



Data-Driven Speech Intelligibility Prediction

Pedersen, Mathias

DOI (link to publication from Publisher):
[10.54337/aau532688918](https://doi.org/10.54337/aau532688918)

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

[Link to publication from Aalborg University](#)

Citation for published version (APA):
Pedersen, M. (2023). *Data-Driven Speech Intelligibility Prediction*. Aalborg Universitetsforlag. Ph.d.-serien for Det Tekniske Fakultet for IT og Design, Aalborg Universitet <https://doi.org/10.54337/aau532688918>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

DATA-DRIVEN SPEECH INTELLIGIBILITY PREDICTION

**BY
MATHIAS BACH PEDERSEN**

DISSERTATION SUBMITTED 2023



AALBORG UNIVERSITY
DENMARK

Data-Driven Speech Intelligibility Prediction

Ph.D. Dissertation
Mathias Bach Pedersen

Dissertation submitted March 3, 2023

Dissertation submitted: March 3, 2023

PhD supervisors: Prof. Jesper Jensen
Aalborg University and Demant A/S

Prof. Zheng-Hua Tan
Aalborg University

Prof. Søren Holdt Jensen
Chora A/S

Asger Heidemann Andersen
WS Audiology A/S

PhD committee: Professor Dorte Hammershøi (chairman)
Department of Electronic Systems
Aalborg University, Denmark

Professor Steven van de Par
Department of Medical Physics and Acoustics
Carl-von-Ossietzky University in Oldenburg, Germany

Professor Fei Chen
Department of Electrical and Electronic Engineering
Southern University of Science and Technology, China

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-733-8

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

© Copyright: Mathias Bach Pedersen, except where otherwise stated.

Printed in Denmark by Stibo Complete, 2023

About the Author

Mathias Bach Pedersen



Mathias B. Pedersen received the B.Sc. and M.Sc. degrees in Mathematical Engineering from Aalborg University, Aalborg, Denmark, in 2016 and 2018 respectively. He is currently pursuing his phd degree at Aalborg University, Denmark with the section of AI and Sound. His main research interests include machine learning, speech & signal processing, mathematical modelling and speech intelligibility prediction.

Abstract

Speech Intelligibility (SI) is a measure of the number of words in a speech signal that are understandable to a group of listeners. Measuring SI is time consuming, because it requires a test involving a panel of human listeners. Predicting SI algorithmically can instead provide estimates of the SI of speech signals significantly faster, which is highly valuable during, e.g., the development of speech communication or enhancement systems and devices. Traditionally, SI predictors use models of the human auditory system and signal features empirically demonstrated to correlate with SI. More recently, data-driven machine learning models have been trained using listening test data to perform SI prediction.

In this thesis we study data-driven SI prediction, and develop several data-driven SI predictors trained on listening test data. We identify that there is currently a critical scarcity of listening test data for the purpose of training data-driven SI predictors. This data scarcity motivates us to investigate and develop training strategies that make more efficient use of the available listening test data.

Specifically, we design and evaluate several data-driven SI predictors, trained using various quantities of listening test data. The evaluations of our first SI predictors help identify that there is a scarcity of listening test data, and that data-driven SI predictors consequentially do not generalize well beyond their training conditions. In conjunction with the development of our later data-driven SI predictors, we investigate strategies to mitigate and circumvent the problems caused by the listening test data scarcity. First, we use a hybrid data-driven and non-data-driven model, which is able to reach good prediction performance with fewer trainable parameters. Secondly, we append a layer of listening test dataset specific logistic mapping functions to a data-driven SI predictor, which allows pooling of heterogeneous listening test datasets. Finally, we train a neural network to estimate Speech Presence Probability (SPP) rather than SI, which requires no listening test data, but only speech and noise data, which is more abundantly available. Subsequently, we map the estimated SPP to SI via a relatively simple algorithm, which achieves good performance even for unseen listening conditions.

Resumé

Taleforståelighed (TF) er et mål for antallet af ord i et talesignal der er forståelige for en gruppe lyttere. Det er tidskrævende at måle TF, fordi det kræver en test der involverer et panel af menneskelige lyttere. I stedet kan algoritmisk prædiktion af TF give estimer af talesignalers forståelighed væsentligt hurtigere, hvilket er meget værdifuldt under, f.eks., udvikling af systemer og enheder til talekommunikation eller -forbedring. Traditionelt set har TF prædiktorer anvendt modeller af den menneskelige høreelse samt signaltræk, som er demonstreret empirisk at korrelere med TF. I senere tid er data-drevne maskinlærings modeller blevet trænet på lyttetestdata til at udføre TF prædiktion.

I denne afhandling undersøger vi data-drevet TF prædiktion, og udvikler adskillige data-drevne TF prædiktorer trænet på lyttetestdata. Vi identificerer at der er en nuværende kritisk mangel på lyttetestdata til brug i træning af data-drevne TF prædiktorer. Denne datamangel motiverer os til at undersøge og udvikle træningsstrategier der gør mere effektiv brug af den tilgængelige lyttetestdata.

Specifikt designer, træner og evaluerer vi adskillige data-drevne TF prædiktorer, trænet på variable mængder af lyttetestdata. Evalueringerne af vores første TF prædiktorer hjælper os med at identificere at der er en mangel på lyttetestdata, og at data-drevne TF prædiktorer som konsekvens ikke generaliserer godt udover deres træningsbetingelser. I forbindelse med udviklingen af vores senere TF prædiktorer undersøger vi strategier til at modkæmpe og omgå problemerne, der følger af manglen på lyttetestdata. Først bruger vi en data-drevet og ikke-data-drevet hybridmodel, der er i stand til at opnå god prædiktionssevne med færre trænbare parametre. Dernæst tilføjer vi et lag af lyttetest datasætafhængige logistiske afbildningsfunktioner til en data-drevet TF prædiktor, hvilket tillader sammenlægning af heterogene lyttetest datasæt. Endelig træner vi et neuralt netværk til at estimere Taletilstedeværelsessandsynlighed (TTS) i stedet for TF, hvilket ikke kræver lyttetestdata, men kun tale- og støjdata, som er mere rigeligt tilgængeligt. Efterfølgende afbilder vi de estimerede TTS til TF via en relativt simpel algoritme, der opnår god prædiktionssevne selv for usete lytteforhold.

Contents

About the Author	iii
Abstract	v
Resumé	vii
Abbreviations	xiii
List of Publications	xv
Preface	xvii
I Introduction	1
1 Speech Communication and Intelligibility	3
1.1 Speech Production, Transmission and Perception	4
1.1.1 Speech Production	4
1.1.2 Speech Transmission	5
1.1.3 Speech Perception	6
1.2 Speech Intelligibility	6
1.3 Speech Intelligibility Prediction	7
2 Neural Network Architectures for SIP	11
2.1 Neural networks	11
2.1.1 Basic Neural Network Architectures	11
2.1.2 Composite Neural Network Architectures	15
2.1.3 Neural Network Training	16
2.2 Training Data Scarcity for Data-Driven SIP	17
3 Speech Intelligibility Predictors	21
3.1 Categorization of SI Predictors	22
3.1.1 Comparative Measures	22

3.1.2	Signal Domain	23
3.2	Non-Data-Driven SI Predictors	28
3.2.1	SNR Based SI Predictors	28
3.2.2	Correlation Based SI predictors	31
3.2.3	Mutual Information Based SI Predictors	33
3.3	Data-Driven SI Predictors	34
3.3.1	SNR Based SI Predictors	35
3.3.2	Correlation Based SI Predictors	35
3.3.3	Learned Comparison Based SI Predictors	37
4	Scientific Contributions	39
4.1	Specific Contributions	39
4.1.1	[A] A Neural Network for Monaural Intrusive Speech Intelligibility Prediction	39
4.1.2	[B] End-to-End Speech Intelligibility Prediction using Time-Domain Fully Convolutional Neural Networks . .	40
4.1.3	[C] Training Data-Driven Speech Intelligibility Predic- tors on Heterogeneous Listening Test Data	40
4.1.4	[D] Data-Driven Speech Presence Probability Estima- tion for Non-Intrusive Speech Intelligibility Prediction .	41
4.2	Summary of Contributions	42
4.3	Directions of Future Research	43
4.3.1	Crowdsourcing SI Data	43
4.3.2	Applying SI Predictors to Speech Enhancement	43
4.3.3	Applying Data-Driven SIP in Portable Devices	43
	References	44

II Papers 55

A	A Neural Network for Monaural Intrusive Speech Intelligibility Pre- diction	57
1	Introduction	59
2	Neural Network SI-Predictor	61
2.1	Preprocessing	61
2.2	Architecture	61
3	Network Training	63
4	Listening Test Data	64
5	Results	65
5.1	Training Details and SIP Performance	65
5.2	Long-Term SIP Performance	66
5.3	Short-Term SIP Performance	66
6	Conclusion	67

References	68
B End-to-end Speech Intelligibility Prediction Using Time-Domain Fully Convolutional Neural Networks	71
1 Introduction	73
2 Data-driven Intelligibility Prediction	74
2.1 Intrusive Speech Intelligibility Prediction	74
2.2 Neural Speech Intelligibility Prediction	75
3 Experimental Design	77
3.1 Training, Validation and Test Data	77
3.2 Cross Validation	77
4 Experimental Results	78
4.1 End-to-end Data-driven Intelligibility Prediction	78
4.2 Data-driven vs. Non-data-driven SIP	79
4.3 Frequency-domain Data-driven SIP	81
5 Conclusion	81
References	81
C Training Data-Driven Speech Intelligibility Predictors on Heterogeneous Listening Test Data	85
1 Introduction	87
2 Related work	90
3 Architecture and mapping functions	92
3.1 Network design	93
3.2 1/3 Octave band transform	94
3.3 CNN layers	94
3.4 ESTOI back-end	94
3.5 Dataset-specific mapping functions	96
4 Dataset description and training procedure	97
4.1 Datasets	97
4.2 Training	98
4.3 Network parameters	99
5 Performance evaluation	99
5.1 Experiment A	101
5.2 Experiment B	105
6 Conclusion	109
References	109
D Data-Driven Non-Intrusive Speech Intelligibility Prediction using Speech Presence Probability	115

Contents

List of Abbreviations

ADFD	Akustiske Databaser For Dansk
AI	Articulation Index
ASR	Automatic Speech Recognition
BBL	Babble
BiSIM	Binaural SI Model
BSTOI	Binaural STOI
CNN	Convolutional Neural Network
CSII	Coherence SI Index
DNN	Deep Neural Network
DSMF	Dataset-Specific Mapping Function
ESII	Extended SI Index
ESTOI	Extended Short-Time Objective Intelligibility
FC	Fully Connected
FNN	Fully connected Neural Network
GMM	Gaussian Mixture Model
HASPI	Hearing Aid Speech Perception Index
HLLR	HMM-based Log Likelihood Ratio
HMM	Hidden Markov Model
IBM	Ideal Binary Mask
ITFS	Ideal Time-Frequency Segregation
KLT	Karhunen-Loève Transform
KNN	K-Nearest Neighbour
LSTM	Long Short-Term Memory
ModA	Modulation Area
MOSA-Net	Multi Objective Speech Assessment Network
MSE	Mean Squared Error
NISA	Non-Intrusive Speech Assessment
NN	Neural Network
NISTOI	Non-Intrusive STOI
NORI	No Reference Intelligibility
NSIP	Neural SI Predictor
PB-STOI	Pitch-Based STOI

