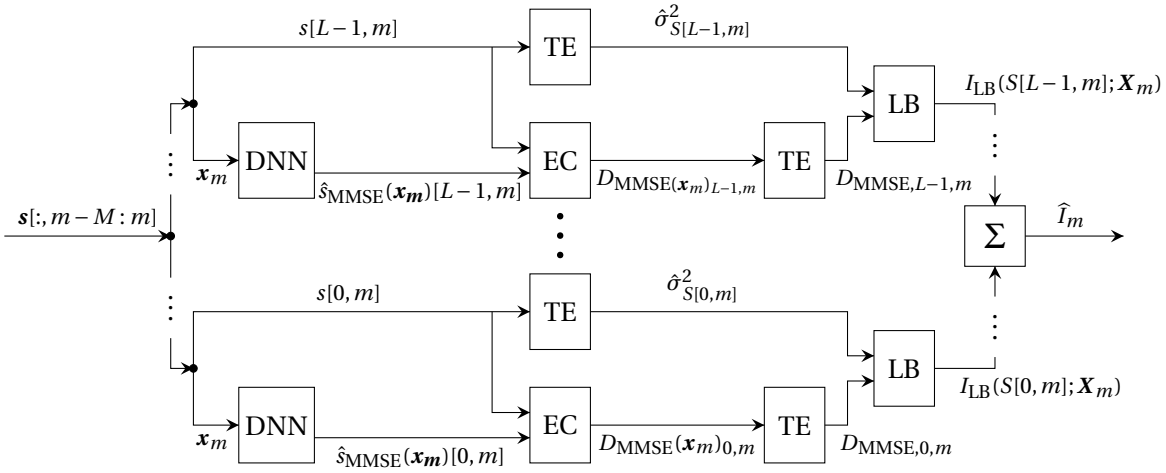


Figure 4.3. Illustration of per-frame mutual information estimate computation at frame m. The following abbreviations are used, TE: temporal average, EC: error computation, LB: lower bound on subband mutual information.

Finally, $I_{\mathrm{LB}}(S[k,m];\boldsymbol{X}_m)$ is computed according to (4.20) and summing over subbands yields the proposed estimate of mutual information $\widehat{I}_m$. The process of computing the per-frame estimate of the mutual information at frame m is illustrated in Figure 4.3. The data $\boldsymbol{s}[:, m-M:m]$ is partitioned into that data needed for each frequency subband. Afterwards, the aforementioned computations are carried out for each frequency subbands. At last, the estimates of mutual information for each frequency subband are summed to obtain the per-frame mutual information estimate $\widehat{I}_m$.

Deep Learning for MMSE Estimation 5

In the derivation of the upper bound on the conditional differential entropy in Section 4.2.1, the MMSE estimator $\hat{\boldsymbol{s}}_{\text{MMSE}}(\boldsymbol{x})$ of $\boldsymbol{S}$ given $\boldsymbol{x}$ emerged. In [9], a linear minimum mean square error (LMMSE) estimator is introduced to replace the MMSE estimator. As the mean square error (MSE) of the LMMSE estimator is at best equal to that of the MMSE estimator, replacing the MMSE estimator in (4.15) with a linear estimator loosens the upper bound on the conditional differential entropy $h(S|\boldsymbol{X})$ in (4.9). I.e., $D_{\text{MMSE}} \leq D_{\text{LMMSE}} \implies h_{\text{MMSE}}(S|\boldsymbol{X}) \leq h_{\text{LMMSE}}(S|\boldsymbol{X})$. In fact, any estimator can replace the MMSE estimator $\hat{s}_{\text{MMSE}}(\boldsymbol{x})$ in (4.15), but the MMSE estimator tightens the bound.

In this thesis, we deviate from the approach of linear estimation in [9], in that we try to realize the MMSE estimator. Specifically, we will formulate this estimator as a DNN, which has the advantage of being a non-linear estimator in which there is no need for analytic expressions. In Section 5.1 and Section 5.2, we present the architecture and training specifications of our proposed neural network, respectively.

5.1 Neural Network Architecture

In designing a neural network, perhaps the simplest architecture that comes to mind is a simple fully connected network. However, in order for the model to better adapt to the spectro-temporal dependencies in the two dimensional signal representation, convolutional layers may be introduced in the network. The use of convolutional layers is supported by [45], in which spectro-temporal patterns are linked to the auditory system. Hence, in this thesis, we opt to build a convolution neural network (CNN) as the estimator $\hat{s}_{\text{MMSE}}(\boldsymbol{x})$ in (4.15). Note that, as this network is trained iteratively, the true MMSE estimator will not be obtained, however, we still refer to our proposed estimator as the MMSE estimator.

In CNNs, the first layers are convolutional layers and pooling layers. The resulting outputs of the convolution layers (referred to as feature maps) are flattened and possibly multiple fully connected layers follows. This common CNN architecture is depicted in Figure 5.1 for one con-

volutional layer, one pooling layer, and one fully connected layer. The aim of the convolutional layers is to learn multidimensional correlation, while pooling layers, amongst other things, aim to subsample the feature maps produced by the convolutional layers. This in done in order to reduce the amount of parameters in the neural network, hence, reducing the training time. Often, multiple sets of a convolutional layer and a pooling layer are used in succession.

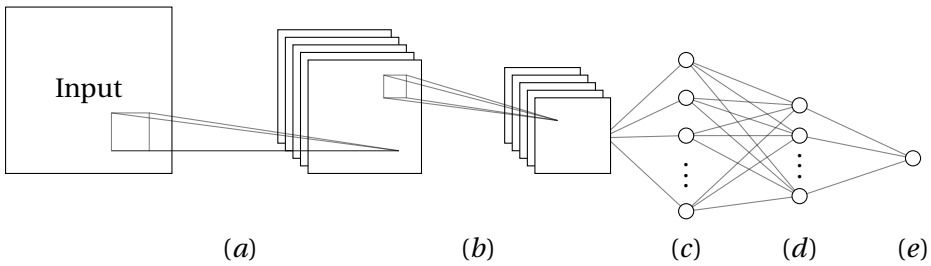


Figure 5.1. Illustration of common CNN architecture. (*a*): convolution operation, (*b*): pooling operation, (*c*): flattening operation, (*d*): fully connected layer, (*e*): output layer.

As mentioned in the introduction of this chapter, our aim is to realize a MMSE estimator as a neural network. To this end, the architecture of the proposed network has been chosen through an iterative design procedure with trail and error. In the following, we will provide an overview of the input feature representation of the network, the output and the proposed architecture.

The first five layers of the proposed network consists of convolutional layers each with (3×3) filters. The number of filters in the five convolutional layers are 20, 40, 80, 160 and 320. Increasing the number of filters with the depth is common practice, as this allows the network to pickup more complex structures in the data by combining the previous feature maps. In the model selection, we examined different number of convolutional layers and filter sizes. However, the aforementioned specifications showed the best performance in terms of the lowest validation error. We opt not to use pooling layers, as the input in itself is rather small. Pooling layers would rapidly reduce the dimensionality to a point in which another convolutional layer could not be used, hence, limiting the complexity in the feature maps. After the convolutional layers, the feature maps are flattened to a vector and followed by three fully connected layers with a decreasing number of neurons. At last, the output layer has a single neuron, as the aim is to estimate a single value.

Between each convolutional layer and each fully connected layer, we include a batch normalization layer. The idea of batch normalization is that the layers in neural networks train faster on standardized data, hence, batch normalization seeks to standardize the input data to a layer.

Besides omitting the pooling layers, another method to retain dimensionality in the convolutional layers in neural networks is zero-padding, which is the process of appending a set amount of bins with the value zero around the border of the input to the convolutional layers as depicted in Figure 5.2. By choosing the amount of zero-padding - both vertically and horizontally - to be one less than the filter size, the natural reduction in dimensionality created by the convolution operation is offset [46, p. 351]. In preliminary experiments, we found that

zero-padding did not improve performance, hence, we opt not to use zero-padding.

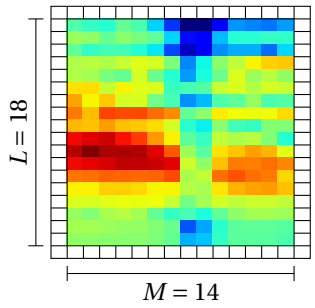
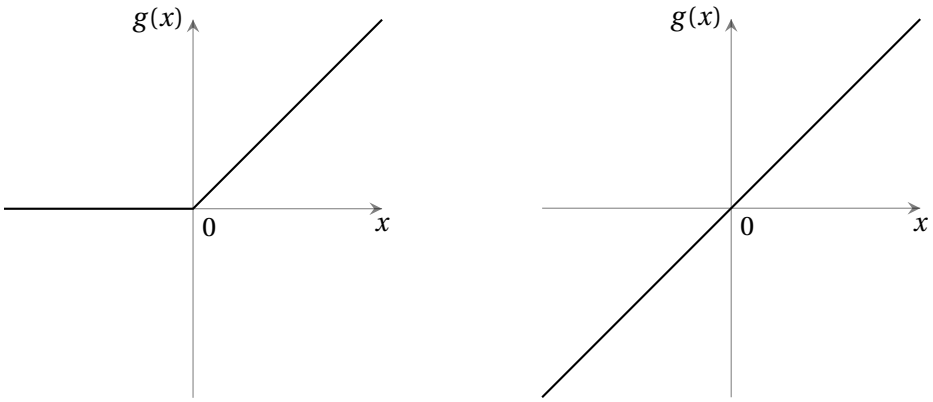


Figure 5.2. Input image for the CNN with one tile of zero-padding both horizontally and vertically. White tiles represent zero-padding.

To introduce non-linearity in neural networks, activation functions are used. In our proposed CNN, for all layers, except the output layer, the rectified linear unit (ReLU) activation function given as $g(x) = \max\{0, x\}$ is used. This is common practice, as the ReLU activation function is computationally efficient while suppressing problems of vanishing and exploding gradients, which other activation functions suffer from [46, Sec. 6.3]. As the aim of the CNN is to predict a continuous value, the estimation problem is treated as regression which entails using the linear activation function given as $g(x) = x$ for the output layer. The activation functions used in the CNN are depicted in Figure 5.3.



(a) ReLU activation function given as $g(x) = \max\{0, x\}$. **(b)** Linear activation function given as $g(x) = x$.

Figure 5.3. Activation functions used in the CNN.

In using neural networks, an important aspect is how well the model generalizes to unseen data. A technique to help a neural network generalize better is dropout which, during training, forces a proportion of the inputs to a layer to be omitted from the computations, and in turn aids in avoiding overfitting [47]. A similar approach is used for convolutional layers, namely spatial dropout, in which entire feature maps are omitted from training. We found that, using spatial dropout with a probability 0.2 of dropping a feature map in each convolutional layer and dropout with a probability 0.2 of dropping a neuron in the fully connected layers, yielded the best generalization performance.

Since an estimator for each of the L subbands is needed, L equivalent networks are trained on

the same input data, however, with different targets. The input feature maps to each of the CNNs is previous spectro-temporal tiles, specifically, across the previous M frames and all L subbands. Hence, the input to the neural networks at frame m is

$$\boldsymbol{x}'_m \triangleq \Big[\boldsymbol{s}[:, m-M] \quad \boldsymbol{s}[:, m-(M-1)] \quad \cdots \quad \boldsymbol{s}[:, m-1]\Big], \tag{5.1}$$

where $\boldsymbol{s}$ is the filter bank processed short-time magnitude spectrum from (4.1) and $M > 0$ denotes the number of previous frames to include in the context. Note that (5.1) is simply (4.2) organized as a matrix. The target for the kth neural network at frame m is $s[k, m]$. The L estimators result in a parallel neural network structure which is illustrated in Figure 5.4.

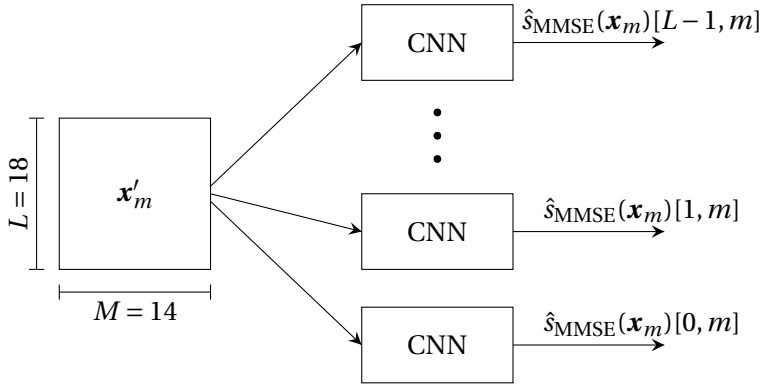


Figure 5.4. Illustration of the parallel neural network structure, where each CNN block constitute a neural network.

The final architecture of the proposed CNNs is illustrated in Figure 5.5 and recapped in Table 5.1. Note that the horizontal lines in Table 5.1 partition the network into overall sections, but is solely for visualization purposes.

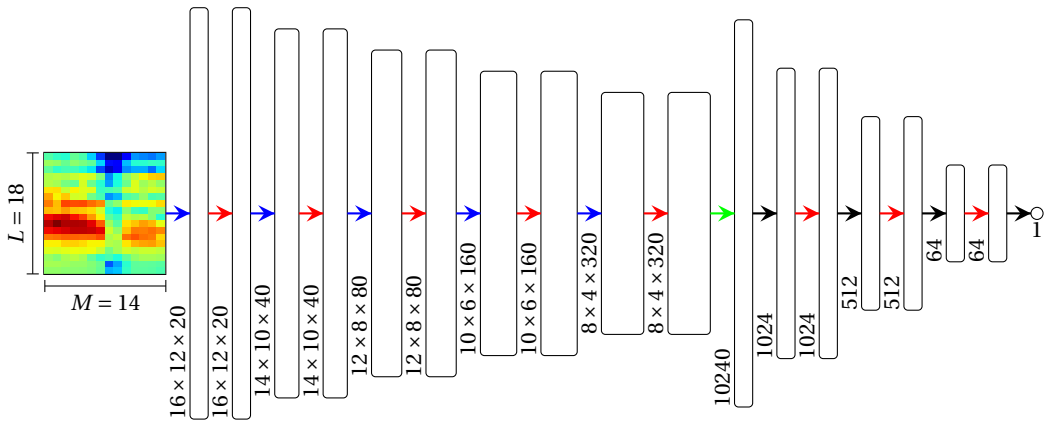


Figure 5.5. Neural network architecture. Arrows represent different layers, while the boxes represent the output of the layers. Blue arrows represent convolutional layers, red arrows represent batch normalization, the green arrow represents a flattening operation and black arrows represent fully connected layers.

Table 5.1. Neural network architecture. Conv2D is a two-dimensional convolutional layer with Conv Filters $X \times (Y \times Z)$ meaning X filters of size $Y \times Z$ (height × width). Horizontal lines partition the network into overall sections, but is solely for visualization purposes.

Layer	Conv Filters	Activation	Output shape	(Spatial) Dropout
Input	-	-	18×14	-
Conv2D	$20 \times (3 \times 3)$	ReLU	$16 \times 12 \times 20$	0.2
Batch Normalization	-	-	$16 \times 12 \times 20$	-
Conv2D	$40 \times (3 \times 3)$	ReLU	$14 \times 10 \times 40$	0.2
Batch Normalization	-	-	$14 \times 10 \times 40$	-
Conv2D	$80 \times (3 \times 3)$	ReLU	$12 \times 8 \times 80$	0.2
Batch Normalization	-	-	$12 \times 8 \times 80$	-
Conv2D	$160 \times (3 \times 3)$	ReLU	$10 \times 6 \times 160$	0.2
Batch Normalization	-	-	$10 \times 6 \times 160$	-
Conv2D	$320 \times (3 \times 3)$	ReLU	$8 \times 4 \times 320$	0.2
Batch Normalization	-	-	$8 \times 4 \times 320$	-
Flatten	-	-	10240	-
Fully Connected	-	ReLU	1024	0.2
Batch Normalization	-	-	1024	-
Fully Connected	-	ReLU	512	0.2
Batch Normalization	-	-	512	-
Fully Connected	-	ReLU	64	0.2
Batch Normalization	-	-	64	-
Fully Connected/Output	-	Linear	1	0.2

5.2 Neural Network Training

In order to assess our proposed measure of mutual information, a listening test is conducted, which will described in detail in Section 6.1. For this listening test, the Dantale II speech corpus [48], [49] is used, hence the CNNs are to used as estimators on this speech corpus. However, as this corpus is rather small in the context of neural network training, the TIMIT speech corpus is used for training of the neural networks. A detailed description of the Dantale II speech corpus is presented in Section 6.1.2.