List of Abbreviations

PESQ	Perceptual Evaluation of Speech Quality
PReLU	Parametrized ReLU
ReLU	Rectified Linear Unit
ResNet	Residual Network
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SDR	Signal-to-Distortion Ratio
sEPSM	speech-based Envelope Power Spectrum Model
SI	Speech Intelligibility
SII	SI Index
SIIB	SI In Bits
SIMI	SI prediction based on Mutual Information
SIP	SI Prediction
SNR	Signal-to-Noise Ratio
SPP	Speech Presence Probability
SRMR	Speech-to-Reverberation Modulation energy Ratio
SSN	Speech Shaped Noise
STFT	Short-Time Fourier Transform
STGI	Spectro-Temporal Glimpsing Index
STMI	Spectro-Temporal Modulation Index
STI	Speech Transmission Index
STOI	Short-Time Objective Intelligibility
STOINET	STOI Network
THMMB-STOI	Twin Hidden Markov Model-Based STOI
VAD	Voice Activity Detector
wSTMI	weighted STMI

List of Publications

The main body (Part II) of this dissertation consists of the following publications:

- [A] M. B. Pedersen, A. H. Andersen, S. H. Jensen and J. Jensen, “A neural network for monaural intrusive speech intelligibility prediction,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 336–340, 2020.
- [B] M. B. Pedersen, M. Kolbæk, A. H. Andersen, S. H. Jensen and J. Jensen, “End-to-end speech intelligibility prediction using time-domain fully convolutional neural networks,” in *Proceedings of Interspeech*, pp. 1151–1155, 2020.
- [C] M. B. Pedersen, A. H. Andersen, S. H. Jensen, Z. -H. Tan and J. Jensen, “Training data-driven speech intelligibility predictors on heterogeneous listening test data,” *IEEE Access*, Vol. 10, pp. 66175–66189, 2022.
- [D] M. B. Pedersen, S. H. Jensen, Z. -H. Tan and J. Jensen, “Data-driven non-intrusive speech intelligibility prediction using speech presence probability,” *submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

List of Publications

Preface

This thesis documents the scientific work carried out as part of the PhD project entitled *Data-Driven Speech Intelligibility Prediction and Listening Test Data Scarcity*. The project was funded by the Independent Research Fund Denmark¹, and carried out between September 2018 and January 2023 within the Artificial Intelligence and Sound Section of the Department of Electronic Systems at Aalborg University. Parts of the work was carried out in collaboration with Oticon A/S, Smørum, Denmark. This collaboration was mainly in the form of on-line meetings and discussions as well as feedback on written draft manuscripts due to the COVID-19 pandemic.

The thesis is structured in two parts: Part I contains a general introduction, and Part II contains a collection of research papers. More specifically, Part I introduces the field of speech intelligibility prediction as well as the field of deep learning. The scientific contributions of the PhD project are also summarized in Part I. Part II contains a collection of four research papers published in or submitted to peer-reviewed conferences and journals.

I would like to thank my supervisors, Jesper Jensen, Søren Holdt Jensen and Zheng-Hua Tan for their dedication to the project, even through several extensions. Also, for their support and expert guidance at every stage of the project. I would also like to thank Asger Heidemann Andersen for excellent collaboration, fruitful discussions and support. Thanks to Morten Kolbæk for collaboration and sparring. Thanks to all my wonderful friends who helped me stay in high spirits. Finally, I would like to extend a very special thanks to my family for your unconditional support.

Mathias Bach Pedersen
Aalborg University, March 3, 2023

¹<https://dff.dk/>

Preface

Part I

Introduction

Chapter 1

Speech Communication and Intelligibility

Language, particularly spoken language, is the primary human means of interpersonal communication. Using language and speech, we are able to share ideas and abstract thoughts and make ourselves understood by our peers. Speech communication is indispensable in modern society, and speech plays a role in almost every facet of our lives, as we use it in our jobs, cooperating with colleagues and interacting with customers or clients. We also use speech for comfort, chatting and connecting with friends or family. A great deal of entertainment and news is delivered to us in the form of speech through television and the Internet. We are also able to speak with other people over long distances by means of telephones or voice over Internet protocol.

The speech communication process can be deteriorated at different stages. In particular, deterioration may be due to the speech production of the talker, e.g., dialect, accent or pronunciation. It may also be due to the medium of speech transmission between talker and listener, such as wired/wireless transmission, signal processing, loudspeaker or hearing aid reproduction of the speech, background noise, reverberation, etc. Finally, speech communication may be deteriorated due to impairment, e.g., of the listeners hearing or cognitive ability.

There are at least three relevant aspects relating to the deterioration of speech communication: Speech intelligibility (SI), speech quality and listening effort. SI is the focus of this dissertation, and it describes how understandable speech is to human listeners on average. SI is measured by way of listening tests, where listeners are asked to reproduce or identify as many words as possible, when listening to deteriorated speech. Speech quality [108] is measured in listening tests in much the same way as SI, but instead of reproducing the words, listeners are asked to rate the quality of the test signals on

a scale from one to five, typically according to how pleasant or annoying the signal is to listen to due to noise and/or distortion. It is interesting to note that SI and speech quality do not necessarily correlate, as low quality speech signals can be highly intelligible [56, 77, 88]. Listening effort is a measure of how much effort a listener must expend in order to pay attention to a given speech signal [120]. Listening effort can be measured using, e.g., observation of pupil dilation [139] or electroencephalograms [1] of the listener, asking the listener to report their own perceived listening effort, etc.

1.1 Speech Production, Transmission and Perception

Speech communication can be modelled by three distinct phases, i.e., speech production, transmission and perception [40, 131]. A basic understanding of the physical systems involved in the communication of speech, and the challenges that arise in each of these phases is helpful in further identifying the factors that determine SI. We will begin by briefly describing each of these phases individually in the following.

1.1.1 Speech Production

Speech production begins with the conception of a message in the brain of the speaker. This message is encoded in the chosen language of the speaker, and the physiological production of speech can begin. To produce a voiced sound, the flow of air from the lungs is constricted by the closing of the vocal folds, which periodically open as air pressure rises [131, Chapter 1] [24] [40, Chapter 2]. The frequency at which the vocal folds open gives rise to the fundamental frequency and harmonics of the voiced sound. The power of these harmonics is shaped spectrally as the sound passes through the mouth and nasal cavity. The spectral shaping can be described in large part by formants, i.e., the peaks in the power spectral density [131, Chapter 3] [40, Chapter 2]. These formants are the result of resonance and thus depend on the shape of the mouth and nasal cavity, the relative position of the tongue and teeth, etc. Unvoiced sounds, on the other hand, are produced in a variety of ways including, for example, continuous noise-like sound resulting from constricted air flow at various points in the mouth, like “F”, “H” and “S”, or a sudden release of built-up air pressure at various positions in the mouth, like “B”, “K”, “P” and “T” [131, Chapter 7] [40, Chapter 2].

1.1.2 Speech Transmission

The basic form of speech transmission is the propagation of sound waves through the air, as in a traditional face-to-face conversation. In this scenario, the communication may be deteriorated by environmental factors, i.e., noise or interfering signals. Background noise is a common problem caused by, e.g., man-made noise pollution such as traffic or natural noise such as wind. Depending on the Signal-to-Noise Ratio (SNR) the speech or parts thereof may be masked by the noise [9, 38, 42, 91] [94, Chapter 3], and become imperceptible to the listener. Interfering signals may come in the form of unrelated speech from competing talkers [15, 27], or reverberation due to sound reflection on surfaces in the environment [18, 28, 95, 111]. Similarly to noise, these interfering signals may have a masking effect on the relevant speech, but can deteriorate SI more than noise of equivalent power [51, 77].

Today, speech is very commonly transmitted electronically as well as acoustically. Electronic transmission of speech typically involves at least a microphone and loudspeaker, and the quality of either can affect the speech signal. Additionally, depending on the situation, the speech signal may also be processed in some manner, for instance by using a statistical model to remove unrelated noise while preserving the speech content [10, 33, 49, 125, 132], or a neural network trained to enhance noisy speech signals [48, 98, 136]. Although processing is often meant to enhance or facilitate speech communication, it can also introduce new challenges in the form of distortions. Distortion of speech due to digital transmission or processing can take many forms including enhancement errors, such as under-suppression of noise, over-suppression of speech, and processing artefacts, i.e., sounds introduced by processing that were not present originally [19, 24, 86]. In telephony or voice over Internet protocol, transmitted speech is compressed in order to reduce the required bandwidth, which commonly introduces artefacts [56, 87]. The transmitted speech may also be subject to de-noising or de-reverberation as well [33, 62, 74, 124].

For users of hearing assistive devices, such as hearing aids, speech is transmitted by both a purely acoustic path and a partly electronic path in parallel. In the electronic path, i.e., the sounds detected and reproduced by the hearing aid, the dynamic range may be altered, shifting and amplifying certain frequencies to compensate for impaired hearing at other, typically higher, frequencies [107, Chapter 13]. Feedback cancellation is also commonly employed in hearing aids due to the close proximity of microphones and loudspeakers. Finally, algorithms to reduce the impact of noise, including beamformers, de-reverberation algorithms, noise reduction algorithms, and speaker separation algorithms, designed to enhance the speech for the listener [48, 98], are also common.

1.1.3 Speech Perception

Speech perception is perhaps the most complex stage of speech communication. Perception begins when the sound wave enters the ear canal, setting the eardrum in motion in concordance with the sound pressure of the speech. The eardrum is connected to a set of three bones, the ossicles, in the middle ear, regulating the strength of the physical vibrations before they are transmitted to the cochlea in the inner ear [106, Chapter 3] [107, Chapter 4]. The cochlea consists of three oblong, coiled, fluid-filled chambers separated by membranes. Rows of outer and inner hair cells, located along one of these membranes, the basilar membrane, play a central role in the transduction from physical vibrations into neural signals [106, Chapter 3] [107, Chapter 4]. The resonance frequency changes along the length of the cochlea due to its tapering shape and stiffness gradient [107, Chapter 5]. This allows the cochlea to perform a decomposition of the frequency content of the incoming sound wave into spatially separated resonances along the basilar membrane [106, Chapter 3] [94, Chapter 3]. The resonances along the cochlea are converted to neural signals through the vibrations induced in the row of inner hair cells on the basilar membrane, and transmitted via the auditory nerve to the brain, where they can be interpreted as perceived sounds and decoded [106, Chapter 4] [107, Chapter 4].

Hearing impairments can make it very difficult, for those affected, to understand speech, particularly if noise or distortion is present [9, 13, 39] [107, Chapter 13]. Hearing impairments can roughly be categorized as either conductive or sensorineural [106, Chapter 10]. Conductive impairments are those that cause the speech signal to be degraded before it reaches the inner ear, whereas sensorineural impairments are the result of reduced function in the cochlea or the auditory nerves beyond. Sensorineural impairments are the most common [107, Chapter 13], and are often caused by damage to the outer hair cells in the cochlea, which can happen due to age, genetic factors or exposure to loud sounds [106, Chapter 10].

1.2 Speech Intelligibility

The main focus of this dissertation is Speech Intelligibility (SI). SI is a measure defined as the proportion or rate of words in a speech signal that are correctly perceived by listeners. Hence, quantitatively, SI is a scalar in the range between zero and one. Functionally, the SI of a specific speech signal is the number of words an average listener is able to understand or reproduce divided by the total number of words in that same signal. It can be of, e.g., scientific or commercial interest, to measure the SI of speech in specific noise and/or processing conditions. As an example, developers of hearing assistive technology, such as head-sets, hearing aids etc. designed to assist people

1.3. Speech Intelligibility Prediction

with perception of speech, would be interested in measuring how well their technology performs in certain situations, say at a dinner party or in a car cabin.

The SI in these situations can be measured by carrying out a listening test, with specific samples of speech under the conditions of interest [91]. A listening test consists of a number of test signals, i.e., speech signals under a specific set of test conditions. During the test, listeners are exposed to the test signals one-by-one and asked to either identify or reproduce the words in each signal. This can be done in different ways. Sometimes the listeners are given lists of candidate words to choose from, which is known as a closed vocabulary test, such as those found in [3, 46, 59]. Otherwise they are asked to write down or repeat verbally the words that they hear, in what is called an open vocabulary test, such as those found in [77, 86]. The correctly identified words are recorded for the purpose of computing SI. This test is repeated for several listeners, who are either exposed to the same test signals, or to different test signals under the same test conditions. The listeners should ideally have identical listening capabilities, e.g., normal hearing or similar level of hearing impairment. Once the test is concluded, the average of the test scores within each test condition can be taken across all participating listeners, yielding the SI measurements for each condition [41]. Note that listening tests may sometimes be scored based on other speech structures than words, such as phonemes or entire sentences. For the purposes of this dissertation, and in most of the listening tests described in the articles enclosed in Part II, SI is defined based on words.

An SI score from a listening test can be interpreted in a number of ways. Obviously, because of how the test is designed, it can be interpreted as the proportion of words in a specific speech signal that listeners will be able to understand on average. Given multiple listeners and signals under the same conditions, this interpretation can be extrapolated to the expected proportion of words a listener will be able to reproduce or understand on average in the given acoustic condition. If speech under the given condition is considered as a communication channel transmitting words from talker to listener, then the SI can be considered the success rate of this channel.

1.3 Speech Intelligibility Prediction

Because of the fact that human participants are required to perform a listening test, they are very time consuming in terms of organization, preparation and conducting. This means that it is not practical to carry out a listening test in every situation where SI measurements are desired, e.g., in the iterative development of algorithms or devices with a focus on improving SI. This problem is the focus of the scientific field of Speech Intelligibility Prediction

(SIP), which is concerned with the development of objective methods, algorithms, that can predict the outcome of a listening test, or more specifically, the SI of a speech signal. Ideally, the prediction produced by an SI predictor is strongly correlated with the SI, as measured in an actual listening test. Two categories of SI predictors are defined by the types of input signals they are given. In particular, so-called *intrusive* SI predictors are given a potentially noisy and/or processed speech signal, along with either a clean reference version of the same speech signal, or the pure noise signal in the case of speech in additive noise, in isolation. Intrusive SI predictors can be applied as a replacement for listening tests in the development process of speech processing algorithms or devices [62, 124], where a clean speech or pure noise signal is available. The other category is *non-intrusive* SI predictors, which are only given the noisy and/or processed speech signal as input [35]. Non-intrusive SI prediction must be expected to be a harder problem than the intrusive counterpart, simply because of the reduction in available information, which may also be the reason why significantly fewer non-intrusive SI predictors have been proposed than intrusive ones, at least in the *non-data-driven* context, cf. Section 3.2. The range of applications for non-intrusive SI predictors extends to situations, where a clean reference is not available. Notably, it is possible to incorporate non-intrusive SI predictors into speech enhancement systems, enabling in principle, portable devices to optimize or adapt their speech processing algorithms according to predicted SI.

In SIP, and other fields where human responses to a test are linked to a physical quantity, psychometric functions arise [89]. A psychometric function is the function that maps human responses in a given test to the quantity that is being measured. In the case of listening tests and SI, the responses are the word scores and the measured quantity is SI. It has been shown that the psychometric functions related to the SI of speech in additive noise are monotonically increasing and s-shaped, relative to the SNR [84]. When it comes to SIP, there is also a psychometric function relating predictions of SI to measurements of SI from a specific listening test [90]. The predictions produced by an SI predictor are generally not identical to and often not even a good estimate of the *absolute* measurements of SI obtained in a listening test. Since absolute SI is indexed by the predicted SI via a function, i.e., the psychometric function, the SI predictions are termed SI indices. The monotonicity of the psychometric function means that an increase in an SI index corresponds to an increase in absolute SI.

The psychometric function of a given listening test depends on a large number of variable factors other than the physical speech signal, collectively referred to as the listening test *paradigm*. The paradigm includes, e.g., language, talkers, lexical redundancy, sentence structure, vocabulary, equipment etc. SI predictors do not have access to the paradigm, which is why the psychometric function must be applied to the SI predictions in order to map

1.3. Speech Intelligibility Prediction

them to absolute SI. The only way to find the psychometric function, however, is through regression based on listening test data, which means that this is only possible if a listening test is performed. Thus, a fundamental limitation of SI indices is that, while they can always be compared within the same paradigm, cross-paradigm comparisons are only meaningful, if the psychometric functions of both paradigms are known, i.e., if listening tests have been performed.

Chapter 1. Speech Communication and Intelligibility

Chapter 2

Neural Network Architectures for SIP

In Chapter 3 and Part II we describe a number of *data-driven* SI predictors. These data-driven SI predictors rely on large parametrized models, Neural Networks (NN), optimized using large datasets. In this chapter we describe variations of neural network architectures that are commonly used for data-driven SI prediction, including Fully connected Neural Networks (FNN), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). We also describe the more specific U-Net and Residual Network (ResNet) architectures, which have been used for data-driven SIP in the research papers presented in Part II of this thesis.

2.1 Neural networks

2.1.1 Basic Neural Network Architectures

Fully Connected Neural Networks

The basic building block of neural networks is the artificial neuron [44, Chapter 6], which consists of an affine and a piece-wise differentiable, non-linear function

$$y = \sigma(\mathbf{w}^\top \mathbf{x} + b), \quad (2.1)$$

where lower-case boldface letters represent column vectors, $\mathbf{x} \in \mathbb{R}^n$ is the input vector, $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are the weights and bias, respectively, of the affine function, $\sigma(\cdot)$ is a non-linear function called the activation function and $y \in \mathbb{R}$ is the output. Artificial neurons are combined, in parallel, to form

a layer of a neural network:

$$\mathbf{y} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}), \quad (2.2)$$

where upper-case boldface letters represent matrices, $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_n] \in \mathbb{R}^{m \times n}$ is a matrix of weights, $\mathbf{y} \in \mathbb{R}^m$ is the output vector, $\mathbf{x} \in \mathbb{R}^n$ is the input vector, $\mathbf{b} \in \mathbb{R}^m$ is the vector of biases and the activation function σ is applied element-wise. Applying layers like these in sequence makes up an FNN architecture [52, 53, 117],

$$\mathbf{y}^{\{l\}} = \sigma(\mathbf{W}^{\{l\}} \mathbf{y}^{\{l-1\}} + \mathbf{b}^{\{l\}}), \quad (2.3)$$

where the superscript $\{l\}$ indicates that the superscripted variable belongs to the l 'th layer of the neural network, with $l = 1, \dots, L$ and $\mathbf{y}^{\{0\}} := \mathbf{x}$. The dimensions, $m^{\{l\}} \times n^{\{l\}}$, of $\mathbf{W}^{\{l\}}$ generally change from layer to layer, subject to the restriction that $m^{\{l\}} = n^{\{l+1\}}$.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) differ from FNNs by replacing the affine part of (2.2) with linear convolutions [8]. For example, a two-dimensional linear convolution operation denoted by the $*$ operator can be defined as

$$(\mathbf{F} * \mathbf{G})[j, k] = \sum_{m=0}^{m_G} \sum_{n=0}^{n_G} \mathbf{F}[m+j, n+k] \mathbf{G}[m, n], \quad (2.4)$$

where $\mathbf{F} \in \mathbb{R}^{m_F \times n_F}$ and $\mathbf{G} \in \mathbb{R}^{m_G \times n_G}$ are matrices with $m_F \geq m_G$ and $n_F \geq n_G$ and $j = 0, \dots, m_F - m_G$, $k = 0, \dots, n_F - n_G$. Note that it is possible to change the boundary conditions of the convolution by padding \mathbf{F} with zeros. Also note that the linear convolution operation can easily be generalized to tensors with a different number of coordinates, by introducing an additional sum over each new coordinate, or removing one sum in the case of one-dimensional linear convolution. A simple CNN can be constructed as a series of L convolutional layers

$$\mathbf{Y}_c^{\{l\}} = \sigma(\mathbf{X}^{\{l\}} * \mathbf{W}_c^{\{l\}} + b_c), \quad (2.5)$$

where $l = 1, \dots, L$ denotes the layer, $\mathbf{X} \in \mathbb{R}^{m_x \times n_x}$ is the input matrix, e.g., a part of a spectrogram, and $\mathbf{W}_c \in \mathbb{R}^{m_w \times n_w}$ is the c 'th convolutional kernel, i.e., a small matrix of weights that are shared across the input signal as a result of the convolution operation, σ is the non-linear, piece-wise differentiable activation function applied element-wise and b_c is the bias corresponding to the c 'th kernel. Several kernels are used in parallel in a single convolutional layer, as indexed by the subscript c . Conveniently, the individual convolutions with each kernel in one convolutional layer can be performed at the same time by

2.1. Neural networks

a single convolution operation with one additional coordinate indexing the kernel number, c . CNN's are more parameter efficient than FNN's as the convolutional kernels are smaller than the input signal. Although CNN's are only 'locally' connected for any single layer, as described by (2.5), they are still able to connect more distant entries of input signals indirectly over multiple layers. Convolutional layers can be made even more efficient by applying stride. Stride can be defined as a down sampling of the convolution, i.e., a stride of s corresponds to only using every s 'th entry along each axis of $(\mathbf{X} * \mathbf{W})$

$$(\mathbf{X} * \mathbf{W})_s[u, v] = (\mathbf{X} * \mathbf{W})[sj, sk], \quad (2.6)$$

where $u = 0, \dots, \lfloor (m_{\mathbf{X}} - m_{\mathbf{W}})/s \rfloor$, $v = 0, \dots, \lfloor (n_{\mathbf{X}} - n_{\mathbf{W}})/s \rfloor$ and u is a positive integer.

In addition to convolutional layers, pooling layers are often employed in CNNs. A pooling layer is designed to compress information in a matrix or tensor, by way of a simple function, typically a maximum or average, applied locally. Max pooling layers are very commonly used in CNNs

$$\text{maxpool}_{m_p \times n_p}(\mathbf{F})[j, k] = \max(\mathbf{F}[j : j + m_p - 1, k : k + n_p - 1]), \quad (2.7)$$

where m_p and n_p denote the dimensions of the pooling operation, the maximum function, $\max(\mathbf{M})$, returns the largest entry of the matrix \mathbf{M} , and $\mathbf{F}[j : j + m_p - 1, k : k + n_p - 1] \in \mathbb{R}^{m_p \times n_p}$ is the sub matrix of $\mathbf{F} \in \mathbb{R}^{m_{\mathbf{F}} \times n_{\mathbf{F}}}$ that contains the j 'th through $(j + m_p - 1)$ 'th rows and the k 'th through $(k + n_p - 1)$ 'th columns. Max pooling layers with stride are used to reduce the number of samples in the input without introducing additional weights. Due to the exploitation of local connections, CNN's have proven effective for problems involving audio and visual data in particular [44, Chapter 9] [80, 109].