5.2.1 Input Data Generation

The TIMIT speech corpus contains recordings of 630 speakers of eight major dialects of American English. For each speaker the corpus contains 10 phonetically rich sentences resulting in a total of 6300 sentences, sampled at a frequency of $f_s = 16$ [kHz] [38]. The sentences have an average duration of 3.08 [s]. Due to memory limitations on the computer on which the neural networks are trained, a subset of the corpus is used for training. For training, $N_{\text{train}} = 2500$ sentences are randomly chosen while $N_{\text{test}} = 750$ different sentences are randomly chosen for validation. We allocate this amount of sentence for training and validation, as this is a fairly common partitioning of training and validation data [46, p. 121]. The computer used for training the neural networks is equipped with a NVIDIA GeForce GTX 1060 6GB graphics processing

unit.

Silent regions of speech contain no information about future regions, hence, does not contribute to the possibly predictive nature of speech [9, p. 432]. Therefore, in the training of the neural networks, silent regions are excluded. This is done with a simple energy based voice activity detector (VAD). Before describing the VAD, we consider the following definition of energy.

Definition 5.1 (Energy)
Let $\boldsymbol{x} \in \mathbb{R}^N$. Then, the energy of $\boldsymbol{x}$ is defined as

$$E = \|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=0}^{N-1} x^2[i]}\,. \qquad \blacktriangle$$

The VAD measures the energy of frames in the time domain. These frames are obtained by sliding a Hann window of length 256 with both hop and overlap of 128 samples, over the signals. Then, the VAD identifies frames where the energy is less than some threshold Δ_E [dB] below the frame with the maximum energy. These identified frames are then excluded. We examined VAD outputs for $\Delta_E = 40$ [dB], as proposed in [10], and deemed the results acceptable, yet no fine-tuning was conducted. Note that this VAD frame exclusion is applied on each sentence individually, as a frame containing large energy could potentially produce an unsuitable threshold for some sentences, if all sentences were subject to the VAD collectively.

After silent regions are removed, the sentences are peak normalized (see Definition 2.1), scaling each sentence into the range $[-1, 1]$. This is done to ensure likeliness of the data which is to be passed as input to the CNNs. However, note that this normalization is essentially nullified, as we perform input standardization of the data to the neural networks, which will be presented in the proceeding.

Next, the training sentences are concatenated to a single signal and the STFT is applied to obtain a spectro-temporal representation. For the STFT, a Hann analysis window of length $N_{\text{frame}} = 256$ samples is chosen with a frame shift of $R = \frac{N_{\text{frame}}}{2} = 128$ samples. These parameters are chosen to give a fine effective temporal resolution of 8 [ms]. Furthermore, using a frame shift equal to half the length of the analysis window, is common practice in speech processing applications [7, p. 35]. A DFT order of $N_{\text{DFT}} = 2N_{\text{frame}} = 512$ is used as this ensures that all subbands have associated frequencies. The same processing is applied to the validation data.

After obtaining the spectro-temporal representation of the signals, a one-third octave filter bank followed by a decibel transformation is applied to the magnitude spectra, as in (4.1). This filter bank is inspired by that used in [9]. However, in [9], signals are resampled to a sampling frequency of 10 [kHz] and 15 filters are used with the lowest center frequency at 150 [Hz]. As we use signals sampled at a frequency of 16 [kHz], a larger number of filters is required to cover the frequencies up to half the sampling frequency. We use $L = 18$ filters, with the lowest center

frequency at 140 [Hz]. By using these filter bank specifications, we obtain a filter bank covering up to half the sampling frequency. We changed the lowest center frequency in order to contain the upper band frequency of the highest frequency filter to under half the sampling frequency.

When training neural networks, it is common practice to either normalize or standardize the input data and targets [50, Sec. 8.1-8.2]. This is done for multiple reasons. Firstly, having all input features on the same scale increases convergence speed. Secondly, typically, the training process becomes more stable. Thirdly, such procedure makes the networks invariant to arbitrary scaling of the input data, which is important as we are to use the networks on a different speech corpus than the one on which it is trained on. We choose to standardize the data.

Next comes the question of whether the data should be standardized according to a global mean and standard deviation (referred to as full-band standardization) or each of the L subbands should be standardized according respective means and standard deviations (referred to as per-band standardization). We choose to apply per-band standardization for a couple of reasons: As low frequency subbands contain more energy than high frequency subbands, the training process of adjusting the weights associated with the low frequency subbands might dominate the training process under full-band standardization, hence disregarding the weights associated with the high frequency subbands. This issue is circumvented with per-band standardization as the dynamic range for each subband is approximately the same. Furthermore, different speech corpora are likely recorded using different recording equipment and the frequency response of microphones varies, leading to different weightings of frequencies across microphones. Per-band standardization nullifies this issue. At last, we conducted experiments suggesting a slight superiority of per-band standardization over full-band standardization.

The per-band standardization is carried out as [50, (8.2)]

$$S_{\text{STD}}[k,m] = \frac{S[k,m] - \hat{\mu}_k}{\hat{\sigma}_k}, \quad k = 0,1,\ldots,L-1, \quad \forall m, \tag{5.2}$$

where $S[k,m]$ is the non-standardized filter bank output, and $\hat{\mu}_k$ and $\hat{\sigma}_k$ are the sample mean and sample standard deviation of the kth subband, estimated from all active frames in the training data, respectively. The validation data is likewise per-band standardized. It should be mentioned that when the neural networks are to be used as estimators on another corpus, the data should be standardized according to the statistics of the corpus in question.

After standardization, the data is restructured as input to the neural networks. At frame m, this is done according to (5.1). Note that no input matrix is formed prior to $m = M$. We set $M = 14$ as this corresponds to approximately 112 [ms] with the given STFT parameters. According to [17], both vowel and consonant sounds are generally captured in this duration.

5.2.2 Training Procedure

When training a neural network, the loss function and the optimization procedure of the weights must be specified. As we aim to build a MMSE estimator, we choose the MSE as loss function.

For the optimization procedure of the weights, the optimization method stochastic gradient descent and multiple alterations of this algorithm is often used. We choose to use the Adam optimizer, as this optimizer is widely used and compares favorably to other stochastic optimization methods [51]. No experiments with the hyperparameters of the Adam algorithm have been conducted and the parameters follows those provided in [51, Alg. 1].

Neural networks are trained in batches, i.e., a number of inputs is propagated through the network, then the weights are updated and the next batch of inputs is propagated through the network. When all inputs have been propagated through the network in batches, this is called an epoch. Choosing the number of inputs in a batch (referred to as batch size) is a trade-off between accuracy and speed. With a large batch size, large amount of data is used to update the weight but updates does not happen as often, hence, accurate but slow. With a small batch size, the weights are updated often based on small amounts of data, hence, less accurate but fast. We use a batch size of $N_{\text{batch}} = 32$, as this is a good default value [52, p. 9], and since we, during model selection, found this to be a good compromise between accuracy and convergence speed.

As discussed in Section 5.1, the ability for a neural network to generalize well is important. Hence, we need to stop the training process at a given point, which is done in the following manner. The loss evaluated on the validation data (referred to as validation loss) is measured at the end of each epoch and as the training proceeds, we expect the model to become more accurate, i.e., for the loss to decrease. However, at some point the model might become overfitted to the training data and the validation loss will increase. If the validation loss has not reach a new minimum for $N_{\text{patience}} = 10$ epochs, the training is stopped and the model weights resulting in the lowest validation loss is saved to be used for future predictions [46, Sec. 7.8].

The parameters presented throughout this section are summarized in Table 5.2

Table 5.2. Parameters presented throughout Section 5.2.

Parameter	Value	Unit
M	14	frames
L	18	frequency subbands
f_s	16000	Hz
Δ_E	40	dB
N_{frame}	256	samples
N_{DFT}	512	samples
R	128	samples
N_{train}	2500	sentences
N_{test}	750	sentences
N_{batch}	32	inputs
N_{patience}	10	epochs

Since our CNNs operate on standardized data, predictions made by the CNNs are likewise standardized. In order to convert predictions to their correct scale, they are multiplied by the es-

timated standard deviation of the corresponding subband of the dataset and afterwards the estimated mean of the corresponding subband is added. This is the inverse procedure of the per-band standardization presented in (5.2).

With the CNNs trained on the TIMIT speech corpus, a set of weights is obtained for each model. These weights might not be near as optimal for the Dantale II speech corpus, as the spoken language, the speakers, the noise, and the recording equipment differs for the two corpora. However, we do not make any effort to adjust the weights to better fit the Dantale II speech corpus.

Experiments 6

In order to evaluate our proposed measure MI which, as mentioned in Chapter 1, is a non-intrusive perceptually relevant estimate of the information theoretical quantity mutual information for discrete time frames of speech signals, a listening test is conducted. This chapter aims to introduce and describe the steps necessary to conduct this experiment. In Section 6.1, the framework of the listening test is presented followed by the results in Section 6.2. In Section 6.3, an additional experiment is presented, which examines the hypothesis that important speech frames are double protected from acoustic noise (see research sub-question 2 in Chapter 1).

6.1 Experimental Framework

For the conducted listening test, we replaced parts of the speech signals with noise and presented these to listeners. This experimental paradigm is reminiscent of that in [17], [27]. However, we replaced single frames of the speech signals, whereas in [17], [27] they replaced segments of successive frames corresponding to 80 [ms] and 112 [ms]. Our choice of replacing single frames is due to the fact that our proposed measure identifies single frames which are important for speech intelligibility.

6.1.1 Listeners

In the listening test, 24 persons (9 female, 15 male) participated. Participants ranged in age from 24 to 65 years with a mean of 35.71 years and a standard deviation of 13.99 years. All participants were native speakers of the Danish language and reported no hearing loss. Participants were asked to rate their prior knowledge of the Dantale II speech corpus on a scale from 1 to 10 with 1 meaning "I do not know what the Dantale II speech corpus is", 5 meaning "In the past couple of years, I have participated in one or more listening tests involving the Dantale II speech corpus", and 10 meaning "Prior to the listening test, I was capable of recollecting most of the words in the Dantale II speech corpus". Of the 24 participants, 21 rated their prior knowledge of the Dantale II speech corpus as 1 while the remaining three participants reported the rates 3, 3, and 7.

6.1.2 Stimuli

For the listening test, the Dantale II speech corpus was used [48], [49]. The sentences in this corpus are Danish and are all spoken by a single female speaker. The corpus consists of 16 lists, from which we used 15, each containing 10 sentences, resulting in a total of 150 sentences. Each sentence is comprised of five words structured in the following manner: name, verb, numeral, adjective and object. Each word is chosen from a set of 10 candidate words, containing one or two syllables, in the given word class. In the verb word class, half the words is past tense while the other half is present tense. Note that the 10 candidate words in each word class are the same for all sentences. The resulting sentences are syntactically correct, however, not necessarily with any meaningful context, e.g., "Per gets twenty old houses" (translated from Danish). The fact that the sentences are without meaningful context, means that listeners are not able to guess words based on previous context.

6.1.3 Signal Processing

In our listening test, the measure of MI (presented in Section 4.2) was examined. As [17], [27] report that the measures of CSE and INT predict speech intelligibility, these measures were included in our listening test for comparison.

In order to compute MI, INT, and CSE for each of the 150 Dantale II sentences, the following processing was applied to each sentence. First, the sentence was resampled from 44100 [Hz] to a sampling frequency of $f_s = 16000$ [Hz] with a polyphase filter [53] with a $\frac{up}{down} = \frac{160}{441}$ conversion rate. Next, as the Dantale II speech corpus is stored as 16-bit integer values, the signal was converted to floating point values as the CNN MMSE estimators are build to operate on floating point values.

After obtaining the resampled sentences represented as floating point values, each sentence was peak normalized according to Definition 2.1 and zero-padded from the front with $(M+1)R = 1920$ samples. This zero-padding ensures that the STFT domain representation will have an additional $M = 14$ frames in the front, which in turn makes it possible to compute MI for all original frames, as the CNN MMSE estimators perform predictions based on the M most recent frames.

In the following, we present how each measure was computed for each sentence.

INT

The measure of INT was computed in the time domain. The time domain signal was subject to the same windowing process occurring in the STFT, with a Hann window of length $N_{\text{frame}} = 256$ samples and an overlap of $R = 128$ samples. The measure INT was obtained by using Definition 4.1 for each frame.

MI

The proposed measure of MI was computed as described in Section 4.2, with the STFT and filter bank parameters specified in Section 5.2. The parameter β associated with the recursive update scheme presented in Section 4.2.4 is an exponentially weighted moving average filter, which means that the time constant describes when the impulse response has decayed by a factor $\frac{1}{e}$. The time constant τ and the parameter β from (4.25) are related as

$$\tau = \frac{-1}{\ln(\beta)} \frac{R}{f_s} \text{ [s]}. \tag{6.1}$$

Using large values for β, results in the measure being slowly varying over time while small values for β allows for more rapid changes. We settled on $\beta = 0.948$ which corresponds to a time constant of $\tau \approx 0.15$ [s]. In (6.1), the first fraction gives the number of needed updates for the impulse response to decay by a factor $\frac{1}{e}$. As successive frames are R samples offset, we multiply the number of updates with R and multiply with the sampling period $\frac{1}{f_s}$ to obtain the result in seconds.

CSE

To compute the measure of CSE, the STFT and filter bank parameters of the measure MI was used, however, the filter bank in (4.1) was computed as

$$S'[i,m] = \sum_{k \in C_i} \left| \tilde{S}[k,m] \right|, \quad i = 0, 1, \ldots, L-1,$$

and the computation of CSE was carried out as in line 6 in Algorithm 1 with $J = 1$. Setting $J = 1$ in line 6 in Algorithm 1 omits the boxcar summation. This summation essentially acts as a low-pass filter which temporally smooths the measure of CSE, hence, smearing how alike two adjacent frame are. Therefore, $J = 1$ was chosen. Also note that the STFT parameters and filter bank used to measure CSE differs from those originally proposed for CSE in [17].

Now that we have ppresented how the measures MI, INT, and CSE were computed for each sentence, we describe how frames were replaced by noise. As in Section 5.2, a VAD was used to identify frames with low energy. For the VAD, we maintained the threshold $\Delta_E = 40$ [dB] from Section 5.2. Frames surpassing this threshold are referred to as active frames.

In [17], [27], three conditions were created for each measure. Based on the value of the measure, segments of successive frames were ordered in: 1) ascending order, 2) descending order and 3) ascending absolute difference from the median. For each ordering, the first predetermined number of segments were replaced with noise. For our listening test, we chose to omit order 3 as this allows for more repetitions of order 1 and order 2. We refer to these methods of replacing frames as MI, INT, or CSE suffixed by -LOW (order 1) or -HIGH (order 2). Whether order 1 or order 2 was used, is referred to as replacement category.

We performed preliminary experiments which indicated that replacing 50% of active frames, barely influenced speech intelligibility for CSE. Replacing 75% of active frames did more heavily influence speech intelligibility, hence, both sentences with 50% and 75% of active frames

replaced, were included in the experiment. Whether 50% or 75% of active frames were replaced is referred to as replacement percentage and denoted $r_{\%}$. Including a control condition with no frames replaced, 13 conditions were included in the experiment. As an example, for the condition MI-LOW-50, the 50% of active frames with lowest MI were replaced.