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are designed for sequentially structured data, e.g., audio waveforms [44, Chapter 10]. RNNs are built around the concept of dynamic systems, i.e., series of random variables $\mathbf{x}[t] \in \mathbb{R}^n$ that each depend on the outcomes of previous variables in the series, $\mathbf{x}[t - 1], \mathbf{x}[t - 2], \dots$. An RNN models such a sequence with a corresponding sequence of hidden states

$$\mathbf{h}[t] = \sigma_{\mathbf{h}}(\mathbf{W}_{\mathbf{h}}\mathbf{h}[t - 1] + \mathbf{W}_{\mathbf{x}}\mathbf{x}[t] + \mathbf{b}_{\mathbf{h}}), \quad (2.8)$$

where $\mathbf{x}[t] \in \mathbb{R}^n$ is the t 'th entry of the input sequence, $\mathbf{h}[t - 1] \in \mathbb{R}^n$ is the vector of hidden states of the $(t - 1)$ 'th entry in the sequence, $\sigma_{\mathbf{h}}(\cdot)$ is the non-linear, piece-wise differentiable activation function applied element-wise that outputs the next state of the RNN, $\mathbf{h}[t] \in \mathbb{R}^n$, $\mathbf{W}_{\mathbf{h}} \in \mathbb{R}^{n \times n}$ and $\mathbf{W}_{\mathbf{x}} \in \mathbb{R}^{n \times n}$ represent the weights of the RNN and $\mathbf{b}_{\mathbf{h}} \in \mathbb{R}^n$ is the bias term [44, Chapter 10]. RNN's can be designed to produce different types of outputs depending

on the use case, e.g., an output for every entry in the input sequence, or only one output at the end of the sequence. For the purpose of illustration, the output of an RNN at entry t would be described as

$$\mathbf{y}[t] = \sigma_{\mathbf{y}}(\mathbf{W}_{\mathbf{y}}\mathbf{h}[t] + \mathbf{b}_{\mathbf{y}}), \quad (2.9)$$

where $\mathbf{y}[t] \in \mathbb{R}^m$ is the output of the RNN at the t 'th entry of the sequence, $\mathbf{W}_{\mathbf{y}} \in \mathbb{R}^{m \times n}$ represents the weights of the RNN, $\mathbf{b}_{\mathbf{y}} \in \mathbb{R}^m$ is the bias term and $\sigma_{\mathbf{y}}(\cdot)$ is the non-linear, piece-wise differentiable activation function applied element-wise [44, Chapter 10]. The weights and biases of an RNN are shared across all time-steps, which enables an RNN to work for variable-length input sequences, such as audio signals. The purpose of the hidden states in an RNN is similar to that of the hidden layers in an FNN, in that the hidden states carry and process information between the input and output of the RNN, but whereas a neuron in an FNN is connected to all neurons in the preceding and following layer, a hidden state of an RNN is connected only to the preceding and following hidden state in the sequence. Like CNNs, by exploiting natural structures in the input data RNNs require much fewer parameters than FNNs for the same size of input.

A popular RNN architecture, which is also used in state-of-the-art data-driven SIP, is the Long Short-Term Memory (LSTM) architecture, which builds on the RNN described by (2.8) and (2.9). The LSTM architecture introduces additional hidden states specifically designed to work as gated memory cells [50], i.e., hidden states with more control over when to accept new information, when to apply it, and when to discard it. The gated memory cells are defined by [50] as

$$\begin{aligned} \mathbf{f}[t] &= \sigma(\mathbf{W}_{\mathbf{h}_f}\mathbf{h}[t-1] + \mathbf{W}_{\mathbf{x}_f}\mathbf{x}[t] + \mathbf{b}_f) \\ \mathbf{i}[t] &= \sigma(\mathbf{W}_{\mathbf{h}_i}\mathbf{h}[t-1] + \mathbf{W}_{\mathbf{x}_i}\mathbf{x}[t] + \mathbf{b}_i) \\ \mathbf{o}[t] &= \sigma(\mathbf{W}_{\mathbf{h}_o}\mathbf{h}[t-1] + \mathbf{W}_{\mathbf{x}_o}\mathbf{x}[t] + \mathbf{b}_o) \\ \tilde{\mathbf{c}}[t] &= \sigma_c(\mathbf{W}_{\mathbf{h}_c}\mathbf{h}[t-1] + \mathbf{W}_{\mathbf{x}_c}\mathbf{x}[t] + \mathbf{b}_c) \\ \mathbf{c}[t] &= \mathbf{f}[t] \cdot \mathbf{c}[t-1] + \mathbf{i}[t] \cdot \tilde{\mathbf{c}}[t] \\ \mathbf{h}[t] &= \mathbf{o}[t] \cdot \sigma_{\mathbf{h}}(\mathbf{c}[t]), \end{aligned} \quad (2.10)$$

where the \cdot operator denotes the element-wise product between vectors, $\mathbf{c}[t]$ is a vector of memory cells, $\mathbf{f}[t] \in \mathbb{R}^n$ is a vector of so-called forget-gates designed to determine which cells of $\mathbf{c}[t-1] \in \mathbb{R}^n$ should be retained or forgotten in $\mathbf{c}[t]$, $\mathbf{i}[t] \in \mathbb{R}^n$ is a vector of so-called input-gates designed to determine which cells of $\mathbf{c}[t]$ should be updated based on the previous state $\mathbf{h}[t-1]$, $\mathbf{o}[t] \in \mathbb{R}^n$ is a vector of so-called output-gates designed to determine which states of $\mathbf{h}[t] \in \mathbb{R}^n$ should be updated based on the memory cells in $\mathbf{c}[t]$, $\tilde{\mathbf{c}}[t] \in \mathbb{R}^n$ is the input to the memory cells in $\mathbf{c}[t]$ controlled by $\mathbf{i}[t]$ and finally $\mathbf{h}[t]$ is the vector of hidden states. The output $\mathbf{y}[t] \in \mathbb{R}^m$ of an LSTM

is described by (2.9). The matrices $\mathbf{W} \in \mathbb{R}^{n \times n}$ with various subscripts as well as the vectors $\mathbf{b} \in \mathbb{R}^n$ with various subscripts in (2.10) are weights and biases of the gates and memory cells, which implies that the behaviour of the gates and cells is learned based on data. The gated memory cells allow LSTM's to retain information across many time steps more effectively than the basic RNN's described in (2.8) [50].

2.1.2 Composite Neural Network Architectures

U-Net

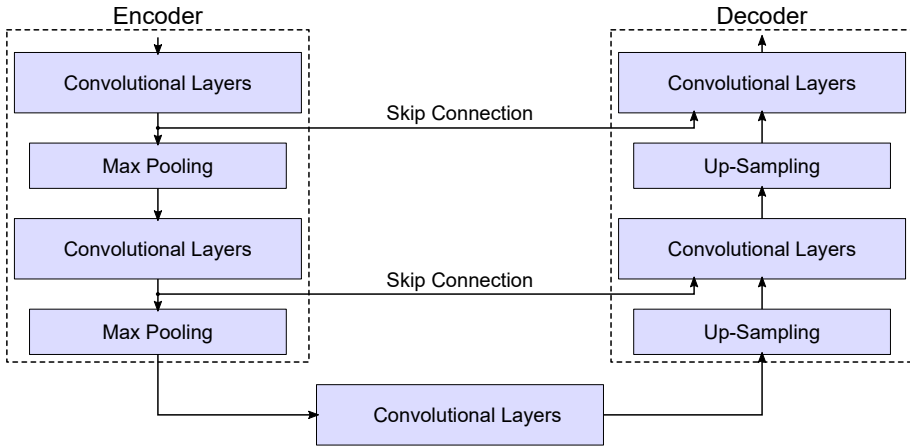


Fig. 2.1: The U-net architecture. Inputs are first passed through convolutional and max pooling layers in the encoder, reducing the dimensions of the input. The decoder then restores the inputs to their original dimensions via up-sampling and convolutional layers. Skip connections between corresponding stages of the encoder and decoder allow the decoder to access the input signal from the encoder directly.

U-net is a CNN architecture, sketched in Figure 2.1 and originally proposed in [116] as an end-to-end fully convolutional model for biomedical image segmentation. The architecture features an encoder and decoder, depicted on the left and right in Figure 2.1, respectively, linked together by skip connections. A skip connection is a connection in a neural network that bypasses one or more layers, and transmits data unmodified. The encoder consists of CNN layers and max pooling layers that gradually reduce the dimensions of the input, whereas the decoder consists of CNN layers and up-sampling to restore the input to the original dimensions. U-net was designed for and demonstrated to work well for scenarios with limited quantities of training data [116]. The U-net architecture has become widely used for segmentation tasks, i.e., partitioning grid-based data like images into segments belonging to different classes. The task could, for example, be to divide a

noisy/processed speech signal into intelligible and unintelligible segments, see Part II Paper B.

ResNet

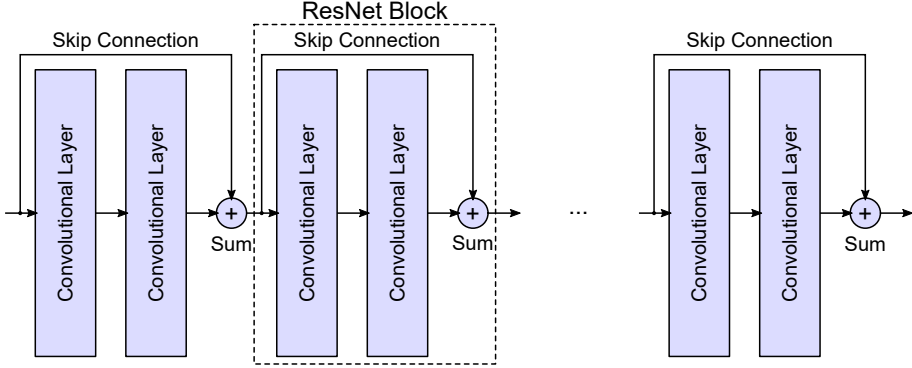


Fig. 2.2: The ResNet architecture. ResNet is designed as a series of small blocks. Each block is designed to compute a residual, i.e., an additive change to the input of the block. The skip connections bypassing each block of convolutional layers help prevent the vanishing gradient problem, and allow for deeper networks to be trained.

The Residual Network (ResNet) architecture, proposed in [47] and sketched in Figure 2.2, is a CNN architecture focused on enabling greater depth, i.e., a larger number of layers. ResNet is constructed as a series of individual blocks, as seen in Figure 2.2, each consisting of two CNN layers in parallel with a skip connection. In a ResNet block the output of the CNN layers is added to their original input, which means that the CNN layers are constrained to computing an additive change, known as a residual, to the input. This constraint makes it possible to successfully optimize deeper networks [47]. Deeper networks can be more powerful than shallower networks with the same number of parameters [47, 52, 53], but can also be harder to train due to what is known as the vanishing/exploding gradient problem [11, 44]. The problem is that the partial derivatives of the weights of a neural network tend to become either vanishingly small or exceedingly large in magnitude, when the number of layers is increased, due to the increased number of multiplicands given by the chain rule. In ResNet, however, due to the series of skip connections the partial derivatives contain a term which is a sum rather than a product, which prevents the vanishing gradient problem.

2.1.3 Neural Network Training

The values of the weights, $\mathbf{W}^{\{l\}}$, and biases, $\mathbf{b}^{\{l\}}$, $l = 1, \dots, L$ of neural networks are determined in a training phase. Often a supervised learning

2.2. Training Data Scarcity for Data-Driven SIP

paradigm is used, where a neural network is trained to solve a problem represented by a dataset of training samples $\{\tilde{\mathbf{x}}\}$ and the corresponding desired outputs known as labels $\{\tilde{\mathbf{y}}\}$. The training samples and labels are used in an optimization procedure to minimize a selected cost function, $d(\cdot)$, e.g., mean squared error, between the outputs of the neural network and the labels of the training samples

$$\{\mathbf{W}^l, \mathbf{b}^l\}_{l=1,\dots,L} = \underset{\{\mathbf{W}^l, \mathbf{b}^l\}_{l=1,\dots,L}}{\operatorname{argmin}} d(\mathbf{y}^{\{L\}}, \tilde{\mathbf{y}}), \quad (2.11)$$

where $\mathbf{y}^{\{L\}}$ is the output of the neural network and thus dependent on the weights and biases $\{\mathbf{W}^l, \mathbf{b}^l\}_{l=1,\dots,L}$. The optimization of the cost function is typically performed iteratively using a variation of gradient descent [118], which requires evaluating the neural network in the points given by the training samples (forward propagation), as well as computing the gradients, $\frac{\partial d(\mathbf{y}^{\{L\}}, \tilde{\mathbf{y}})}{\partial \mathbf{W}^{\{l\}}}$ and $\frac{\partial d(\mathbf{y}^{\{L\}}, \tilde{\mathbf{y}})}{\partial \mathbf{b}^{\{l\}}}$ (backward propagation). Since a neural network is a composite function, backward propagation involves computing partial derivatives by repeated application of the chain rule [119]. Forward and backward propagation is typically performed on small batches of the training data. Repeating this process, known as stochastic gradient descent, ideally leads the neural network to the neighbourhood of a minimum of the cost function [44, Chapter 6].