The noise used to replace frames in the sentences was speech shaped noise (SSN). This type of noise is created by finding a set of coefficients for an all-pole filter through linear predictive coding. As described in [43, Sec. 18.7], using a 12th-order filter is common practice at 8 [kHz], hence, we used a 24th-order filter as we operate at 16 [kHz]. The coefficients were computed on the basis of 10 Dantale II sentences, as this number is the default for the listening test software, which will be presented in Section 6.1.4.

With the coefficients for the all-pole filter found, the SSN for one frame was generated as follows. A unit-variance white Gaussian noise sequence of length $10 \cdot N_{\text{frame}}$ was generated. This noise sequence was passed through the all-pole filter and all but the last N_{frame} samples were discarded. This was done to allow the filter to reach steady state. The resulting noise sequence of length N_{frame} was multiplied with a Hann window of equal length, as the frame which was to be replaced was extracted from the original signal with a Hann window. Overlapping Hann windows have the benefit of smoothing transitions between frames.

Before proceeding, we consider the power of a vector $\boldsymbol{x} \in \mathbb{R}^N$.

Definition 6.1 (Power)
Let $\boldsymbol{x} \in \mathbb{R}^N$. Then, the power of $\boldsymbol{x}$ is defined as

$$P = \frac{1}{N}\|\boldsymbol{x}\|_2 = \frac{1}{N}\sqrt{\sum_{i=0}^{N-1} x^2[i]}\,. \qquad \blacktriangle$$

Let P_N and P_S denote the power of the SSN sequence and the frame which was to be replaced, respectively. Multiplying each sample of the SSN sequence with $\sqrt{\frac{P_S}{P_N}}$ results in the SSN sequence having power equal to the frame which was to be replaced. The SSN sequence was then substituted into the signal. The keen listener might have noticed a rasping sound in successive SSN frames, which was caused by the fixations of power levels of the frames. The power level of frames in a SSN sequence is generally varying and adjusting each frame to have a specific power level introduced this artefact.

The reason for using SSN as opposed to white Gaussian noise is that the latter brings discomfort to humans, possibly exhausting participants. SSN has, as the name suggests, the same distribution of frequencies as speech and results in a more pleasant listening experience.

After $r_{\%}$ of active frames were replaced with SSN, the sentence was compiled of the overlapping frame. The signal was resampled back to a sampling frequency of 44100 [Hz] with a polyphase filter with a $\frac{up}{down} = \frac{441}{160}$ conversion rate. The sentence was scaled back to the original dynamic range and converted to 16-bit integer values.

6.1.4 Listening Test Procedure

The stimuli was presented for the participants through MATLAB software provided by Oticon A/S. When executing the software, a graphical user interface (GUI) was presented for the participant [13, p. 11]. This user-operated test has the advantage, compared to traditional listening tests where subjects orally repeats the perceived words, of avoiding the need for an experimenter to be present, while providing similar results [54].

The GUI presented a column for each of the five word classes. Each column contains the 10 candidate words for the given word class. The GUI is depicted in Figure 6.1. The GUI is equipped with a button labeled START to start the listening test. When started, the participant was presented with a sentence randomly subject to one of the conditions presented in Section 6.1.3. Then, on the GUI, the participant marked which words they heard. By clicking the button labeled NÆSTE (next in Danish), the next sentence was presented for the participant, unless the target number of sentences had been presented, then the listening test terminated.

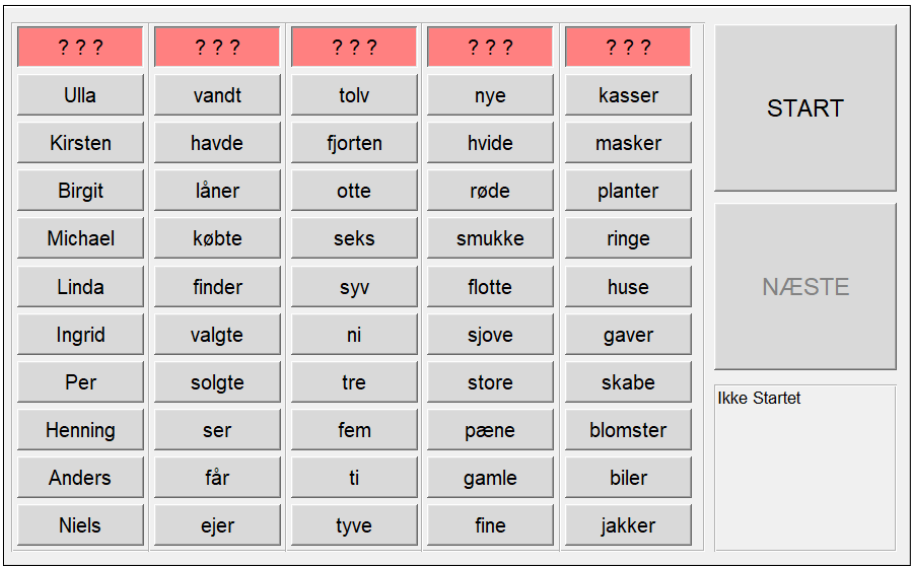


Figure 6.1. GUI used in the listening experiment [13, Fig. 8].

The stimuli was presented for the participants at a sample rate of 20 [kHz] at a comfortable volume (adjusted by the author but fixed across the participants, except for two participants who requested a slight increase in volume) with the equipment presented in Table 6.1. Optimally, a listening test would be conducted in a sound-proof chamber. Also, the volume would be adjusted to a specific value for the sake of reproducibility, e.g., in [27], the overall RMS sound pressure level of the speech corpus was adjusted to 57 [dB]. The sound pressure level measured in decibels entrails using a reference sound pressure in Definition 2.3 of $p_0 = 20$ [μPa] [28, Sec 1.9]. However, as described in Chapter 1, due to inaccessibility of laboratories at the time of need, this has not been possible.

Before the experiment, each participant underwent a training session, in which they were presented with 26 sentences in order to become somewhat familiar with the framework, condi-

Table 6.1. Equipment used in the listening experiment.

Equipment	Model	Settings
Sound card	Focusrite Scarlett Solo 2nd Generation	Direct Monitor: OFF
Headphones	AKG K 240 DF	-

tions and GUI. The training session lasted 6 minutes and 53 seconds on average with a standard deviation of 47 seconds. The training session was comprised of two control sentences and four sentences from each of the other conditions with $r_{\%} = 50$. The reason for not using $r_{\%} = 75$, was to not overwhelm the participants. If no training session was conducted, participants would presumably get much better at classifying words throughout the experiment. The training session aims to diminish this learning bias, however, such learning bias is never complete eliminated.

After completing the training session, the experiment was conducted. Each participant was presented with 182 sentences in such a manner, that each condition was presented 14 times. The experiment lasted 41 minutes and 11 seconds on average with a standard deviation of 5 minutes and 43 seconds. The order in which the sentences were presented to the participants was randomized. The participants were instructed that halfway through the experiment (after 91 sentences), they were allowed to take a short break. This break was recommended in order to avoid listening fatigue, which could influence results in the latter part of the experiment. After the break, the participants were presented with the remaining 91 sentences, after which the experiment concluded.

6.2 Experimental Results

In this section, results from the listening test are presented. First, in Section 6.2.1, the proportion of correctly identified words for each condition is presented with the aim of answering the hypothesis if MI is able to identify frames which are important for speech intelligibility. Afterwards, in Section 6.2.2, correlations between the measures of MI, INT, and CSE are examined. In Section 6.2.3, the distribution of misclassifications over the five word classes is examined, and finally, in Section 6.2.4, the extend to which replaced frames are grouped is examined for the different conditions. The results presented in this section will be elaborated upon and discussed in Chapter 7.

Before proceeding to the presentation of the results, we present an example of the waveform of a Dantale II sentence along with MI, INT, and CSE measured for this sentence. This example in Figure 6.2, serves the purpose of visualizing typical behavior for the measures. Note that in the bottom pane in Figure 6.2, the measures are individually normalized to the range $[0,1]$ as we are interested in the behavior of the measure rather than the magnitude. From Figure 6.2, we see that the measures MI, INT, and CSE differ quite a lot. First of all, we see that CSE is characterized by many spikes. These spikes generally occur at the start and end of phonemes,

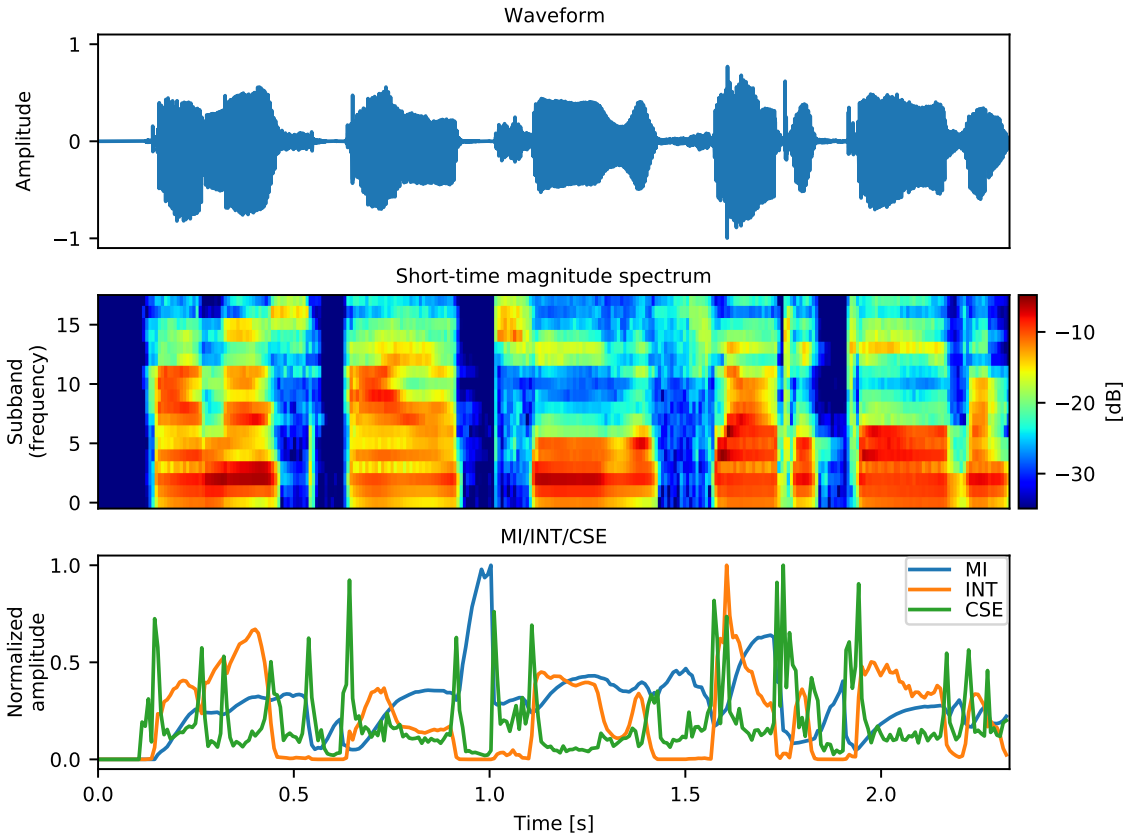


Figure 6.2. The top pane illustrates the waveform of a sample sentence of the Dantale II speech corpus. The middle pane illustrates the short-time magnitude spectrum - processed by a one-third octave filter bank with $L = 18$ filters - of the sentence. The bottom pane illustrates the measures MI, INT, and CSE measures across the sentence. Note that the measures have individually been normalized to the range $[0, 1]$.

which make sense as the spectral difference is large in these circumstances. Furthermore, we see that the measures of MI and INT are generally more slowly varying compared to CSE.

To support Figure 6.2, in Figure 6.3 we present waveforms and filter bank processed short-time magnitude spectra of the same sentence from the Dantale II speech corpus, with frames replaced with SSN according to the conditions MI-LOW-50, INT-HIGH-50, and CSE-HIGH-50.

From Figure 6.3, is seems that frames replaced according to the condition CSE-HIGH-50, tend to more spread out across the sentence, as opposed to the conditions MI-LOW-50 and INT-HIGH-50, where it seems that contiguous frames are replaced.

6.2.1 Proportion Correctly Classified Words

In this section, results from the listening test are presented. Each of the 24 participants was presented with 14 repetitions of the 13 conditions, resulting in a total of 182 sentences. The speech intelligibility score (SIS) was then computed for each participant as the percentage of correctly classified words in the given condition. These SISs are then averaged over all 24 participants and the results are depicted in Figure 6.4a. Conditions (disregarding replacement percentage)

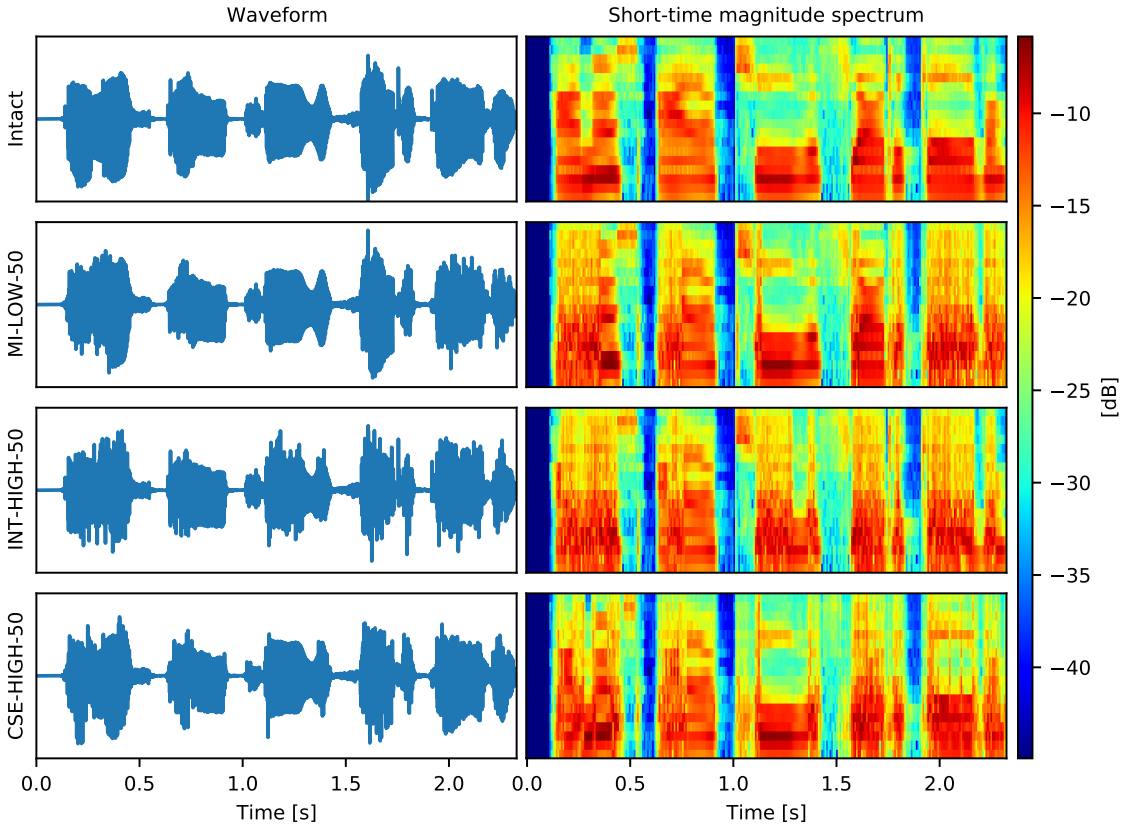
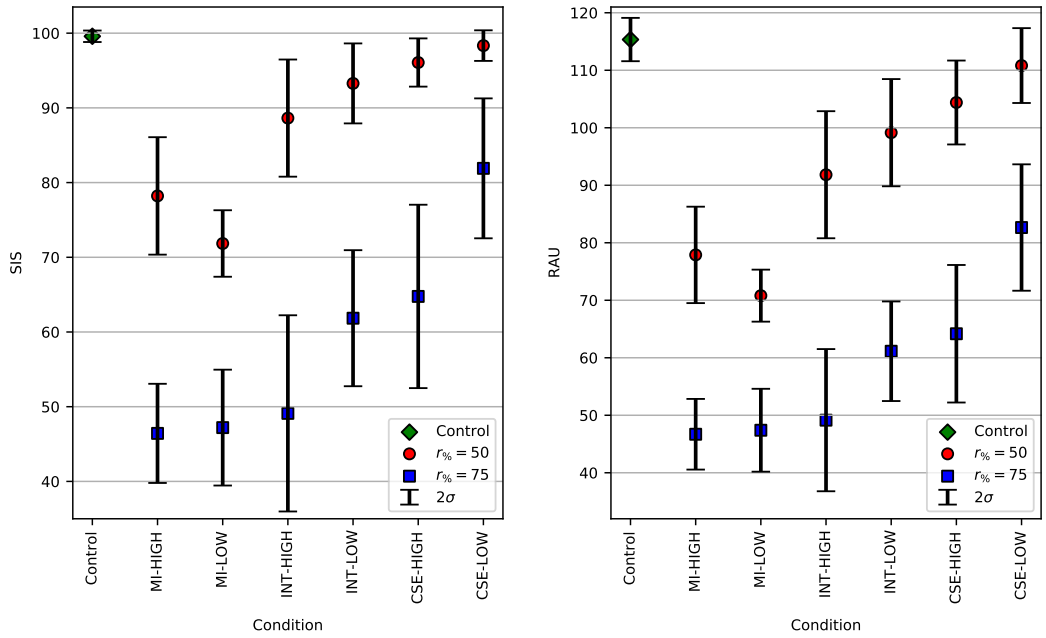