It is worth noting that the minimum of the cost function found through gradient descent is specific to the training data, and may not correspond to a low value of the cost function on new data. This can cause problems once the network is employed after the training phase, since the performance generally drops, relative to the training phase. In fact, since training data can be considered as a finite set of discrete points in a continuous space, a network with a sufficient number of trainable parameters may eventually find ways to lower the cost in these specific points very slightly, while greatly increasing the cost in other regions that might be relevant, but not represented in the training set. This phenomenon is known as overfitting [44, Chapter 7] and it is of particular concern when training data is scarce.

2.2 Training Data Scarcity for Data-Driven SIP

As mentioned in Section 1.3, conducting listening tests is a time consuming process, which makes it difficult to collect a sufficient quantity of data for the training of a large-scale data-driven SI predictor. We found in papers A and B, presented in Part II of this thesis, that even pooled collections of listening test datasets that were considered large from a non-data-driven point of view, were insufficient in size and diversity from a machine learning point

of view. Training data scarcity is not only a problem because of the limited quantity, but also because of the limited diversity in terms of listening conditions, i.e., types of noise and processing deteriorating the speech. Our research indicates that diversity is a very important aspect of the training data for data-driven SIP, because applying data-driven SIP algorithms to listening conditions not seen during network training can lead to a substantial performance drop, as we note in paper B in Part II. The conclusion is that there is a severe scarcity of listening test data for the purpose of data-driven SIP.

Overfitting is a common problem faced when training data-driven models on limited amounts of training data, which is certainly the case for data-driven SI prediction. The size of the model versus the size and variety of the training set is ultimately the deciding factor in the problem of overfitting. Sufficiently small models, relative to the size of the training data set, may not overfit, but also may not be able to model the nuances in the training data, causing the end-performance to drop as a result. On the other hand, networks that are large compared to the amount and variety of training data can use their excess parameters to gain very small improvements, specific to the data samples in the training set, at the cost of potentially much poorer performance on new data not represented in the training set. There are standard approaches in the field of machine learning that may be utilized to mitigate the effects of overfitting. Reducing the number of trainable parameters can prevent overfitting, by limiting the complexity the model is able to achieve, however this may also lead to lower performance at test time. Alternatively, early stopping can be employed, where a separate validation dataset is used to detect the point during training at which the model begins to overfit and halt the optimization. Dropout [44, Chapter 7], is another method that aims to constrain the model during training to limit its capacity to overfit, without reducing the number of parameters. Data augmentation is a technique where the existing training samples are perturbed in any number of ways, e.g., by rotating, mirroring, shifting, scaling etc. This can help mitigate overfitting through increased coverage of the desired space of input signals.

As mentioned, one can reduce the number of parameters to avoid overfitting when training data is scarce, but this may limit the expressive power of the model and negatively impact the achievable performance. In some cases, this problem can be mitigated through what is known as transfer learning, effectively standing on the shoulders of existing successful models. Transfer learning involves taking, as a starting point, an existing neural network that has been trained for a different task with similar input data, and re-training the final layers to perform the desired task instead. In the case of data-driven SIP one could make use of an architecture trained on a speech-processing task, where labelled data is not as scarce. For example, Automatic Speech Recognition (ASR) systems have been used in data driven SIP, which

is described in detail in Section 3.3. Alternatively hybrid data-driven and non-data-driven architectures can be used. So long as a non-data-driven series of computational steps consists of piece-wise differentiable functions, the chain rule can be applied and back-propagation is possible. Such a hybrid architecture is presented in Part II Paper A.

Another simple, yet quite popular approach to data-driven SIP, that attempts to circumnavigate the data scarcity, is to create new data-driven non-intrusive predictors by utilizing existing non-data-driven intrusive SI predictors to label training data. A speech dataset is labelled with intrusive predictions of SI, and subsequently used to train a machine learning model to non-intrusively predict the outcome of, or emulate, the non-data-driven intrusive predictor. We describe a number of these emulators in Section 3.3. The main advantage of these emulators is that a training dataset can be constructed to any desired size and diversity, since the labels are produced by an algorithm and the labelling process can be automated. There is, however, a limitation inherent to these data-driven SI predictors that emulate existing non-data-driven predictors, in that any flaws in the non-data-driven predictor are inherited by the data-driven emulator.

Chapter 3

Speech Intelligibility Predictors

SIP is an active research field and the human perception of speech is not yet fully understood, particularly the processes beyond the peripheral auditory system [94, 106, 107]. There are many different methods developed for SI prediction, some of which are inspired by models of speech production, transmission and perception. They may work well in various circumstances, but less well in others. Very broadly, SI predictors can be categorized as *intrusive* or *non-intrusive* as already mentioned in Section 1.3. Furthermore, they can be categorized into *non-data-driven* or *data-driven* predictors. Non-data-driven SI predictors, we define as being based on hand crafted features, as opposed to machine learning models trained on labelled data. Data-driven SI predictors, on the other hand, we define as being based, at least in part, on machine learning models where the model parameters are optimized using large labelled datasets [70].

It is worth noting that SI predictors can also be categorized as monaural or binaural, referring to whether they operate on the same signal for both ears, diotic listening, or a separate signal for each ear, dichotic listening. Dichotic listening brings with it a unique impact on SI, as explained by factors such as inter-aural phase-, time- and level differences between the signals reaching each ear [14, 15, 25, 27]. Typically, binaural SI predictors handle the impact of dichotic listening separately, e.g., by using the equalization-cancellation model [29], which in signal processing terms is an interference cancelling beamformer using the left and right ear signals as input, see for instance [6, 12, 16, 137]. In early stages of the research underlying this dissertation, we realized that binaural listening test data is even more scarce than monaural data, and for this reason decided to focus on monaural SI prediction.

3.1 Categorization of SI Predictors

3.1.1 Comparative Measures

SI predictors are designed to compute a *comparative measure* between a clean reference and the corresponding noisy/processed test signal in a given signal domain. Broadly, we can list four types of comparative measures utilized in the design of existing SI predictors, namely *SNR*, *correlation*, *mutual information* and *learned comparisons*, as well as five signal representation domains, namely the *frequency* domain, *time-frequency* domain, *temporal modulation* domain, the *spectro-temporal modulation* domain, and finally *learned* domains:

- **SNR.** The effects of additive, non-modulated noise on SI is among the earliest studied, and is quite well understood today [42]. The impact of additive non-modulated noise on SI can be modelled by the summation of independent SNR-based contributions to the overall SI from a range of frequency bands of varying importance. This is evidenced, for example, by the Articulation Index (AI) [42], the Speech Intelligibility Index (SII) [57] as well as the Extended SII (ESII) [113].
- **Correlation and coherence.** Non-linear processing can potentially have a great impact on the SI of speech signals [18, 19, 75, 86, 88]. For non-linearly processed speech, it is not straight-forward to compute SNR, because the signal and noise components in the processed signal are generally not easily separable, but non-linear processing can have. Predictors like the Short-Time Objective Intelligibility (STOI) [133], the Extended STOI (ESTOI) [59] or the Coherence SII (CSII) [71] make use of other statistics than SNR, i.e., cross correlation and coherence in particular, which can be computed for non-linearly processed signals.
- **Mutual information.** While the coherence and correlation measures employed by, e.g., CSII and (E)STOI allow for application in a much more general context than the original AI and SII, they are chosen, at least in part, for their mathematical simplicity. More precisely, coherence and correlation both measure the linear dependency between random variables or processes [73, Chapter 7]. Mutual information is a more general measure of the dependency between two random variables or processes. The mutual information between a clean reference and noisy/processed test signal is therefore potentially better suited as a comparative measure since it captures higher orders of statistical dependencies. Various estimators of mutual information have been applied in SI predictors, such as the mutual information sub-band measure [135], Speech Intelligibility based on Mutual Information (SIMI) [61] and Speech Intelligibility In Bits (SIIB) [81].

- **Learned comparisons.** Data-driven SI predictors are often not end-to-end, but rely on data-driven and non-data-driven parts. Typically the data-driven part of a data-driven SI predictor is used to learn and compute a comparison between the clean reference and noisy/processed test signals, or to estimate such a comparison in the absence of a clean reference, based on features extracted by the non-data-driven part. For example, the data-driven SI predictor proposed in [5] was designed to learn its own comparative measure in the time-frequency domain.

Note that these categories of comparative measures are not completely mutually exclusive. For example, it turns out that the AI, though SNR based, is in fact an estimator of mutual information for the specific case, where the additive noise comes from a memoryless Gaussian process [2, 82]. Other methods like the speech-based envelope power spectrum model [63, 64] combine two of the categories, namely SNR and spectro-temporal modulation factors for the prediction of SI.

We find that even non-intrusive SI predictors, though they don't have access to the clean reference signal, can still be categorized by the same taxonomy of comparative measures as intrusive SI predictors, because non-intrusive predictors are designed with an intended comparison that they aim to estimate. This will become clear as we describe non-intrusive SI predictors in Section 3.2.

3.1.2 Signal Domain

SI predictors can be further categorized by the domain in which they analyse the speech signal.

- **Frequency domain.** Early SI predictors, for instance the Articulation Index [42], Speech Transmission Index [54] and Speech Intelligibility Index [57], operate completely in the frequency domain. They do so by computing long-term averages of relevant statistics across time, e.g., long-term speech and noise power spectral densities, cf. Figure 3.1 (c) or signal-to-noise ratios within each frequency sub-band.
- **Time-frequency domain.** While frequency domain based SI predictors can work well for non-modulated additive noise, they perform poorly in modulated noise conditions [9, 13, 23, 26, 37–39, 45, 92, 96, 129]. To perform better in modulated noise conditions, SI predictors like the Extended SII (ESII) [59] and the glimpse proportion model [21], were designed to work in the time frequency domain, i.e., on signals decomposed into spectrograms, cf. Figure 3.1 (b), rather than frequency bands.

- **Temporal modulation domain.** Some SI predictors, like the speech-based Envelope Power Spectrum Model (sEPSM) [63, 64] can make use of temporal modulation frequencies, which are characteristic to speech signals, by operating in the temporal modulation domain. The temporal modulation domain representation of a signal can be reached, e.g., by filtering the individual frequency sub-bands in a time-frequency log magnitude spectrogram using a bank of modulation filters, such as is done in [63], [64] and [110].
- **Spectro-temporal modulation domain.** Spectro-temporal modulations are inherent in speech and important to SI [20]. The Spectro-Temporal Modulation Index (STMI) [32] and the Spectro-Temporal Glimpsing Index (STGI) [31] both predict SI through signal analysis in the spectro-temporal modulation domain. More specifically, these methods further decompose log magnitude spectrograms into spectro-temporal modulation frequency channels. This is achieved by convolving the input spectrogram with 2-dimensional filter kernels corresponding to a range of combinations of temporal and spectral modulation frequencies. Figure 3.2 shows an example of a speech signal decomposed in this way.
- **Learned domain.** While many SI predictors operate in physically motivated domains like frequency, time-frequency or modulation domains, some SI predictors operate in a learned domain. This is particularly the case for recent data-driven SI predictors. An example of this is the CNN based SI predictor proposed in [122], which is trained to estimate the speech transmission index and designed to learn its own domain representation.

The SI predictors described in this chapter, as well as those presented in Part II, are categorized by comparative measure and signal domain in Tables 3.1 and 3.2, which encompass intrusive and non-intrusive predictors, respectively. We separate intrusive and non-intrusive SI predictors in different tables, because non-intrusive SI predictors do not compute comparative measures directly. Instead we found that they are designed to estimate an intended comparative measure. Conveniently, this separation also serves as an overview of intrusive and non-intrusive SI predictors.