Figure 6.3. The figure illustrates the same sample sentence of the Dantale II speech corpus as in Figure 6.2. The left column depicts the waveforms in the time domain while the right column depicts the short-time magnitude spectra processed by a 18-filter one-third octave filter bank. The top row illustrates the intact sentence, i.e., where no frames have been replaced. From top to bottom, the last three rows present the sentence with frames replaced according to the conditions MI-LOW-50, INT-HIGH-50 and CSE-HIGH-50, respectively.

are represented on the horizontal axis while the SIS is represented on the vertical axis. The green diamond represents the control condition with no frames replaced, the red circles represent a replacement percentage of 50 and the blue squares represent a replacement percentage of 75. The error bars represent ±1 standard deviation of the mean SIS of the participants.

First, a set of key observations from the results in Figure 6.4 is presented. Afterwards, these observations are supported by analysis of variance (ANOVA).

As expected, sentences are generally more intelligible when $r_{\%} = 50$ compared to when $r_{\%} = 75$. In line with the results of [17], [27], the replacement category HIGH generally resulted in larger degradation of speech intelligibility compared to the replacement category LOW for both INT and CSE. This trend was apparent for both replacement percentages. In contrast to the measures of INT and CSE, we expect larger degradation in speech intelligibility for MI-LOW than for MI-HIGH. This is in fact the case for $r_{\%} = 50$, however, for $r_{\%} = 75$ the SISs for MI-LOW and MI-HIGH are almost identical with MI-HIGH showing slightly larger degradation. On this note, is it worth mentioning that the measures of MI and CSE are more valid for $r_{\%} = 50$ than for $r_{\%} = 75$. This is because these measures require knowledge of past frames. The measure



(a) Scores depicted as SISs.

(b) Scores depicted as RAUs.

Figure 6.4. Measured SISs and RAUs. On the horizontal axis conditions are represented. The markers represent the average score across all 24 participants and the error bars represent ±1 standard deviation of the mean. The green diamond represents the control condition, the red circles represent a replacement percentage of 50 and the blue squares represent a replacement percentage of 75.

of CSE requires knowledge of the previous frame while the measure of MI requires knowledge of the previous $M = 14$ frames. As this past knowledge might not be presented for the participants, as some of these frames might have been replaced by noise, the participants are not necessarily presented with the same information as the measures are computed upon. This aspect becomes more severe with an increased replacement percentage.

Across replacement percentages and replacement categories, a general tendency arises in that SISs for MI are lower than those of INT, which in turn are lower than those of CSE. In [17], [27], segments of successive frames are replaced and neighboring segments are not eligible for replacement. In our experiment no such ineligibility for replacement is considered, meaning that shorter or longer consecutive segments of speech might be replaced. One can certainly imagine that, whether many small segments or a few large segments are replaced in a sentence, generally affects the SIS of the measure. The grouping of replaced frames for the conditions are further examined in Section 6.2.4.

In this thesis, we attempt to identify the frames which are most important for speech intelligibility. We hypothesize that in the condition MI-LOW, the frames which are most important for speech intelligibility are replaced with noise, hence this condition should have a SIS lower than that of INT-HIGH and CSE-HIGH. As evident from Figure 6.4a, this is in fact the case for both replacement percentages. By hypothesizing that the condition MI-LOW contains the frames

most important for speech intelligibility, comes the implicit hypothesis that the condition MI-HIGH contains the frames least important for speech intelligibility. Hence, the condition MI-HIGH should in this case have a SIS higher than that of INT-LOW and CSE-LOW. As seen in Figure 6.4a, this is not the case for neither of the two replacement percentages.

Next, a set of ANOVAs are presented to support the aforementioned observations. However, in performing an ANOVA, data within each group is assumed to be a realization of a Gaussian distribution and the variance within each group is assumed equal [35, Sec. 7.5.1]. These assumptions are often invalidated for proportion score such as measured speech intelligibility, especially at the upper and lower end of the scale [55, p. 455]. To combat these issues, the authors of [55] proposed the rationalized arcsine transform, which produces the rationalized arcsine unit (RAU) as an alternative to the proportion score.

Definition 6.2 (Rationalized Arcsine Unit)
Let N_C and N_T denote the number of correct classifications and the total number of classifications, respectively. Then, the rationalized arcsine unit is defined as [55, pp. 456-457]

$$\theta = \arcsin\left(\sqrt{\frac{N_C}{N_T+1}}\right) + \arcsin\left(\sqrt{\frac{N_C+1}{N_T+1}}\right)$$
$$\text{RAU} = \frac{146}{\pi}\theta - 23. \qquad \blacktriangle$$

For each condition, all 24 participants SISs are converted to RAUs according to Definition 6.2 and are presented in Figure 6.4b. To examine the effect of the rationalized arcsine transform, we examine, if the data within each group is likely to be a realization of a Gaussian distribution for both the SISs and RAUs. For this examination, the Shapiro–Wilk test [56] is considered. In the Shapiro-Wilk test, the null-hypothesis is that data is a realization of a Gaussian distribution, hence for small p-values the null-hypothesis is rejected and there is evidence that the data is not a sample from a Gaussian distribution. For all non-control conditions, the test statistics W and corresponding p-values for the Shapiro-Wilk tests are presented in Table 6.2 for both the SISs and RAUs. Aside from the normality testing via the Shapiro-Wilk test, the proceeding statistical procedure is largely inspired by that of [27]. A description of the Shapiro-Wilk test can be found in Appendix B.

From the p-values presented in Table 6.2, we see, with a significance level of $\alpha = 0.05$, that without the rationalized arcsine transform, there is evidence against the null-hypothesis that data is a realization of a Gaussian distribution for the conditions MI-HIGH-50, INT-HIGH-50, INT-LOW-50, CSE-HIGH-50, and CSE-LOW-50. Also, for the condition CSE-LOW-75, the p-value is on the verge of being statistically significant. For the RAUs, we see from the p-values presented in Table 6.2, that we now only reject the null-hypothesis, that data is a realization of a Gaussian distribution, for the condition CSE-LOW-50. With the RAUs, we are ready to perform ANOVA, however, evidence suggests that the RAUs for the condition CSE-LOW-50 are not likely to be realizations from a Gaussian distribution, hence results involving this particular condition might not be valid.

Table 6.2. Shapiro-Wilk test statistic W and corresponding p-value for each non-control condition with scores represented as both SISs and RAUs.

Condition	SIS		RAU	
	Statistic	p-value	Statistic	p-value
MI-HIGH-50	$W = 0.8900$	0.0133	$W = 0.9198$	0.0578
MI-HIGH-75	$W = 0.9675$	0.6054	$W = 0.9677$	0.6111
MI-LOW-50	$W = 0.9692$	0.6480	$W = 0.9729$	0.7377
MI-LOW-75	$W = 0.9797$	0.8895	$W = 0.9787$	0.8717
INT-HIGH-50	$W = 0.9116$	0.0381	$W = 0.9571$	0.3824
INT-HIGH-75	$W = 0.9680$	0.6182	$W = 0.9694$	0.6524
INT-LOW-50	$W = 0.8797$	0.0082	$W = 0.9621$	0.4824
INT-LOW-75	$W = 0.9550$	0.3458	$W = 0.9601$	0.4399
CSE-HIGH-50	$W = 0.8674$	0.0047	$W = 0.9320$	0.1082
CSE-HIGH-75	$W = 0.9634$	0.5108	$W = 0.9724$	0.7258
CSE-LOW-50	$W = 0.7587$	< 0.0001	$W = 0.8395$	0.0014
CSE-LOW-75	$W = 0.9176$	0.0516	$W = 0.9674$	0.6042

First, a three-way repeated measures ANOVA is performed with the RAUs as the dependent variable and the independent variables being replacement percentage (50 or 75), replacement category (HIGH or LOW), and measure (MI, INT, or CSE). In the tables presented in this section, the factor HL represents replacement category (shorthand for HIGH/LOW) and $A \times B$ denotes interaction effect between A and B. Results for the three-way repeated measures ANOVA are presented in Table 6.3.

Table 6.3. ANOVA table for three-way repeated measures ANOVA with RAUs as the dependent variable and with replacement percentage, replacement category, and measure as the independent variables. In the column labeled Statistic, $F(X,Y) = Z$ denotes the F-distribution with X and Y degrees of freedom and test statistic Z. The p-value is the probability of observing values greater or equal to Z from the $F(X,Y)$-distribution.

Effect	Statistic	p-value	Significance level
$r_{\%}$	$F(1,23) = 1535.10$	< 0.001	$\alpha' = 0.0024$
HL	$F(1,23) = 82.61$	< 0.001	$\alpha' = 0.0033$
measure	$F(2,46) = 487.37$	< 0.001	$\alpha' = 0.0023$
$r_{\%} \times$ HL	$F(1,23) = 55.89$	< 0.001	$\alpha' = 0.0036$
$r_{\%} \times$ measure	$F(2,46) = 22.50$	< 0.001	$\alpha' = 0.0038$
HL $\times$ measure	$F(2,46) = 44.27$	< 0.001	$\alpha' = 0.0029$
$r_{\%} \times$ HL $\times$ measure	$F(2,46) = 2.60$	0.0850	$\alpha' = 0.0250$

For the ANOVAs, a significance level of $\alpha = 0.05$ is desired. However, when testing multiple hypotheses, the likelihood of incorrectly rejecting a null-hypothesis increases. Hence, the significance level should be corrected to account for this phenomenon. One method of correcting the significance level is the Bonferroni-Holm correction as described in Algorithm 2.

The ANOVAs presented in Tables 6.3 to 6.5, contain testing of 22 hypotheses. The Bonferroni-Holm correction, presented in Algorithm 2, is used to correct the significance levels for these

Algorithm 2 Bonferroni-Holm Correction [35, Sec. 6.6.4]

Input: Sorted p-values in ascending order $p_0, p_1, \ldots, p_{N-1}$ with associated null-hypotheses $H_0, H_1, \ldots, H_{N-1}$, desired significance level α

1: **for** $n = 0, 1, \ldots, N-1$ **do**
2: $\alpha' = \frac{\alpha}{N-n}$
3: **if** $p_n < \alpha'$ **then**
4: reject null-hypothesis H_n
5: **else**
6: accept null-hypotheses $H_m,\ m = n, n+1, \ldots, N-1$ and terminate algorithm
7: **end if**
8: **end for**

hypotheses with a desired significance level of $\alpha = 0.05$. The corrected significance levels are reported in the ANOVA tables as α'.

Consider Table 6.3, as expected, a significant main effect is present for the replacement percentage. Furthermore, significant main effects are also found for both replacement category and measure. Next, consider the interaction effects for the ANOVA in Table 6.3. These effects describe if the effect of one factor is dependent on the level of another factor, e.g., the effect $r_{\%} \times$ HL describes if the difference between the replacement categories HIGH and LOW is dependent on the value of the replacement percentage. Significant interaction effects are found for all three pairwise interactions. Finally, the three-way interaction $r_{\%} \times$ HL $\times$ measure proved statistically insignificant, meaning that the pairwise interaction HL $\times$ measure does not vary significantly across replacement percentages.

To further examine the interactions, a two-way repeated measures ANOVA is performed for each of the measure MI, INT, and CSE, separately. Hence, for each of the three two-way repeated measures ANOVAs the dependent variable is the RAUs observed for the given measure and the independent variables are the replacement percentage and the replacement category. The results for these three two-way repeated measures ANOVAs are depicted in Table 6.4.

Table 6.4. ANOVA tables for three separate two-way repeated measures ANOVAs, one for each of the measure MI, INT, and CSE. For the three ANOVAs, the dependent variable is the RAUs for the given measure and the two independent variables are replacement percentage and replacement category.

Measure	Effect	Statistic	p-value	Significance level
MI	$r_{\%}$	$F(1,23) = 1174.44$	< 0.001	$\alpha' = 0.0025$
	HL	$F(1,23) = 11.31$	0.003	$\alpha' = 0.0100$
	$r_{\%} \times$ HL	$F(1,23) = 16.43$	< 0.001	$\alpha' = 0.0083$
INT	$r_{\%}$	$F(1,23) = 538.98$	< 0.001	$\alpha' = 0.0026$
	HL	$F(1,23) = 42.19$	< 0.001	$\alpha' = 0.0042$
	$r_{\%} \times$ HL	$F(1,23) = 3.52$	0.073	$\alpha' = 0.0125$
CSE	$r_{\%}$	$F(1,23) = 490.36$	< 0.001	$\alpha' = 0.0028$
	HL	$F(1,23) = 104.99$	< 0.001	$\alpha' = 0.0045$
	$r_{\%} \times$ HL	$F(1,23) = 34.37$	< 0.001	$\alpha' = 0.0167$

For the three separate two-way repeated measures ANOVAs presented in Table 6.4, there is a significant main effect of replacement percentage for all three measures as expected. Furthermore, for the replacement category, significant main effects are present for all three measures. Finally, for the measures MI and CSE, a significant interaction effect between replacement percentage and replacement category was found. For this reason, separate comparisons between the replacement categories are carried out for each measure and replacement percentage. This is done with paired t-tests, which are equivalent to performing one-way repeated measures ANOVAs with the t-statistic being the square root of the F-statistic. For these six separate paired t-tests the dependent variable is once again the RAUs for the given measure and replacement percentage and the independent variable is the replacement category. The results for these paired t-tests are presented in Table 6.5.

Table 6.5. Paired t-tests for the six combinations of measure and replacement percentage. For the paired t-tests, the dependent variables is the RAUs for the given measure and replacement percentage and the independent variable is the replacement category.

Measure	$r_{\%}$	Effect	Statistic	p-value	Significance level
MI	50	HL	$t(23) = 5.34$	< 0.001	$\alpha' = 0.0056$
	75	HL	$t(23) = -0.51$	0.616	$\alpha' = 0.0500$
INT	50	HL	$t(23) = -4.334$	< 0.001	$\alpha' = 0.0071$
	75	HL	$t(23) = -5.54$	< 0.001	$\alpha' = 0.0050$
CSE	50	HL	$t(23) = -5.07$	< 0.001	$\alpha' = 0.0063$
	75	HL	$t(23) = -9.94$	< 0.001	$\alpha' = 0.0031$

From the results in Table 6.5, it is evident that for all six combinations of measure and replacement percentage except for MI at $r_{\%} = 75$, a significant effect of replacement category is found. Hence, the remaining five combinations of measure and replacement percentage does contain information about the importance of speech frames. For MI, the SIS - and in turn the RAU - decreases with the replacement category from HIGH to LOW, while for the measures INT and CSE, the SIS increases with the replacement category from HIGH to LOW.

6.2.2 Correlation Analysis

As originally presented in [27] and reproduced in Section 4.1.1, we consider correlation between measures in order to examine their relationships. For each of the 150 sentences in the Dantale II speech corpus, we compute the Spearman's correlations coefficient (see Definition 4.3) between MI and INT, between MI and CSE, and between CSE and INT. Note that the Spearman's correlations coefficients are computed solely on active frames as silent frames may alter the correlation artificially. The mean and standard deviation of the Spearman's correlations coefficients across the 150 sentences are presented in Table 6.6.

From Table 6.6, little to no correlation is present for either of the three pairings of measures. Worth noting is that the measures of MI and INT are essentially completely uncorrelated. However, perhaps the most interesting result is the correlation between CSE and INT. In [27], the

Table 6.6. Mean and standard deviation of the 150 Spearman's correlations coefficients between MI and INT, between MI and CSE, and between CSE and INT.

	MI vs. INT	MI vs. CSE	CSE vs. INT
μ_{ρ_S}	−0.012	−0.135	−0.201
σ_{ρ_S}	0.184	0.136	0.115

authors reported an average Spearman's correlations coefficient of $\rho_S = 0.926$ between CSE and INT, which we in this thesis (see Section 4.1.1) reproduced on a different speech corpus with a resulting average Spearman's correlations coefficient of $\rho_S = 0.829$. In this present experiment, however, we report an average Spearman's correlation coefficient of $\rho_S = -0.201$ between CSE and INT. Not only is the magnitude significantly lower but in fact the value is negative.

There are a few key differences to keep in mind between the two correlation experiments in [27] and Section 4.1.1 and this present correlation experiment. First, different speech corpora are used, with the Dantale II speech corpus (used in this present experiment) being Danish and the AZBio (used in [27]) and TIMIT speech corpora (used in Section 4.1.1) being American English. Secondly, the acoustic preprocessing differ quite a lot. In [27] and Section 4.1.1, a non-overlapping rectangular analysis window was used in the STFT and a 33 ROEX filter bank was applied, while in this present experiment, a Hann analysis window with overlap equal to half the window length is used in the STFT and a 18 one-third octave filter bank is applied. Finally, perhaps the most prominent difference is that in [27] and Section 4.1.1, successive non-overlapping frames were grouped together and used for computations, whereas in this present experiment, individual frame are used for computation.

As the computations of CSE in this thesis and in [27] are dissimilar, we do not claim to disprove the findings of [27]. However, our findings raise concern with the close connection between INT and spectral difference in terms of CSE proposed in [27].

6.2.3 Distribution of Misclassifications

In order to examine which parts of speech that are most affected by the measures of MI, INT, and CSE, we consider the distribution of misclassifications across the five word classes in the sentences. For each non-control condition, the number of misclassified words in each of the five word classes, was measured for all 24 participants. The proportion of misclassifications that each word class was accountable for is presented in Figure 6.5 for each non-control condition. From left to right the columns represent the measures MI, INT, and CSE. From top to bottom the rows represent replacement of the: 1) 50% active frames with the highest value of the given measure, 2) 75% active frames with the highest value of the given measure, 3) 50% active frames with the lowest value of the given measure, and 4) 75% active frames with the lowest value of the given measure. In each of the 12 panes, the five bars represent the five word classes from left to right.

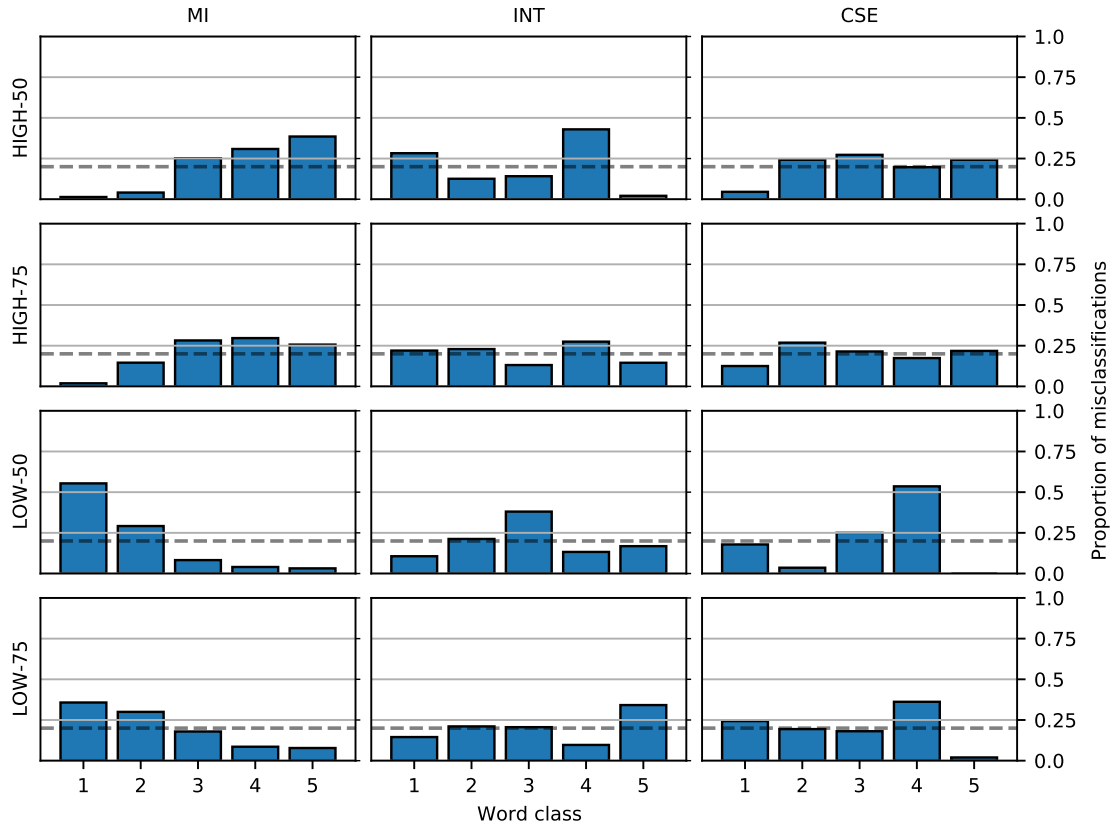


Figure 6.5. The proportion of misclassifications that each word class is accountable for, for each of the 12 non-control conditions. The dashed line represent 0.2 which corresponds to a uniform distribution of misclassifications across the five word classes.

With no prior knowledge, one would expect a uniform distribution of misclassifications across the five word classes, as the words in the Dantale II sentences are adjusted to be equally difficult to perceive [49, p. 11]. From the results presented in Figure 6.5, for some conditions this does not seem to be the case, as the dashed lines represent the uniform distribution. To measure to which amount the distribution of misclassifications diverge from the uniform distribution, the Kullback-Leiber divergence is considered (see Definition 3.6). As the Kullback-Leiber divergence is interpreted as the distance from the true distribution p_X to the assumed distribution q_X, we measure it with p_X as the distributions presented in Figure 6.5 and q_X as the uniform distribution. The Kullback-Leiber divergences for the 12 non-control conditions are presented in Table 6.7. In the computations of the Kullback-Leiber divergence, the base 2 logarithm is used in (3.13), resulting in the unit bits.

From Figure 6.5 and Table 6.7, multiple interesting tendencies are observed. In general the distribution of misclassifications more closely resembles the uniform distribution when $r_{\%} = 75$ compared to when $r_{\%} = 50$.

For MI-LOW (both for $r_{\%} = 50$ and $r_{\%} = 75$), misclassifications are heavily represented in the start of the sentences, specifically in the first word class. In fact, for MI-LOW-50, the first word class is accountable for over half the misclassifications. On the contrary, few misclassifications

Table 6.7. Kullback-Leiber divergence between distribution of misclassifications and the uniform distribution. In relation to Definition 3.6 the former is p_X while the latter is q_X.

HL	$r_{\%}$	Measure		
		MI	INT	CSE
HIGH	50	0.4940 [bits]	0.3911 [bits]	0.1551 [bits]
	75	0.2703 [bits]	0.0540 [bits]	0.0432 [bits]
LOW	50	0.6903 [bits]	0.1540 [bits]	0.7240 [bits]
	75	0.2355 [bits]	0.1198 [bits]	0.2780 [bits]

occur in the first two word classes for MI-HIGH. These observations are supported by the fact that in Table 6.7, the Kullback-Leiber divergences are relatively large for the conditions MI-HIGH-50 and MI-LOW-50. One possible explanation for this phenomenon is that the measure MI uses temporal estimates. As described in Section 6.1.3 these recursive updates are described by the time constant of $\tau \approx 0.15$ [ms], meaning that the value of MI might be inaccurate in the beginning of the sentences.

A similar tendency to that of MI is observed for the measure CSE, in that few misclassifications are reported for the first word class for CSE-HIGH-50 and for the last word class for CSE-LOW (both for $r_{\%} = 50$ and $r_{\%} = 50$). In fact, across all 24 participants, not a single misclassification of the last word class was observed for CSE-LOW-50. This is also reflected in Table 6.7, in that CSE-LOW-50 is accountable for the largest Kullback-Leiber divergence. However, this tendency cannot be caused by inaccurate estimates, as the measure CSE takes only the previous frame into account.

6.2.4 Frame Replacement Adjacency

As described in Section 6.2.1, SISs for sentences processed with the MI measure was generally lower than those processed with the INT measure, which in turn generally were lower than those processed with the CSE measure. In this section, we examine how the replaced frames are distributed across the sentences.

To consider the distribution of the replaced frames, we consider a sample sentence from the Dantale II speech corpus. For each of the 12 non-control conditions, the replaced frames are depicted in Figure 6.6. Note that silent frames are included in this figure, and simply fall under the "Kept" category.

Figure 6.6 suggests that for the measure MI, frames tend to be replaced in larger segments compared to the measures INT and CSE. Moreover, replaced frames tend to be more grouped for the measure INT than for the measure CSE. To examine whether this tendency is present throughout the Dantale II speech corpus, the lengths of the replaced segments are considered. For all 150 Dantale II sentences, the lengths of the replaced segments are recorded for each of the 12 non-control conditions. The average length of the replaced segments for the 12 non-control

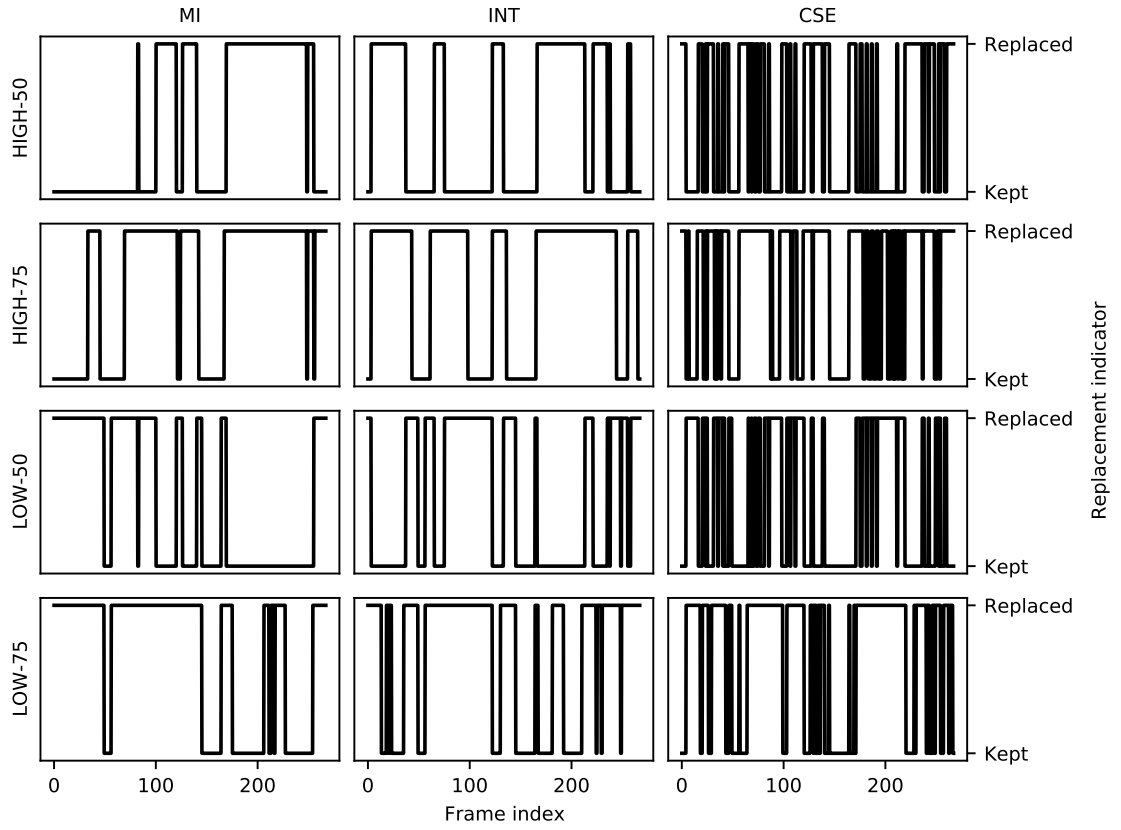


Figure 6.6. Sample sentence from the Dantale II speech corpus exemplifying which frames are replaced for the 12 non-control conditions.

conditions are presented in Table 6.8. Note that a frame effectively amounts to 8 [ms].

Table 6.8. Average length of the replaced segments for the 12 non-control conditions.

HL	$r_{\%}$	Measure		
		MI	INT	CSE
HIGH	50	134.43 [ms]	122.43 [ms]	31.52 [ms]
	75	198.70 [ms]	140.59 [ms]	54.95 [ms]
LOW	50	140.33 [ms]	75.25 [ms]	32.53 [ms]
	75	189.23 [ms]	124.41 [ms]	56.69 [ms]

From Table 6.8, it is evident that the average length of replaced segments is on average larger for the measure MI than for the measure INT, which in turn is larger on average than for the measure CSE. This fact supports the observation regarding Figure 6.6, namely that replaced frames tend to be more grouped for the measure MI compared to the measure INT and CSE, and in turn replaced frames tend to be more grouped for the measure INT compared to the measure CSE. Interesting to note is the average length of INT-LOW-50 segments, as this is much smaller than the average length for INT-HIGH-50 segments. This is likely explained by the fact that frames adjacent to silent frames will have low INT, hence, when replacing INT-LOW segments, the first frames to be replaced will likely be frames adjacent to silent frames, which necessarily

are located apart.

To comprehend the impact of the durations presented in Table 6.8, we compare them to the average phoneme duration. The Dantale II speech corpus is not annotated with phoneme start and end points, thus, we are not able to obtain the average phoneme duration for this corpus as the process of annotating a speech corpus is both extensively time consuming and without proper experience prone to produce errors. Alternatively, the TIMIT speech corpus is annotated with phoneme start and end points. Though, these corpora differ quite a bit, we choose to use the TIMIT speech corpus to measure average phoneme duration.

The average phoneme duration across all 6300 sentences in the TIMIT speech corpus is 74 [ms]. From Table 6.8, we see that for the measure CSE the duration of replaced segments are on average shorter than the average phoneme duration. For the measure INT, the average durations of replaced segments range from one to nearly two times the average phoneme duration. Lastly, for the measure MI, average durations of replaced segments constitute approximately two to three times the average phoneme duration.

To support the aforementioned observations of the average length of replaced segments, we present another method to quantify to which extend replaced frames are grouped across the entire Dantale II speech corpus for the three measures. To do this, we compute the empirical cumulative distribution function (CDF) for the set of values with zeros at indices of frames which are kept and values at indices which are replaced. For each sentences, this empirical CDF is computed for the 12 non-control conditions. The more a empirical CDF resembles the CDF of the uniform distribution with equal support, the more equally spread the replaced frames must be. Hence, small resemblance indicates more grouping of the replaced frames.