3.1. Categorization of SI Predictors

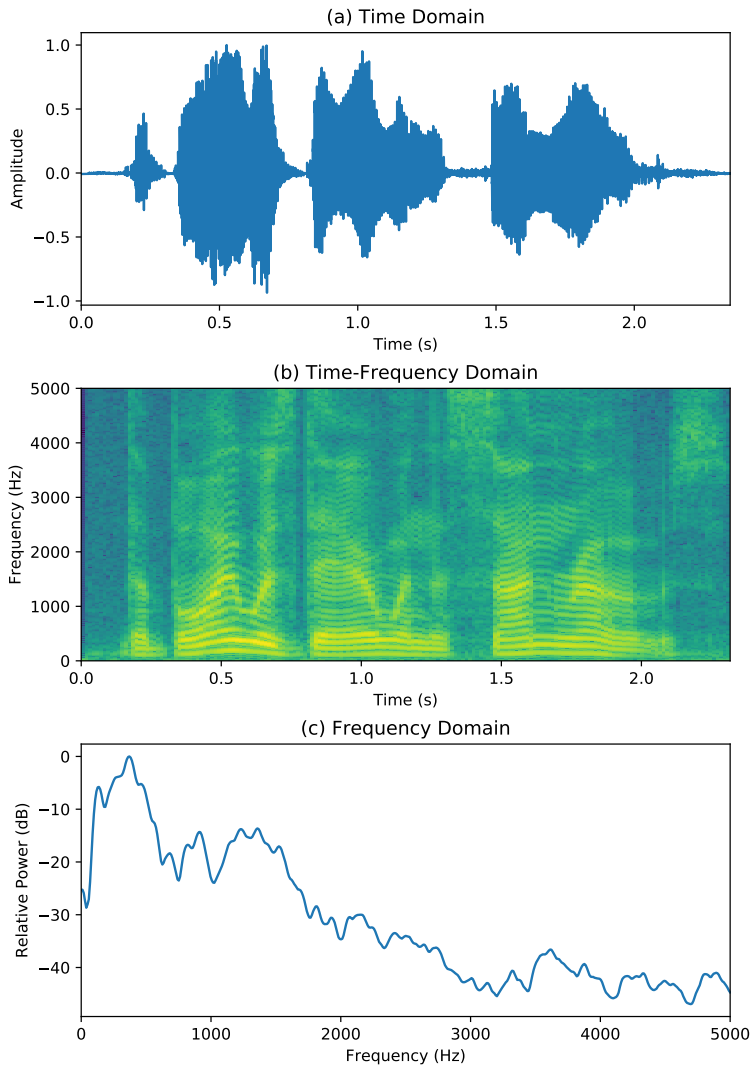


Fig. 3.1: The time domain, time-frequency domain and frequency domain representations of a speech signal produced by a male speaker, saying the sentence "The boy was there when the sun rose." The time domain curve shows the waveform of the speech signal. The time-frequency domain spectrogram shows the log magnitude of time frequency tiles obtained via a short-time Fourier transform. Finally, the frequency domain curve shows the average power over time in the individual Fourier frequency bands.

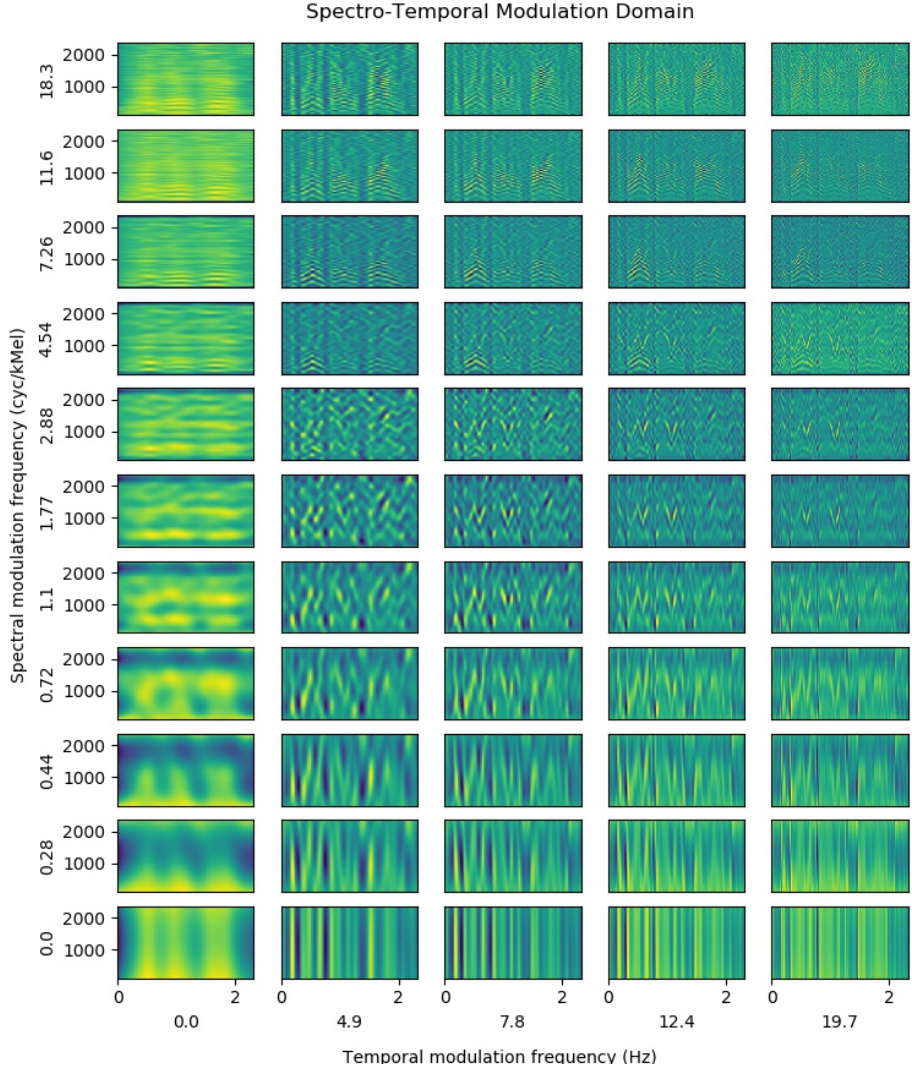


Fig. 3.2: Spectro-temporal modulation domain representation of the same speech signal as in Figure 3.1. The spectral and temporal modulation frequencies of the filter kernels used in the transform to this domain are identical to those used in [30].

3.1. Categorization of SI Predictors

Table 3.1: Intrusive SI predictors arranged vertically by comparative measures; SNR, Correlation (Corr.), Mutual Information (MI), and Learned Comparison (LC), as well as horizontally by signal domains; Frequency Domain (F), Time-Frequency domain (TF), Temporal Modulation domain (TM), Spectro-Temporal Modulation domain (STM), and Learned Domain (LD). Predictors proposed in the papers included in Part II of this thesis are marked in bold face.

Comparative Measure	Signal Domain				
	F	TF	TM	STM	LD
	SNR	AI [42] SII [57]	Glimpse prop. [21] ESII [113] NN [79]	STI [54] sEPSM [63] wSTMI [30]	
	Corr.	CSII [71]	STOI [133] ESTOI [59] THMMB-STOI [65]	HASPI [72] STGI [31]	
	MI		MI-subband [134] SIIB [81] SIMI [61]		
LC		ASR [121] ASR [7] CNN [102] CNN [104]	ASR [128] HLLR [69]		

Table 3.2: Non-intrusive SI predictors arranged vertically by estimated comparative measures; SNR, Correlation (Corr.), Mutual Information (MI), and Learned Comparison (LC), as well as horizontally by signal domains; Frequency domain (F), Time-Frequency domain (TF), Temporal Modulation domain (TM), Spectro-Temporal Modulation domain (STM), and Learned Domain (LD). Predictors proposed in the papers included in Part II of this thesis are marked in bold face.

Comparative Measure	Signal Domain				
	F	TF	TM	STM	LD
	SNR		SRMR [34] ModA [17]		
	Corr.	NISTOI [4] STOINET [143] LSTM [141] LSTM [76] CNN [103]	NISA [123]		
	MI				
LC		CNN [5] NORI [66] LSTM [36]			CNN [122]

3.2 Non-Data-Driven SI Predictors

3.2.1 SNR Based SI Predictors

The Articulation Index

The earliest attempt to algorithmically predict SI was through the Articulation Index (AI) [42] proposed in 1947, cf. Table 3.1. The AI operates in the frequency domain and is built on the assumption that SI contributions from separate frequency bands of a noisy speech signal are independent, and can be added to estimate the overall SI of said noisy speech signal. This is summarized by the following equation:

$$A = \sum_n W_n A_n, \quad (3.1)$$

wherein A is the AI, W_n is the so-called band importance, i.e., the largest possible contribution to the AI from the n 'th frequency band, and A_n is a number between 0 and 1, computed based on the given noise and speech signals, that determines the actual contribution of the n 'th frequency band. The AI uses 20 frequency bands in the range from 250 to 7000 Hz., with specific bandwidths chosen such that W_n was the same value for all frequency bands, i.e., the bands were of equal importance [42]. The computations of the values A_n take into account the long term average SNR, as well as the speech reception threshold and masking effects of noise in the corresponding frequency bands.

The AI successfully predicts the effects of non-modulated noise masking on SI, but it has two important limitations. First, electronic computing technology was barely in its infancy in 1947, and the AI was designed to be computed graphically by hand. Secondly, the AI depends on the long-term characteristics of additive noise, and does not predict the effects of short-term noise fluctuations or distortions on SI [113].

The Speech Transmission Index

In order to more accurately predict the effects of non-linear distortions, arising as a result of the transmission of speech through communications channels, the Speech Transmission Index (STI) [54] was proposed. The STI is a tool to evaluate transmission channels, rather than specific speech signals as most of the other SI predictors mentioned in this dissertation.

The idea behind the STI is that transmission channels can cause difficulties in distinguishing between speech sounds, i.e., mistaking one speech sound for another. The STI attempts to quantify this effect through a bank of artificial, speech-like probe signals, designed to mimic the temporal modulations of natural speech [130]. The probe signals are decomposed into

N separate frequency bands and sent through the investigated transmission channel in pairs. The ratios of sound pressure level differences between the probe signals before and after transmission is computed as,

$$\text{STI}_{i,j} = \frac{\sum_{n=1}^N |L'_{i,n} - L'_{j,n}|}{\sum_{n=1}^N |L_{i,n} - L_{j,n}|}, \quad (3.2)$$

where $L'_{i,n}$ and $L'_{j,n}$ denote the sound pressure levels in dB of the n 'th band of the i 'th and j 'th probe signals after transmission, and $L_{i,n}$ and $L_{j,n}$ similarly denote the sound pressure levels of the probe signals before transmission. The idea is that if the differences between a particular pair of probe signals are significantly reduced by the transmission, i.e., the ratio in (3.2) is low, then the speech sounds represented by the probe signals are likely to be mistaken for one another after transmission. Finally, the average across all probe signal pairs yields the STI.

Extended versions of the STI, using other types of probe signals, including actual speech, have since been studied [43, 100, 101].

The Speech Intelligibility Index

The Speech Intelligibility Index (SII) [57] builds on the same assumption as the AI that SI can be predicted by independent, additive contributions from separate frequency bands:

$$\text{SII} = \sum_n I_n A_n, \quad (3.3)$$

where $0 \leq I_n \leq 1$, the band importance, indicates the importance of the n 'th frequency band, and A_n indicates the contribution to SI of the n 'th band, based on the SNR and perception threshold in the n 'th frequency band [57]. The advantage of the SII, over the AI, is that A_n can be computed electronically. In the computation of $0 \leq A_n \leq 1$ the perception threshold can also be adjusted to account for the hearing profile of a specific listener.

The SII has since been shown to underestimate the SI of speech in temporally modulated noise types [113]. A modified version of the SII, the Extended SII (ESII) [113, 114], was proposed as a way of extending the scope of the SII to include temporally modulated noise. The idea behind the ESII is quite simple: The ESII is computed by applying the SII individually to short segments of the signal under test, ranging from 35 to 10 ms depending on the frequency band, and computing the average of the results.

The Spectro-Temporal Modulation Index

As with the STI, the Spectro-Temporal Modulation Index (STMI) [32] is an SI predictor that builds on the idea of measuring the rate of mistaken speech

sounds by human listeners. The STMI makes use of a spectro-temporal modulation decomposition of speech, cf. Figure 3.2, to compare the clean reference and noisy test speech signals, as opposed to the STI which compares their temporal envelopes.

In addition to predicting the SI of specific noisy test signals, the STMI can also, like the STI, be used to evaluate the effect of a transmission channel on SI. For this purpose, the specific speech signals are replaced by a pre-constructed array of spectro-temporal modulation patterns passed through the transmission channel, which the STMI is subsequently applied to.

The Glimpse Proportion Model

The Glimpse proportion model of SI proposed in [21] shares similarities with ESII in that it is based on SNR in different time-frequency regions of the test signal. Instead of summarizing weighted contributions from the frequency bands, however, the Glimpse model simply counts the number of so-called glimpses where the SNR is locally, in the spectro-temporal sense, high. The Glimpse model is perceptually motivated, operating under the assumption that SI is related to the rate of glimpses in the noisy speech, which is linked to the hypothesis that humans can understand speech in noise, by listening to glimpses, i.e., time-frequency regions with high SNR [21, 55, 85, 105].

The Speech-to-Reverberation Modulation Energy Ratio

The Speech-to-Reverberation Modulation energy Ratio (SRMR) [34] is a non-intrusive SI predictor that utilizes temporal modulation energy similarly to STI. Because of the constraints of non-intrusivity, the modulation energies of the noisy/processed test signal can not be directly compared to those of the clean reference signal, but SRMR utilizes the observation that speech modulation energy tends to be concentrated at modulation frequencies below 20 Hz, whereas the modulation energy of reverberations is scattered across a wider range of modulation frequencies [34]. Reverberations can reduce the SI of a speech signal, particularly when the delay of reflections is greater than 50 milliseconds [28, 93, 95, 111, 112]. SRMR is defined as a ratio of modulation energy below and above approximately 20 Hz.

$$\text{SRMR} = \frac{\sum_{k=1}^4 \bar{\varepsilon}_k}{\sum_{k=5}^K \bar{\varepsilon}_k}, \quad (3.4)$$

where $\bar{\varepsilon}_k$ is the average energy in the k 'th modulation band.

Speech-based Envelope Power Spectrum Model

The Speech-based Envelope Power Spectrum Model (sEPSM) [63] predicts SI as a sum of SNR values similarly to the AI and SII. However, the sEPSM com-

putes the SNR values in the temporal modulation domain. The sEPSM was demonstrated to work well for speech contaminated with stationary additive noise, reverberation and speech denoised by spectral subtraction. A multi-resolution extension to the sEPSM was later proposed in [64] to improve the prediction performance for non-stationary additive noise.

Modulation Area

Similarly to SRMR, the Modulation Area (ModA) [17] method utilizes temporal modulation energy, but rather than a ratio, cf. (3.4), ModA computes the area under the temporal modulation energy curve of the noisy/processed test signal as a non-intrusive predictor of SI

$$\text{ModA} = \frac{1}{N} \sum_n^N A_n, \quad (3.5)$$

where A_n is the modulation energy in the n 'th modulation frequency band. ModA was developed following the underlying observation that noise and particularly reverberations reduce the area under the temporal modulation curve.

Weighted Spectro-Temporal Modulation Index

Band-importance, or frequency weighting, plays a central role in the AI and SII, and the weighted Spectro-Temporal Modulation Index (wSTMI) [30] takes inspiration from this fact. Where the original STMI computes a uniform average of the predicted SI contributions by each spectro-temporal modulation band, wSTMI employs a sparse set of weights to compute a linear combination instead. The weights used by wSTMI were fitted to intelligibility listening test data using L_1 regularized optimization, and showed similarities to modulation transfer functions important to the human auditory system. This alignment with human perception is perhaps what allows the wSTMI to predict SI more accurately as compared to the original STMI and other existing SI predictors, particularly for highly modulated noise and distortions [30].

3.2.2 Correlation Based SI predictors

Coherence Speech Intelligibility Index

The AI and SII algorithms compute the SNR of a noisy test signal, which means that these algorithms rely on an implicit assumption of speech in additive noise, which allows SNR to be computed. However, for speech signals that have been subject to non-linear processing, SNR is not straightforward to compute. For this situation, algorithms such as the Coherence SII (CSII) [71]

have been proposed. The CSII is a modification of the SII, that uses a coherence based Signal-to-Distortion Ratio (SDR) in the computation of the band-wise contributions to SI. CSII is computed using the clean reference and noisy/distorted test signal,

$$\text{CSII} = \sum_j \frac{\sum_k W_j(k) |\gamma(k)|^2 S_{yy}(k)}{\sum_k W_j(k) (1 - |\gamma(k)|^2) S_{yy}(k)}, \quad (3.6)$$

where

$$\gamma(k) = \frac{S_{xy}(k)}{\sqrt{S_{xx}(k) S_{yy}(k)}} \quad (3.7)$$

is the coherence function, $W_j(k)$ is the frequency domain representation of a band pass filter corresponding to the j 'th sub-band used in the SII [71], $S_{yy}(k)$ is the power spectral density of the noisy/processed test signal, $S_{xx}(k)$ is the power spectral density of the clean reference signal and $S_{xy}(k)$ is the cross power spectral density of the clean reference and noisy/processed test signals. Notably, this computation does not require any prior knowledge of the noise or distortion characteristics. Consequently, the applicability of CSII extends beyond speech in additive noise.

Short-Time Objective Intelligibility

Short-Time Objective Intelligibility (STOI) [133] was proposed as an SI predictor designed for noisy speech processed by single-microphone noise reduction algorithms using time-frequency domain multiplicative masks [56, 60, 75, 77].

Rather than SNR, STOI computes sample correlation values between 384 ms segments of the temporal magnitude envelopes in one-third octave bands of the clean reference and noisy/processed test signal,

$$d_{j,m} = \frac{(\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}})^\top (\mathbf{y}_{j,m} - \mu_{\mathbf{y}_{j,m}})}{\|\mathbf{x}_{j,m} - \mu_{\mathbf{x}_{j,m}}\| \|\mathbf{y}_{j,m} - \mu_{\mathbf{y}_{j,m}}\|}, \quad (3.8)$$

where $\mathbf{x}_{j,m}$ is m 'th segment in the j 'th band of the clean reference signal, similarly $\mathbf{y}_{j,m}$ is the m 'th segment in the j 'th band of the noisy/processed test signal subject to a clipping procedure, and μ denotes the sample mean of the subscripted vector. The use of a correlation measure, as opposed to SNR, means that STOI can be applied to non-linearly processed signals, requiring only that the clean reference and noisy/processed test signals are time-aligned [133].

Similarly to the ESII, the ESTOI [59] algorithm is an extension of STOI that expands the scope of noise/processing conditions for which the SI can be accurately predicted. The extension in ESTOI is that the correlations are

computed between normalized spectro-temporal envelopes, rather than temporal envelopes only, as the case is for the baseline STOI [59]. This extension was shown to allow for greatly increased accuracy in SI predictions of speech signals contaminated by temporally modulated noise [59].

The Non-Intrusive STOI (NISTOI) [4] algorithm is a non-intrusive extension of STOI that estimates the clean reference from the noisy/processed test signal, and subsequently applies the original STOI algorithm to the noisy test and estimated clean reference signals. To estimate the clean reference signal, NISTOI projects the noisy/processed test signal into a low-dimensional subspace determined via principal component analysis of generic speech, in the temporal modulation domain. Other variations of this strategy using different estimators for the clean reference have been proposed, such as the Pitch-Based STOI (PB-STOI) [126, 127]

The Hearing Aid Speech Perception Index

The Hearing Aid Speech Perception Index (HASPI) [72] predicts SI using a model of the human auditory system that takes a combination of spectro-temporal modulation and coherence into account. Using this model, HASPI is also able to account for the specific hearing profile of an individual listener. The index is computed as a linear combination of coherences and spectro-temporal correlations between the clean reference and the noisy/processed test signals, at the output of the auditory model.

Spectro-Temporal Glimpsing Index

Inspirations were drawn from STMI, the Glimpse proportion model, and ESTOI in the development of the Spectro-Temporal Glimpsing Index (STGI) [31]. This index works in the spectro-temporal modulation domain, like STMI, and makes short-term comparisons between the clean reference and noisy/processed test signals using normalized correlation coefficients, like ESTOI. STGI uses a threshold on these correlation coefficients to detect glimpses, like the Glimpse model. The original Glimpse model uses SNR to detect glimpses, which restricts it to speech in additive noise, but this restriction does not apply to STGI since the glimpses are detected using correlation coefficients instead.

3.2.3 Mutual Information Based SI Predictors

While the correlation based SI predictors described in the previous Section 3.2.2 rely on second order statistics, the class of mutual information based SI predictors generalizes this idea further and make comparisons in terms of higher order statistics. It is hard, if not impossible, to compute mutual information directly for SI prediction, since the definition of mutual information

includes the joint probability density function of the two signals, in this case the clean reference and noisy/processed test signals. Instead, the mutual information between the clean reference and noisy/processed test signals may be estimated.

Mutual Information Sub-band Measure

The mutual information sub-band measure proposed in [134] and further explored in [135] employs a K-Nearest Neighbour (KNN) mutual information estimator between the segmented one-third octave envelopes of the clean reference and noisy/processed signals. Thus the signal domain of this SI predictor is the same as that of STOI and ESTOI, but the comparative measure is estimated mutual information, as opposed to sample correlation.

Speech Intelligibility In Bits

The Speech Intelligibility In Bits (SIIB) algorithm [81] uses a KNN mutual information estimator, but applies a pre-whitening transform, the Karhunen-Loève Transform (KLT), to segments of the clean reference and noisy/processed test signals in the time-frequency domain. This transform eliminates correlations across time and frequency in the speech signal and even though the KNN mutual information estimator assumes statistical independence, and not only uncorrelation, this results in improved estimates of the mutual information [81].

Speech Intelligibility Based on Mutual Information

The Speech Intelligibility based on Mutual Information (SIMI) algorithm [61] uses a lower bound estimate of mutual information that relies on the linear minimum mean squared error estimator of the clean speech signal given the noisy/processed test signal. The lower bound estimator is computationally simpler than the KNN estimator, and SIMI can be computed significantly faster than the mutual information sub-band measure and SIIB as a result.

3.3 Data-Driven SI Predictors

Data-driven methods, such as neural networks and hidden Markov models, have been adopted across many research fields, due to their power and versatility. The field of SI prediction is no exception and there has been a rapid development of data-driven SI predictors within the recent decade. In this section we give an overview of the state-of-the-art data-driven SI predictors, which includes various types of neural networks, SI predictors based on au-

tomatic speech recognition and data-driven emulators of non-data-driven SI predictors.

Automatic Speech Recognition (ASR) systems [70, 140] are used to produce text transcripts of speech signals, and data-driven ASR systems can do so with great accuracy. Though humans and ASR systems do not necessarily recognize speech using the exact same mechanisms, i.e., human SI may be quite different from the SI of an ASR system [66], the use of ASR systems in SI prediction is motivated by the fact that it enables simulating a subjective listening test by replacing human listeners with a machine listener.

A common strategy, which has produced a large number of non-intrusive SI predictors, is to use an intrusive non-data-driven SI predictor, like STI or STOI to generate the labels for a training set of noisy/processed speech signals, and then train a data-driven non-intrusive model to emulate the non-data-driven SI predictor in question. This approach bears the advantage that any desired amount of training data can be generated quickly and efficiently because listening tests are not required. However, these emulators inherit any flaws and limitations inherent in the emulated SI predictor.