The resemblance between CDFs are measures as follows. First, let $\Delta_{i,j}$ denote the index set of the replaced frames for the ith sentence restricted to values less than or equal to j. Next, let Q_i denote the total number of frames in the ith sentence and let Q_i' denote the number of replaced frames in the ith sentence. Then, the empirical CDF for the ith sentence and the CDF of the uniform distribution for the ith sentence are given as

$$F_i[j] = \sum_{k \in \Delta_{i,j}} \frac{1}{Q_i'}, \quad i = 0, 1, \ldots, 149,$$

$$G_i[j] = \sum_{k=0}^{j} \frac{1}{Q_i}, \qquad i = 0, 1, \ldots, 149,$$

respectively. Subsequently, the resemblance for the ith sentence is computed as

$$A_i = \frac{1}{Q_i} \sum_{j=0}^{Q_i} \left| F_i[j] - G_i[j] \right|, \quad i = 0, 1, \ldots, 149. \tag{6.2}$$

As an example, Figure 6.7 depicts the empirical CDFs for MI-LOW-50, INT-LOW-50, and CSE-LOW-50 as well as the CDF for the uniform distribution for the same sentence used in Figure 6.6. Alongside the CDFs, resemblance computed according to (6.2) is presented.

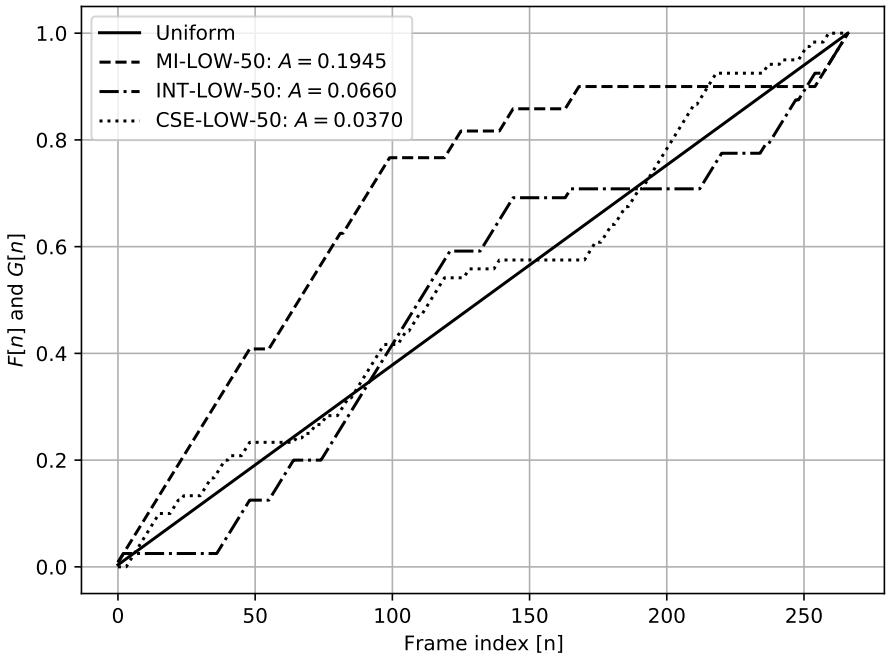


Figure 6.7. Empirical CDFs for MI-LOW-50, INT-LOW-50, and CSE-LOW-50 as well as CDF for the uniform distribution for the sample sentence of the Dantale II speech corpus used in Figure 6.6.

From Figure 6.7, it seems that the CDF of CSE-LOW-50 more closely resembles the uniform CDF compared to the CDFs of MI-LOW-50 and INT-LOW-50. Furthermore, the CDF of MI-LOW-50 diverges more from the uniform CDF than the CDF of INT-LOW-50 does. This observation is supported by the resemblance scores presented in Figure 6.7 and coincides with the observations made in relation to Figure 6.6.

For each of the 12 non-control conditions, the resemblance from (6.2) is computed for all 150 sentences of the Dantale II speech corpus and the averages are presented in Table 6.9.

Table 6.9. Resemblance between empirical CDF and CDF of uniform distribution for each of the 12 non-control conditions averaged over the 150 Dantale II sentences.

HL	$r_{\%}$	Measure		
		MI	INT	CSE
HIGH	50	0.1568	0.0634	0.0407
	75	0.0862	0.0336	0.0260
LOW	50	0.1594	0.0556	0.0462
	75	0.0714	0.0431	0.0248

From Table 6.9, we see that within each replacement percentage a trend is apparent, in that resemblance is larger - regardless of the replacement category - for MI than for INT and CSE, and in turn resemblance is larger for INT than for CSE. This result suggests that, in general, the replacement of frames is more grouped for the measure MI than for the measures INT and CSE, and in turn, replaced frames are more grouped for the measure INT, than for the measure CSE.

This difference in grouping of the replaced frames for the three measures likely take part in explaining the overall difference in SISs between the three measures presented in Section 6.2.1.

6.3 Predictability of High Intensity Frames

In this section, we aim to answer the second research sub-question presented in Chapter 1, namely "Are the speech frames most important for speech intelligibility double protected from acoustic noise, by being characterized by both high sound intensity and high predictability?" We examine this, in the following manner. In a given sentence, we first quantify the predictability of all frames. This could be done using MI, however, since the value of MI is difficult to interpret, we use the CNNs presented in Chapter 5 to extrapolate all frames based on intact context. Next, consider the log-spectral distortion (LSD).

Definition 6.3 (Log-Spectral Distortion)
Let $\boldsymbol{P}, \widehat{\boldsymbol{P}} \in \mathbb{R}^L$ denote two magnitude spectra on a logarithm scale. Then, the log-spectral distortion is defined as [16, Sec. 4.5], [57, Sec. II]

$$\mathrm{LSD}\left(\boldsymbol{P}, \widehat{\boldsymbol{P}}\right) = \sqrt{\frac{1}{L}\sum_{l=0}^{L-1}\left(P[l] - \widehat{P}[l]\right)^2}. \qquad \blacktriangle$$

The LSD is used to measure the distortion of using an extrapolated frame $\widehat{\boldsymbol{S}}[:, m]$ instead of the true frame $\boldsymbol{S}[:, m]$.

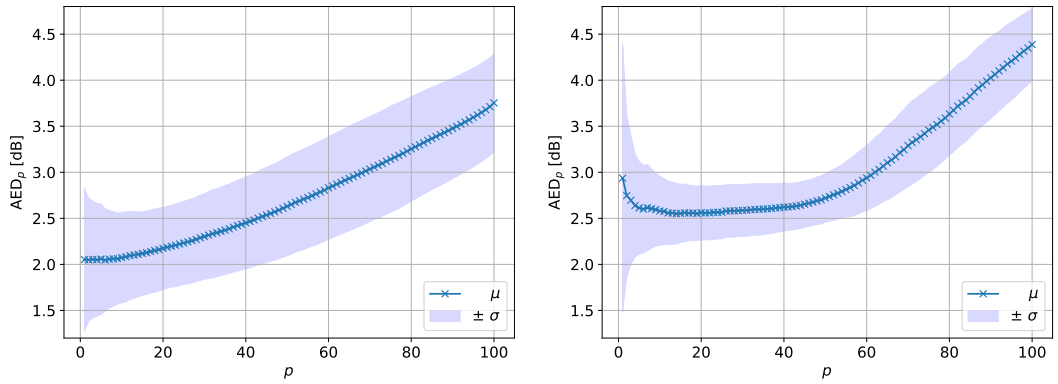
Next, to compute the average distortion of extrapolating multiple frames, consider the average extrapolation distortion (AED).

Definition 6.4 (Average Extrapolation Distortion)
Let $\boldsymbol{P}, \widehat{\boldsymbol{P}} \in \mathbb{R}^{L\times N}$ denote two short-time magnitude spectra on a logarithm scale. Let p and Z_p denote the percentage of frames which are extrapolated and the index set of these frames, respectively. Then, the average extrapolation distortion is defined as

$$\mathrm{AED}_p = \frac{1}{\mathrm{Card}\left(Z_p\right)}\sum_{n\in Z_p}\mathrm{LSD}\left(\boldsymbol{P}[:, n], \widehat{\boldsymbol{P}}[:, n]\right). \qquad \blacktriangle$$

In Figure 6.8, we present AED_p averaged over all 150 Dantale II sentences and over 320 TIMIT sentences for the $p = 1, 2, \ldots, 100$ percent of active frames characterized by highest INT. Note that for the TIMIT sentences, 40 sentences were chosen randomly from each demographic region and none of these sentences were used in the training of the neural networks.

From Figure 6.8a, we see that the AED_p is approximately monotonically increasing with increasing values of p. This tendency is also present in Figure 6.8b beginning at $p = 14$, however, before this value of p, we see a decrease in extrapolation distortion with increasing values of p. This general tendency indicates that the frames characterized by high INT are more predictable than frames characterized by low INT, as the inclusion of low INT frames increases extrapolation distortion.



(a) TIMIT speech corpus.

(b) Dantale II speech corpus.

Figure 6.8. AED$_p$ versus percentage of highest INT frames extrapolated averaged over **(a)** 320 sentences from the TIMIT speech corpus and over **(b)** the 150 Dantale II sentences.

It is worth examining what causes the decrease in AED$_p$ with increasing p in Figure 6.8b, for $p \in [1, 13]$, as this is not the case in Figure 6.8a. The neural networks used to extrapolate frames, are trained on the TIMIT speech corpus. This makes the models more accustomed to test sentences of the TIMIT speech corpus than to the Dantale II sentences, and perhaps, leads to unforeseen difficulties regarding extrapolation in certain conditions. As the AED$_p$ in Figure 6.8a is generally stagnant for $p \in [1,6]$, this difference in familiarity of the models and data might explain this behavior. Furthermore, it worth noting that the AED$_p$ is smaller for all values of p in Figure 6.8a compared to Figure 6.8b, which again likely is to be explained by the aforementioned difference in familiarity of the models and data.

Besides the shape of the AED$_p$ curves in Figure 6.8, it would be beneficial to be able to evaluate if the extrapolation distortion of high INT frames is low enough to be considered acceptable. In [58, p. 6], the authors report that maintaining the following three conditions produces no noticeable difference between the extrapolated frames and the original frames: 1) the average LSD - hence AED - is below 1 [dB], 2) there is no outlier frames having LSD larger than 4 [dB], and 3) less than 2% of frames have a LSD in the range $2-4$ [dB]. The thresholds in the above conditions are, however, constructed for spectra of all-pole models. All-pole models spectra are generally smoother than DFT spectra, resulting in all-pole model spectra having lower LSD [16, pp. 158]. However, we cannot guarantee that this is the case in our setup, as our DFT spectra are processed by a one-third octave filter bank. Though possibly justifying the rather high values of AED$_p$ in Figure 6.8 compared to the thresholds presented in [58], we are unaware of any sensible way of evaluating the magnitudes of AED$_p$.

7 Discussion

The aim of this chapter is to discuss the results and specifications presented in Chapter 6 as well as to discuss several aspects of the MI measure proposed in this thesis. In Section 7.1, the performance of the measures MI, INT, and CSE is discussed, while the limitations of the measures are discussed in Section 7.2. In Section 7.3, the likeliness of the measures are discussed and finally, in Section 7.4, the experiment presented in Section 6.3 is discussed.

7.1 Performance of Measures

In this thesis, we examine whether a predictive measure such as MI or CSE, or a more simple measure such as INT is the most prominent contributor to speech intelligibility. In order to examine this, a listening test has been conducted. The framework for this test was outlined in Section 6.1 and the results were presented in Section 6.2.

From the results of the listening test depicted in Figure 6.4, it was seen the that CSE and INT are able to predicts speech intelligibility as expected from [17], [27]. This is the case as we saw a larger degradation in the SIS for the replacement category HIGH than that of the replacement category LOW. On the other hand, for the proposed measure MI, we expected larger degradation in SIS for the replacement category LOW, than for the replacement category HIGH. This behavior was the case for $r_{\%} = 50$, however, for $r_{\%} = 75$, the difference between the means of MI-HIGH and MI-LOW SISs was found to be statistically insignificant. This indicates that at least in certain conditions, MI is able to predict speech intelligibility.

The previous paragraph suggests that, for MI, the replacement percentage affects the magnitude of the difference between RAUs for the replacement categories HIGH and LOW. This fact is supported by a two-way repeated measures ANOVA (see Table 6.4), as the interaction effect $r_{\%} \times$ HL is statistically significant. In fact in the three-way repeated measures ANOVA in Table 6.3 the $r_{\%} \times$ HL interaction effect is statistically significant, suggesting that, in general, the replacement percentage affects the difference between HIGH and LOW within the three measures. As for MI, we found a statistically significant interaction effect $r_{\%} \times$ HL for the measure CSE (see Table 6.4), and these observations are evident from Figure 6.4. Contrary to MI, for CSE a statistical significant effect of replacement category for both $r_{\%} = 50$ and $r_{\%} = 75$ was

found in (see Table 6.5), suggesting that CSE is able to predict speech intelligibility regardless of the replacement percentage, however, proves better at $r_{\%} = 75$. For INT, a statistical significant effect of replacement category was found for both $r_{\%} = 50$ and $r_{\%} = 75$ (see Table 6.5). However, the interaction effect $r_{\%} \times$ HL in the two-way repeated measures ANOVA in Table 6.4 did prove statistically insignificant, suggesting that INT does predict speech intelligibility and does so equally well, regardless of replacement percentage.

All three measures, MI, INT, and, CSE are able to predict speech intelligibility, at least in certain conditions. Next comes the question of which measure is the most prominent contributor to speech intelligibility. As results differ quite a lot for the two replacement percentages, we start by considering them individually.

As found in Section 6.2 for $r_{\%} = 75$, MI is unable to predict speech intelligibility with an insignificant difference in SISs between MI-LOW and MI-HIGH of 0.77 corresponding to 0.70 [RAUs]. For INT, the replacement category HIGH resulted in a larger SIS degradation of 12.74 corresponding to 11.99 [RAUs] compared to LOW, while this value for CSE was 17.14 (18.47 [RAUs]). Note that here, the absolute differences are reported and direction is inferred from Figure 6.4. These results indicate that for replacement percentage for $r_{\%} = 75$, CSE better predicts speech intelligibility than INT does, whilst MI is unable to do so.

For $r_{\%} = 50$, all three measures are able to predict speech intelligibility as supported by the t-tests presented in Table 6.4. The difference between replacement categories within each measure, in terms of RAUs, are 7.09 [RAUs], 7.31 [RAUs], and 6.42 [RAUs] for MI, INT, and CSE, respectively. This suggests a slight superiority of INT over MI followed by CSE in terms of RAUs. On that note, it is important to notice the relationship between SISs and RAUs, as especially the edges of the scales are dissimilar. This is seen in Figure 7.1, which depicts the difference between SISs and RAUs for a sample size of $N_T = 70$, as used in the listening test (five words times 14 sentences). From the results of the listening test, it was found that SISs for MI are

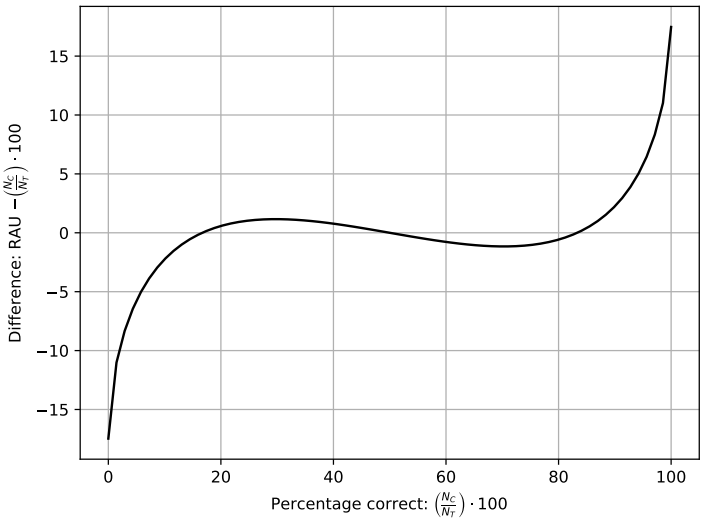


Figure 7.1. Difference between SIS and RAU with $N_T = 70$. After [55, Fig. 1].

generally closer to the middle of the scale than those of INT, which in turn are closer to the middle of the scale than those of CSE. This means larger dissimilarity between SISs and RAUs for CSE than for INT, and larger dissimilarity between SISs and RAUs for INT than for MI. The difference between replacement categories within each measure in term of SISs are 6.37, 4.64 and 2.26 for MI, INT and CSE, respectively. Contrary to the RAUs, we saw from the SIS, that MI is superior to INT, which in turn is superior to CSE.

In general, based on the results of the listening test presented in this thesis, no measure can be crowned superior, as MI provided best predictive abilities at $r_{\%} = 50$, CSE did so at $r_{\%} = 75$ and INT performed well across both replacement percentages.

7.2 Limitations of Measures

In this section the limitations and validity of the measure MI, INT, and CSE are discussed.

First of all, the value of MI at a given frame of the speech signal is computed based upon the previous $M = 14$ frames, CSE requires that only the previous frame is intact, and INT requires no knowledge of past frames. In relation to the listening test, the aforementioned requirements provide an inherit advantage in terms of validity of INT compared to MI and CSE. This is because the MI and CSE values computed throughout the Dantale II sentences are computed with the required intact context. However, this context is not necessarily presented for the participants. In fact, since large percentages of active frames are replaced, this raises concerns to whether MI and CSE are suitable for a listening test as conducted in this thesis. To account for this issue, either the test methodology or the definition of the measures could be altered.

In the following, we discuss the test methodology. In [17], [27], the required context is preserved in the listening test. However, segments of contiguous frames are replaced whereas we replace singles frames. If we were to preserve required context, this would results in replacement of single frames which effectively amounts to 8 [ms] in between intact context, which most likely will not affect speech intelligibility. This is supported by the fact that SISs for CSE - regardless of replacement percentage and replacement category - are in general higher than those for MI and INT, while results presented in Section 6.2.4, suggest smaller grouping of replaced frames for CSE than for MI and INT. Hence, replacing only single frames would probably not be a feasible solution. However, the effective length of frames could be enlarged by increasing the size of the analysis window in the STFT to a point at which speech intelligibility is affected by the replacement of a single frame. E.g., the frame length could correspond to the average phoneme length found in Section 6.2.4 to be 74 [ms] for the TIMIT speech corpus. Keeping in mind that such a frame length is very large in the field of speech processing, hence, it might not be an ideal option as the temporal resolution is severely reduced. In addition, this alteration of test methodology may also alters the measures.

Let us now discuss possible options to alter the measures. As CSE is nothing more than the spectral difference between two frames summed across subbands, this difference could be re-

computed, if the context frame has been replaced by noise. Two method of recomputing CSE comes to mind. First, a given CSE value could be computed based on the most previous intact frame. However, if multiple frames are replaced in succession, CSE might be computed upon two frames which are temporally distant in such a way, that they are essentially uncorrelated. Therefore, in such scenario we would expect large values of CSE. Secondly, CSE could be computed based upon the SSN frame which has replaced the required context. In this way, CSE is still computed upon two consecutive frames and reflects the stimuli that the participants perceive. This means that the replacement procedure for a sentence becomes as follows. First, the CSE is computed for all frames, the frame with the largest/smallest value of CSE (dependent on the condition in question) is replaced and omitted from replacement consideration. Recomputing CSE with either of the aforementioned methods, replaces a frame and repeats until a desired amount of frames have been replaced. Using this method would result in no violation of the assumption of required context for CSE.

The measure MI is more complex than CSE, hence, the alterations suggested for CSE cannot simply be applied for MI. The measure MI is - among other things - comprised of a set of CNN MMSE estimators, which perform computations on segments of $M = 14$ frames. The reasoning for choosing $M = 14$ was that this amounts to approximately 112 [ms], as this duration generally contains both vowel and consonant sounds [17, p. 12388]. If smaller M was chosen, the likelihood of violating the assumption of intact context would be decreased, however, possibly affecting the performance of the neural networks. Moreover, the neural networks could be trained on damaged data, i.e., data in which random frames are replaced by SSN, in order to make the networks robust to the occurrence of SSN frames in the test data. Besides neural networks, the measure of MI contains multiple temporal estimates, for which the replacement of frames must also be addressed, such that SSN frames are included in these estimates if they are contained in the context. This method would entail the compute/replace cycle of the CSE from the previous paragraph, in that, after a frame has been replaced with noise, the MI should be recomputed for the sentence before another frame is replaced. Yet, another method of correcting the aforementioned violation regarding intact context for MI, is to revisit how we defined the measure. MI could be measured between two segments of successive frames where the newest (in relation to time) segment would span a duration large enough to degrade speech intelligibility. This way, a segment of frames could be replaced while the context would remain intact.

A few more concerns must be addressed for the measure MI. In regards to the CNN MMSE estimators, the TIMIT speech corpus, which is American English, was used for training while the Dantale II speech corpus, which is Danish, was used in the listening test. Furthermore, both male and female speakers constitute the TIMIT speech corpus while a single female speaker utters the entire Dantale II speech corpus. Also, the noise level and the recording equipment differ for the two corpora. These differences may introduce errors in the neural networks, and ideally, the training process should be more reminiscent of the testing process. The effect of

this was seen in Figure 6.8, where the CNN MMSE estimators were used to extrapolate frames of TIMIT and Dantale II sentences. The AED was used to measure the average distortion of using the extrapolated frames. We generally saw smaller AED for the TIMIT speech corpus than for the Dantale II speech corpus, suggesting larger familiarity of the CNNs with the former speech corpus. Transfer learning is the technique of training a model on one task and retraining the model on a different but related task [46, p. 539]. Hence, transfer learning could be employed by training the model on the TIMIT speech corpus, to hopefully learn general spectro-temporal structures in speech, and then retraining the model on the Dantale II speech corpus to learn features more specifically related to this corpus. However, problems arise with the small amount of data in the Dantale II speech corpus, and therefore, utilizing transfer learning is probably not possible with the use of the Dantale II speech corpus. If another speech corpus was to be used for testing, transfer learning might be feasible.

Another concern with the proposed measure of MI is that - as we saw in Section 6.2.3 - we found that for MI-LOW the majority of misclassifications occurred in the first word class whilst nearly no misclassifications occurred in this word class for MI-HIGH. These observations suggest that MI is generally low in the beginning of sentences. The measure MI uses temporal averages to estimate variance components. These estimates are prone to large variance in the beginning of sentences as the recursive filters have yet to reach steady-state. This fact may in part explain the tendency of low MI in the start of sentences and to circumvent this issue, the time constant τ, which control the temporal dependency of the temporal estimates, could be adjusted. Another option is to omit the first word class from the Dantale II speech corpus from the frame replacement procedure. In this way, for a given sentence, MI - as well as INT and CSE - would be measured for all frames, but only frames contained in the last four word classes would be considered for replacement. Then, the entire sentence would be presented for the participants. This process would allow the temporal filters to reach steady-state before any frames are replaced.

On another subject, it is worth noting the computational complexity of the three measures as these differ quite a lot. The measure INT is computationally light as the processing is a sum of squared time domain amplitudes. The measure CSE requires a bit more processing as it involves the STFT, however, the FFT is an immensely efficient algorithm resulting in a computationally light measure. Lastly, MI is far more computationally heavy compared to INT and CSE. Just as for CSE, the processing occurs in the spectro-temporal domain but what really makes the difference is the CNN MMSE estimators. To obtain a value of MI for a given frame, the previous $M = 14$ frames are propagated through the $L = 18$ CNNs. The architecture, size and hyperparameters of each of the neural networks were chosen on the basis of relatively small preliminary tests. However, other configurations - not tested in this thesis - may produce equally good or better results with fewer parameters to train, resulting in a less computationally heavy model.

In [27], the authors raised concern with CSE, in that it is computed upon the linear scale mag-

nitude spectra. The authors suggest that the measure would be more relatable to the human auditory system if the magnitude spectrum was represented on a logarithmic scale, such as the decibel scale. To that end, in [27], the authors introduced CSE computed on the decibel scale termed dB-CSE, however, dB-CSE did not prove able to predict speech intelligibility. As for dB-CSE, the measure MI is computed upon the decibel scaled magnitude spectra, which suggests an advantage of MI over CSE in terms of perceptual relevance. The use of the decibel scale is not a necessity as other well accepted speech processing algorithms do not apply this transformation, e.g., [8], [9], [10]. However, transformation to the decibel scale is generally a good approach as, this scale is more closely related to human perception than a linear scale [27].

Regardless of the measure, a energy-based VAD was used to exclude silent frames from the replacement procedure. The threshold parameter Δ_E, which determines the activity allowed in a frame in order to be classified as silent, has not been optimized. We have used $\Delta_E = 40$ [dB] as proposed in [8], [10] in which the Dantale II speech corpus was used. However, this value for Δ_E might not be appropriate for the TIMIT speech corpus and should possible have been altered. As speech does not start and end instantaneous, the process of optimizing the value of Δ_E is difficult. Inappropriate choices for Δ_E include choosing a too conservative value resulting in near-silent frames being included in the replacement procedure while a too aggressive value could exclude speech-active frame, e.g., onsets and offsets. We examined VAD outputs for $\Delta_E = 40$ [dB] and deemed the results acceptable, yet no fine-tuning was conducted. The threshold will have a direct impact on the measure INT, as frames with INT just clearing the threshold, Δ_E, will be the active frames with the lowest INT. This means that if an unsuitable threshold is chosen, the INT measure might be distorted as the INT-LOW conditions might contain near-silent frames, which should have been excluded by a properly working VAD, possible inflating the difference between the replacement categories HIGH and LOW for the measure INT.

To round off this section, the control condition - in which no frames were replaced - of the listening test is discussed. No comparisons have been made to this condition as this would produce no relevant information. If a revised listening test were to be conducted, a control condition in which random frames were replaced, would be better suited for the framework of the test. The amount of frames replaced at random should then correspond to the replacement percentage as for the three measures. We would expect that such a control condition would harm speech intelligibility less than MI-LOW, INT-HIGH, and CSE-HIGH, as these conditions should represent frames which are most important for speech intelligibility. On the other hand, we would expect that it would harm speech intelligibility more than MI-HIGH, INT-LOW, and CSE-LOW, as these conditions should represent frames which are least important for speech intelligibility.

Based on the limitations presented in this section, it is deemed that when it comes to the ability of identifying frames most important for speech intelligibility, we cannot declare whether

MI, CSE, or INT is best suited. We can, however, conclude that CSE and especially MI have drawbacks which should be dealt with in order to more sensible evaluate the measures.

7.3 Likeliness of Measures

As discussed in Section 7.1, all three measures are able to predict speech intelligibility, at least in certain conditions. In Section 6.2.2, a correlation experiment was conducted to examine how closely related the measures are. The average Spearman's correlation coefficient was computed over the Dantale II speech corpus for the three pairing of the measures where we found all three pairings to be essentially uncorrelated. Parts of this result is surprising, as [27] and [59] found correlation coefficients of 0.926 and 0.79 between CSE and INT, respectively, whereas we found a correlation coefficient of -0.201 between CSE and INT. This difference in results might in part be explained by the fact that we use a different front-end processing, i.e., STFT parameters and choice of filter bank, than the one originally proposed in [17]. However, presumably a more prominent contributing factor to this difference in the results, is that we omit the boxcar summation in CSE, i.e., we let $J = 1$ in line 6 in Algorithm 1. As mentioned in Section 6.1, the reason for omitting this boxcar summation is that it essentially acts as a low-pass filter resulting in a more slowly varying measure but otherwise serves not perceptually relevant purpose.

In Section 6.2.4, the average length of replaced segments for the 12 non-control conditions was examined. This experiment was supported by observations of to which extend frames were grouped when replaced according to the non-control conditions. The results of these two experiments coincided, in that we found that, in general, MI is a more rapidly varying measure than INT, which in turn is a more rapidly varying measure than CSE, while frames are replaced in larger groups for MI than for INT, and frames are replaced in larger groups for INT than CSE. The relatively large discrepancy in average length of replaced segments for the three measures - presented in Table 6.8 - may well explain the absence of correlation between the measures.

An interesting result to notice is the fact that all measure are able to predict speech intelligibility while being essentially uncorrelated, suggesting that one measure contains information about speech intelligibility that is not present in the other measures.

7.4 Double Protection from Acoustic Noise

In Section 6.3, an experiment was conducted in order to answer our second research subquestion from Chapter 1, namely if frames characterized by high INT - which we found in the listening test to contribute to speech importance - are more predictable than frames characterized by low INT. We measured the AED when $p = 1, 2, \ldots, 100$ percent of active frames with highest INT were extrapolated using the CNN MMSE estimators.

For the TIMIT speech corpus, we found a monotonically (if noisy oscillations are ignored) increasing relationship between AED_p and p, suggesting that frames characterized by high INT

are indeed more predictable than frames characterized by low INT. Note that only a subset of 320 sentences were used from the TIMIT speech corpus due to large time consumption associated with the CNN MMSE estimators used for predictions. Ideally, a larger amount of sentences should be used to increase validity of this experiment.

For the Dantale II speech corpus, the AED_p was monotonically increasing with increasing p, when $p \geq 14$. We expect this discrepancy to be a result of the fact that the CNN MMSE estimators were trained on the TIMIT speech corpus, hence a matter of not being properly accustomed to specific features of the Dantale II speech corpus.

Despite the inconclusiveness of the values of AED_p for $p < 14$ for the Dantale II speech corpus, our results strongly supports the hypothesis that frames characterized by high INT are more predictable than frames characterized by low INT, hence acting as a double protection from acoustic noise.

Conclusion 8

We have in this thesis investigated the potentials of using an information theoretical measure to quantify predictability and in turn rank importance of speech frames. Specifically, the main research question of this thesis is:

How can a non-intrusive objective measure of predictability of speech be operationalized in terms of mutual information in order to locate elements of speech, which contribute most to speech intelligibility?

In order to answer this question, a formulation of mutual information, referred to as MI, inspired by that of [9], was proposed. The proposed measure contrast that of [9] in three major ways: 1) our proposed measure is non-intrusive whereas the method of [9] measure mutual information between a processed and a clean signal, hence, is intrusive, 2) in the formulation of mutual information a MMSE estimator emerges, where we in this work, have proposed to realize the MMSE estimator as a system of CNNs, and whereas in [9], the MMSE estimator is approximated using a LMMSE estimator, and 3) in our proposed measure, the magnitude spectra are represented on the decibel scale, since this scale is more closely related to the human perception system, as opposed to the linear scale used in [9].

To evaluate the performance of our proposed measure, a listening test was conducted in which frames characterized by either high or low mutual information were replaced with speech shaped noise. Likewise, cochlea-scaled spectral entropy (CSE), originally presented in [17] but adapted to our experimental framework, and sound intensity (INT) were included in the listening test. In each sentence used in the listening test, either 50% or 75% of frames were replaced based on the three measures MI, INT, and CSE. When 50% of frames were replaced, results from our experiments, presented in Section 6.2, indicated that our proposed measure better identified frames contributing to speech intelligibility than INT and CSE did. However, when 75% of frames were replaced, our proposed measure was unable to identify important speech frames. Following this, in Chapter 7, it was discussed that this could be due to the fact, that MI, for a given frame, requires that the previous 112 [ms] of the signal are intact. This assumption was neither upheld for 50% nor 75%, but, was more invalidated for the latter. The invalidation of

this assumption, meant that in general, participants were not presented with the same stimuli as the measure MI was computed upon.

We conclude that our proposed measure of MI did prove favorable to CSE and INT in certain conditions. However, in order to more sensible assess our proposed measure MI, we deem that the aforementioned considerations about validity must be addressed.

In this thesis, two research sub-questions have been considered, the first of which is:

In terms of speech intelligibility, is importance of speech frames governed by how predictable they are from past context?

In the listening test, our proposed measure MI and CSE were included as measures of predictability. As both measure were - at least in certain conditions - able to identify frames important to speech intelligibility, we conclude that importance of speech frames are governed by how predictable they are from past context. However, as INT was also able to identify important speech frames, while proving essentially uncorrelated with MI and CSE, we argue that predictability is not the sole contributing factor to speech intelligibility.

The second research sub-questions that has been considered in the thesis is:

Are the speech frames most important for speech intelligibility double protected from acoustic noise, by being characterized by both high sound intensity and high predictability?

To answer this questions, we measured the average log-spectral distortion of extrapolating frames characterized by different amounts of INT. This extrapolation was carried out by deep convolutional neural networks which are embedded in MI. We found a general tendency in that frames characterized by high INT resulted in smaller average log-spectral distortion than frames characterized by low INT. From this result, along with the fact that INT was also able to identify important speech frames in the listening test, we conclude that speech frames important for speech intelligibility are double protected from acoustic noise by being characterized by both high INT and high predictability.

Further Development 9

In Chapter 7, we discussed aspects of our proposed measure, including limitations which should be dealt with, in order to more sensible assess the measure. If we were to further develop our proposed measure, these limitations should be considered, however, more general alterations could also be considered. In this chapter, we present thoughts about these alterations.

One of the general premises of our proposed measure is that we aim to measure predictability of frames. As we are working in a spectro-temporal domain, we see potential in measuring predictability of tiles, i.e., the frequency components that constitute frames. This is because distinct frequency components are necessarily not equally prominent for all sounds. In fact, our proposed measure estimates mutual information of a tile and spectro-temporal context, ultimately summing the estimates of all tiles in a frame leading to the per frame mutual information estimate.

In this thesis, we have solely considered predictability in terms of extrapolation. However, as described in [60, pp. 372-375], the human brain and auditory system is able to interpolate missing information rather than extrapolating it. A frame will certainly become easier to predict if information on both sides (temporally) is included in the context. Hence, adapting our proposed measure to become interpolating rather extrapolating, certainly seems interesting.

Finally, in Section 6.3, we examined whether frames characterized by high sound intensity are also characterized by high predictability. Results were computed on fairly small amounts of data. Hence, to further validate the results, the test should be conducted on a larger amount of data, preferably on multiple speech corpora to simulate different speaker conditions and environments. In such case, the predictor should be qualified to work on distinct speech corpora. Either a non-data-driven predictor could be used (one that is not tuned to the specific speech corpus as our CNN MMSE estimators are), or a neural network trained on data from multiple speech corpora could be used.

Bibliography

[1] GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet*, vol. 388, pp. 1545–1602, Oct. 2016.

[2] C. Elberling and K. Worsoe, *Fading Sounds - About Hearing and Hearing Aids*. The Oticon Foundation, 2006. ISBN: 87-991301-0-6.

[3] V. Cardin, "Effects of aging and adult-onset hearing loss on cortical auditory regions," *Frontiers in Neuroscience*, vol. 10, p. 199, 2016.

[4] F. R. Lin, K. Yaffe, J. Xia, Q. L. Xue, T. B. Harris, E. Purchase-Helzner, S. Satterfield, H. N. Ayonayon, L. Ferrucci, E. M. Simonsick, A. B. Newman, D. Ives, S. Satterfield, J. Elam, S. R. Cummings, M. C. Nevitt, S. M. Rubin, T. B. Harris, and M. E. Garcia, "Hearing loss and cognitive decline in older adults," *JAMA Intern Med*, vol. 173, pp. 293–299, Feb 2013.

[5] G. R. Popelka, B. C. J. Moore, R. R. Fay, and A. N. Popper, *Hearing Aids*. Springer, 2016. ISBN: 978-3-319-33036-5.

[6] P. C. Loizou, *Speech Quality Assessment*, pp. 623–654. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[7] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, second ed., 2017. ISBN: 978-1-138-07557-3.

[8] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[9] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 430–440, 2014.

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

[11] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, pp. 2224–37, 05 2005.

[12] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[13] H. Andersen, Asger, *Speech Intelligibility Prediction for Hearing Aid Systems*. PhD thesis, Aalborg University, 2017. ISBN: 978-87-7210-050-0.

[14] H.-W. Lee, K. Rayner, and A. Pollatsek, "The relative contribution of consonants and vowels to word identification during reading," *Journal of Memory and Language*, vol. 44, no. 2, pp. 189–205, 2001.

[15] J. Shimron, "The role of vowels in reading: A review of studies of english and hebrew," *Psychological Bulletin*, vol. 114, no. 1, pp. 52–67, 1993.

[16] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. USA: Prentice-Hall, Inc., 1993. ISBN: 0-13-015157-2.

[17] C. E. Stilp and K. R. Kluender, "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," *Proceedings of the National Academy of Sciences*, vol. 107, no. 27, pp. 12387–12392, 2010.

[18] B. C. Moore and B. R. Glasberg, "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," *Hearing Research*, vol. 28, no. 2, pp. 209 – 225, 1987.

[19] D. Kewley-Port, T. Burkle, and J. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 122, pp. 2365–75, 11 2007.

[20] R. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, pp. 853–856 vol. 2, 1996.

[21] K. Kluender, J. Coady, and M. Kiefte, "Sensitivity to change in perception of speech," *Speech Communication*, vol. 41, pp. 59–69, 08 2003.

[22] G. L. Dannenbring, "Perceived auditory continuity with alternately rising and falling frequency transitions," *Canadian journal of psychology*, vol. 30, no. 2, pp. 99–114, 1976.

[23] J. Taghia, R. Martin, and R. C. Hendriks, "On mutual information as a measure of speech intelligibility," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 65–68, March 2012.

[24] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, 2009. Visual Attention: Psychophysics, electrophysiology and neuroimaging.

[25] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," *IEEE International Conference on Intelligent Robots and Systems*, pp. 1173–1179, 09 2011.

[26] B. Schauerte and R. Stiefelhagen, ""wow!" bayesian surprise for salient acoustic event detection," *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 6402–6406, 10 2013.

[27] A. J. Oxenham, J. E. Boucher, and H. A. Kreft, "Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy," *The Journal of the Acoustical Society of America*, vol. 142, no. 3, pp. EL264–EL269, 2017.

[28] D. A. Bies and C. H. Hansen, *Engineering Noise Control: Thoery and Practice*. Spon Press, second ed., 2003. ISBN: 0-203-11665-8.

[29] M. Shelvock, "Audio mastering as musical practice," 2012. The University of Western Ontario.

[30] B. C. Moore, *An Introduction to the Psychology of Hearing*. Brill, sixth ed., 2013. ISBN: 978-90-04-25242-4.

[31] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Pearson, 2014. ISBN: 978-1-292-02572-8.

[32] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 3809–3812, May 2010.

[33] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," 1988.

[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, second ed., 2006. ISBN-13: 978-0-471-24195-9.

[35] P. Olofsson and M. Andersson, *Probability, Statistics, and Stochastic Processes*. Wiley, second ed., 2012. ISBN: 978-0-470-88974-9.

[36] M. Taboga, *Lectures on Probability Theory and Mathematical Statistics*. CreateSpace Independent Publishing Platform, second ed., 2012. ISBN-13: 9781480215238.

[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. ISBN: 978-0-521-83378-3.

[38] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," 1993. https://catalog.ldc.upenn.edu/LDC93S1, Philadelphia: Linguistic Data Consortium.

[39] C. E. Stilp, M. Kiefte, J. M. Alexander, and K. R. Kluender, "Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences.," *The Journal of the Acoustical Society of America*, vol. 128 4, pp. 2112–26, 2010.

[40] A. J. Spahr, M. F. Dorman, L. M. Litvak, S. Van Wie, R. H. Gifford, P. C. Loizou, L. M. Loiselle, T. Oakes, and S. Cook, "Development and validation of the AzBio sentence lists," *Ear Hear*, vol. 33, no. 1, pp. 112–117, 2012.

[41] V. Aubanel, M. Cooke, C. Davis, and J. Kim, "Temporal factors in cochlea-scaled entropy and intensity-based intelligibility predictions," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL443–EL448, 2018.

[42] A. Marsiglietti and V. Kostina, "A lower bound on the differential entropy of log-concave random vectors with applications," *CoRR*, vol. abs/1704.07766, 2017.

[43] S. Kay, *Intuitive Probability and Random Processes Using MATLAB*. Springer, 2006. ISBN: 978-0-387-24157-9.

[44] R. G. Brown, *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs, N.J., Prentice-Hall, 1963. ISBN 13: 9780138153083.

[45] S. Shamma, *Spectro-Temporal Receptive Fields*, pp. 1–6. New York, NY: Springer New York, 2013.

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

[47] 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*, vol. 15, pp. 1929–1958, 2014.

[48] M. Hansen and C. Ludvigsen, "Dantale II: Danske hagerman sætninger," 2001. https://audiologi.dk/wp-content/uploads/2011/05/Dantale-II-rapport1.pdf.

[49] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a danish sentence test in noise: Diseño, optimización y evaluación de la prueba danesa de frases en ruido," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[50] C. M. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, 1995. ISBN-13 : 978-0198538646.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[52] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," *CoRR*, vol. abs/1206.5533, 2012.

[53] P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial," *Proceedings of the IEEE*, vol. 78, pp. 56–93, Jan 1990.

[54] E. R. Pedersen and P. M. Juhl, "User-operated speech in noise test: Implementation and comparison with a traditional test," *International Journal of Audiology*, vol. 53, no. 5, pp. 336–344, 2014.

[55] G. Studebaker, "A "rationalized" arcsine transform," *Journal of speech and hearing research*, vol. 28, pp. 455–62, 10 1985.

[56] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[57] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 380–391, October 1976.

[58] K. Paliwal and B. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 3–14, Jan 1993.

[59] Y. Shu, X.-x. Feng, and F. Chen, "Comparing the perceptual contributions of cochlear-scaled entropy and speech level," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. EL517–EL521, 2016.

[60] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990. ISBN: 0-262-52195-4.

[61] J. P. Royston, "An extension of shapiro and wilk's w test for normality to large samples," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 31, no. 2, pp. 115–124, 1982.

Conditional Expectation Equals Minimum Mean Square Error Estimator

A

Theorem A.1 (Conditional Expectation Equals Minimum Mean Square Error Estimator) Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^N$ be random vectors. Then, the conditional expectation

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]$$

is equal to the MMSE estimator $\hat{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ given $\boldsymbol{X}$. ▲

Proof

Consider an estimator $\hat{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ as a function of $\boldsymbol{X}$. Then, the MSE of the estimator is given as

$$\mathbb{E}\left[\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)\middle|\boldsymbol{X}\right]\right],$$

by the law of total expectation.

First observe that if $\hat{\boldsymbol{Y}}$ minimizes $\mathbb{E}\left[\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)\middle|\boldsymbol{X}\right]$, then, the MSE $\mathbb{E}\left[\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)\right]$ is likewise minimized. Hence, to find an expression for $\hat{\boldsymbol{Y}}$ what minimizes the MSE, we differentiate $\mathbb{E}\left[\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)\middle|\boldsymbol{X}\right]$ with respect to $\hat{\boldsymbol{Y}}$ and equate to zero, to obtain

$$\begin{aligned} 0 &= \frac{\partial \mathbb{E}\left[\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)\middle|\boldsymbol{X}\right]}{\partial \hat{\boldsymbol{Y}}} \\ &= \frac{\partial \mathbb{E}\left[\boldsymbol{Y}^T\boldsymbol{Y}-2\hat{\boldsymbol{Y}}^T\boldsymbol{Y}+\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}\middle|\boldsymbol{X}\right]}{\partial \hat{\boldsymbol{Y}}}. \end{aligned} \tag{A.1}$$

By moving the differentiation inside the expectation, as the expectation is with respect to $\boldsymbol{Y}$, (A.1) becomes

$$\begin{aligned} 0 &= \mathbb{E}\left[\frac{\partial(\boldsymbol{Y}^T\boldsymbol{Y}-2\hat{\boldsymbol{Y}}^T\boldsymbol{Y}+\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}})}{\partial \hat{\boldsymbol{Y}}}\middle|\boldsymbol{X}\right] \\ &= \mathbb{E}\left[-2\boldsymbol{Y}+2\hat{\boldsymbol{Y}}|\boldsymbol{X}\right] \\ &= -2\mathbb{E}\left[\boldsymbol{Y}|\boldsymbol{X}\right]+2\mathbb{E}\left[\hat{\boldsymbol{Y}}|\boldsymbol{X}\right]. \end{aligned} \tag{A.2}$$

Finally, as the expectation is with respect to $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ is a function of $\boldsymbol{X}$, (A.2) becomes

$$0 = -2\mathbb{E}\left[\boldsymbol{Y}|\boldsymbol{X}\right] + 2\hat{\boldsymbol{Y}}. \tag{A.3}$$

Furthermore, as the second derivative of $\mathbb{E}\left[\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)\middle|\boldsymbol{X}\right]$ with respect to $\hat{\boldsymbol{Y}}$ is positive, the estimate $\hat{\boldsymbol{Y}}$ is a minimum [34, Th. 2.6.1]. From (A.3), it is evident that letting

$$\hat{\boldsymbol{Y}} = \mathbb{E}\left[\boldsymbol{Y}|\boldsymbol{X}\right],$$

minimizes the MSE, thus, concluding the proof.

■

Shapiro-Wilk Test B

The Shapiro-Wilk test for normality is used to test whether the ordered observation $\boldsymbol{y} \in \mathbb{R}^N$ is a sample from a Gaussian distribution with unknown mean and variance parameters. This appendix is based on [56, Sec. 2.2].

Let $\boldsymbol{X} \in \mathbb{R}^N$ denote an ordered random sample from $\mathcal{N}(0,1)$. This mean that $X[0] \leq X[1] \leq \cdots \leq X[N-1]$. Let

$$\boldsymbol{m} = \begin{bmatrix} \mathbb{E}[X[0]] & \mathbb{E}[X[1]] & \cdots & \mathbb{E}[X[N-1]] \end{bmatrix}^T$$

$$\boldsymbol{V} = \begin{bmatrix} \operatorname{Cov}[X[0],X[0]] & \operatorname{Cov}[X[0],X[1]] & \cdots & \operatorname{Cov}[X[0],X[N-1]] \\ \operatorname{Cov}[X[1],X[0]] & \operatorname{Cov}[X[1],X[1]] & \cdots & \operatorname{Cov}[X[1],X[N-1]] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}[X[N-1],X[0]] & \operatorname{Cov}[X[N-1],X[1]] & \cdots & \operatorname{Cov}[X[N-1],X[N-1]] \end{bmatrix}$$

denote the mean vector and covariance matrix of the ordered statistic $\boldsymbol{X}$, respectively with $\operatorname{Cov}[Z,Z']$ denoting the covariance between Z and Z'.

Next, let

$$\boldsymbol{a} = \frac{\left(\boldsymbol{m}^T \boldsymbol{V}^{-1}\right)^T}{\left(\boldsymbol{m}^T \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m}\right)^{-1/2}},$$

then, the Shapiro-Wilk test statistic W is defined as

$$W = \frac{\left(\sum_{n=0}^{N-1} a[n]\, y[n]\right)^2}{\sum_{n=0}^{N-1} \left(y[n] - \overline{\boldsymbol{y}}\right)^2}.$$

The value of the W statistic ranges from $\frac{Na^2[0]}{N-1}$ to 1, where values close to 1 represent high likelihood of $\boldsymbol{y}$ being a sample from a Gaussian distribution [56, Lemmas 2-3].

To compute critical values for the test statistic W, a transformation is applied to obtain a variable with an approximate Gaussian distribution [61, Sec. 4]. From this approximate Gaussian distribution, p-values can be obtained.