3.3.1 SNR Based SI Predictors

Neural Network for Binaural SI Prediction

A data-driven binaural SI predictor, proposed in [79], uses a neural network to map a number of SNR-based features in two different perceptually based frequency decompositions, namely the critical frequency bands used in the AI, and frequency bands on the Mel scale [99]. The so-called Better Ear model, Band-Wise Better Ear model and Pooled Channel model were used to combine the left and right signals in different ways. The better ear models select either the left or the right channel, band-wise or fully, based on which has the highest SNR, whereas the pooled channel model presents both channels to the neural network. The network was found to have good SI prediction performance, but the training and testing datasets were relatively small, so the generalizability of the network is unclear.

3.3.2 Correlation Based SI Predictors

Non-Intrusive Speech Assessment

The Non-Intrusive Speech Assessment (NISA) [123] method predicts STOI scores without the use of a clean reference signal, i.e., it is a STOI emulator. NISA extracts a range of both short and long-term features, notably the temporal modulation envelopes, of a noisy speech signal and employs tree based regression to estimate STOI scores based on these features. NISA was trained

and tested on speech in additive noise and speech distorted by telecommunication channels. It was shown that NISA is able to predict STOI with very high accuracy for a broad range of SNRs [123].

Twin Hidden Markov Model-Based STOI

The Twin Hidden Markov Model-Based STOI (THMMB-STOI) [65] is a non-intrusive STOI emulator that utilizes a so-called twin HMM consisting of two series of observations, sharing the same hidden states. One half of this twin HMM is trained as an ASR system and the other half as a speech synthesis system. By using the text transcript of the clean speech reference signal, the hidden states of the HMM can be estimated and subsequently used to synthesize an estimation of the underlying clean speech signal. Finally, the synthesized clean speech signal and the noisy/processed test signal are fed to the STOI algorithm to produce an SI prediction. Although the THMMB-STOI method does not require access to the clean reference signal, it does require a transcript of the clean reference signal, and can therefore not be seen as a fully non-intrusive SI predictor. A modification to THMMB-STOI was proposed in [68], where an estimated transcript, produced by the ASR part of the twin HMM, is used instead of the ground-truth transcript.

STOINET

STOINET [143] and the related Multi Objective Speech Assessment Net (MOSA-Net) [142], are trained to emulate STOI non-intrusively, and in the case of MOSA-Net other measures, such as the Perceptual Evaluation of Speech Quality (PESQ) [115] as well. STOINET combines a convolutional neural network with a bidirectional Long-Short-Term Memory (LSTM), a type of RNN architecture, and is trained to emulate STOI. STOINET contains twelve CNN layers, followed by a bidirectional LSTM, and finally an FNN applied frame-wise to obtain estimated STOI scores. The input to the network is a noisy/processed speech signal in the Short-Time Fourier Transform (STFT) time-frequency domain. The network was trained and tested using speech contaminated by several types of additive noise, as well as speech processed by a de-noising NN. Accurate STOI emulation performance was reported for seen noise/processing conditions, whereas a drop in this performance was observed for unseen noise/processing conditions [143].

Two other variants of LSTM architectures were trained as emulators of STOI in [141] and [76].

3.3.3 Learned Comparison Based SI Predictors

HMM-based Log Likelihood Ratio

In [69], a Hidden Markov Model (HMM) based ASR system, trained in the temporal modulation domain, is employed for SI prediction. The ASR system produces a predicted text transcript of the noisy processed speech signal, which is used to compute the log-likelihood ratio between the predicted text transcript and the ground-truth transcript. This ratio is the proposed SI-predictor, the HMM-based Log Likelihood Ratio (HLLR). Since the HLLR method requires a ground truth transcript, it is an intrusive SI predictor.

No Reference Intelligibility

Building on the HLLR method, an ASR system is trained to mimic human word recognition performance in [67]. Two features are then extracted from the ASR system and used as SI predictors: The normalized likelihood difference and the time alignment difference. Note that computing these features requires access to the ground-truth transcript and that the method in [67] is thus not fully non-intrusive. The NO Reference Intelligibility (NORI) method [66], however, extends the method in [67] by computing three additional features extracted from the ASR system, namely the entropy, the log-likelihood ratio and the dispersion, all of which can be obtained non-intrusively. Where NORI uses a word level ASR system, i.e., an ASR system which is trained to recognize a specific set of words, a phoneme level ASR system was used to predict SI in [7].

Matrix Sentence HMM-SI

Another HMM-ASR system based, intrusive SI predictor was proposed in [121]. This predictor uses an ASR system trained on matrix sentences, with limited lists of words and uses the number of correctly identified words by the ASR system as a predictor of SI. The method is designed to be used in conjunction with matrix test sentences that have a limited vocabulary, such as the Dantale sentences [97].

Matrix Sentence DNN-SI

Similar to the matrix sentence HMM-SI described above, a deep neural network based ASR system for intrusive SI prediction is proposed in [128]. This deep neural network is trained for ASR. The network takes as input noisy/processed speech signals in the temporal modulation domain, along with a ground-truth text transcript of the speech. The proportion of correctly recognized words is computed and used as a prediction of SI. This method was shown to predict SI well for a range of speech-like noise types, although

a drop in performance was observed for mismatched noise types between training and testing [128].

STI Emulator CNN

A non-intrusive STI emulator is proposed in [122], which uses a deep CNN trained on speech convolved with artificially generated room impulse responses, and labels generated by STI. The STI emulator is given time domain speech signals as input and is trained end-to-end. Hence, the method operates in a learned domain. It was demonstrated to predict STI scores of reverberant speech signals with very high accuracy [122].

CNN-Based SI Predictor

A CNN for non-intrusive SI prediction was proposed in [5]. The CNN architecture was designed based on the hypothesis, at the time, that listening test data was too scarce to allow for a large network with millions of parameters, a hypothesis that is validated by the research presented in this dissertation. It was also hypothesized that SIP is a relatively simple problem which can be solved by a small neural network. This may be true when a specific, constrained set of listening conditions are considered, but in general the problem appears more complex [104]. The CNN was trained using noisy/processed test signals in the one-third octave band time-frequency domain with measurements of SI from three listening tests as labels. The CNN consisted of one convolutional layer, followed by a global temporal average pooling operation, and finally three FNN layers. The CNN-based SI predictor was shown to predict SI well for unseen stationary additive noise, but less well for fluctuating additive noise and speech processed by single microphone de-noising algorithms [5].

LSTM-Based SI Predictor

A Long Short-Term Memory (LSTM) neural network SI predictor trained on listening test data is proposed in [36]. This LSTM operates in the time-frequency domain, and employs an attention mechanism in order to allow different time frames of the input signal to contribute unequally to the overall SI. The LSTM based SI predictor was trained and tested on a dataset of talkers suffering from dysarthria, demonstrating significantly improved performance over a baseline support vector machine based SI predictor that was trained and tested on the same data.

Chapter 4

Scientific Contributions

Part II of this dissertation consists of four research papers. These papers contain the scientific contributions of this dissertation to the field of data-driven SI prediction. In particular, the central problem of listening test data scarcity is identified, and initial steps towards solutions are proposed. The focus is always on comparison to baseline state-of-the-art SI predictors, and evaluation on listening conditions outside the scope of the training set.

4.1 Specific Contributions

Here, we will briefly summarize the contributions of each research paper found in part II.

4.1.1 [A] A Neural Network for Monaural Intrusive Speech Intelligibility Prediction

In this paper we propose to train a neural network for the specific task of monaural, intrusive SIP. At the time, in 2020, existing neural network based SI predictors were either non-intrusive or binaural. The neural network is trained and tested on an aggregated dataset of listening test data from four different tests. We investigate the possibility and effects of training the network on individually labelled words, as opposed to averaged SI scores across whole noise/processing conditions, and analyse the prediction performance of the trained neural network as a function of input signal duration.

We find that the proposed network is able to achieve higher performance than state-of-the-art baseline SI predictors STOI [133] and ESTOI [59], albeit in noise/processing conditions, which are present in the network training dataset. Additionally, we find that the proposed SI predictor is able to produce accurate predictions with shorter durations of input than the baseline

SI predictors.

4.1.2 [B] End-to-End Speech Intelligibility Prediction using Time-Domain Fully Convolutional Neural Networks

In this paper we identify that scarcity of listening test data is a primary limiting factor in the development of data-driven SI predictors. Specifically, we analyse the existing body of data-driven SI predictors and find that many of them are either emulators of non-data-driven SI predictors or trained on relatively little listening test data, i.e., one or two listening tests. In order to investigate the severity of the data scarcity problem, we propose a fully data-driven end-to-end SI prediction scheme, consisting of a time-domain convolutional neural network using the U-Net architecture, cf. Section 2.1.2. Motivated by the success of this architecture for image segmentation, we apply it to speech segmentation into segments of various levels of intelligibility, and compute the average over time to obtain predictions of SI. We train the proposed SI predictor on data from a relatively large set of listening tests and investigate the generalizability to unseen talkers and noise/processing conditions as compared to the baseline SI predictors, STOI [133], ESTOI [59], HASPI [72] and SIIB [81].

We find that the proposed SI predictor is able to reach higher than baseline performance for seen talkers and noise/processing conditions, but that it does not generalize as well to unseen conditions. In other words, predictions under unseen conditions are below the baseline. This cements that the listening test data scarcity is a major limitation in the further development of data-driven SI predictors.

4.1.3 [C] Training Data-Driven Speech Intelligibility Predictors on Heterogeneous Listening Test Data

In this paper we propose a training strategy to solve a problem that arises when data-driven SI predictors are trained on aggregated sets of different listening tests. The problem is that data-driven SI predictors become overly specialized to the listening tests contained in the training data, resulting in poorer performance when trying to predict the results of listening tests that have not been trained on. This specialization is caused by the listening test paradigms, i.e., the talkers, languages, vocabularies, equipment etc., cf. Section 1.3, that the SI predictor does not have access to when predicting SI. The effects of the listening test paradigms of the training data are internalised by the SI predictor during training, because this leads to better predictions on the training data. This internalisation is detrimental to the general performance, i.e., when predicting the SI of speech signals in new paradigms that are not part of the training set.

We find that our proposed training strategy, which involves appending a temporary layer of sigmoidal mapping functions unique to each individual listening test in the training set to the end of the network, improves the performance of a data-driven SI predictor significantly on unseen noise/processing conditions. The purpose of the appended test-specific layer is to model the influence of listening test paradigms on SI for each listening test individually, which means this layer can be discarded after training to remove the specialization to these listening tests from the network.

4.1.4 [D] Data-Driven Speech Presence Probability Estimation for Non-Intrusive Speech Intelligibility Prediction

In this paper we propose a novel approach to data-driven SI prediction that circumvents the listening test data scarcity problem. The approach is driven by the hypothesis that SI is strongly linked to Speech Presence Probability (SPP), defined in the time-frequency domain on a tile-by-tile basis as the probability that the per-tile SNR is above a fixed threshold. Our proposed SI predictor, which we call Deep Speech Presence, is composed of a neural network SPP estimator, and a post-processing stage mapping estimated SPPs to SI predictions. The SPP estimator is trained on a dataset of speech in additive noise and automatically computed labels. This type of data and labels are much easier to obtain than listening test data, which is a major motivation for this method. The post processing stage for mapping estimated SPP's to SI uses what we call top- p percent average, which involves finding the p percent of tiles with the highest estimated SPP's and computing the average of just those tiles. We explain the efficacy of this particular post-processing step as the result of removing estimations with high uncertainty, namely those close to an estimated SPP of 0.5. The proposed approach bears similarities to the glimpse proportion SI predictor [21], but whereas glimpse proportion is intrusive and deals with SNR, our approach is non-intrusive and relies on probabilities, SPP's, rather than SNR. We compare the proposed SI predictor to the baseline non-intrusive predictors NISTOI [4], SRMR [34] and a data-driven emulator of STOI [133] trained on the same dataset as Deep Speech Presence. We find that a deep neural network trained in the proposed manner, to estimate SPP's, can be used as an accurate predictor of SI in a great variety of conditions outside the scope of the training data, i.e., even though the SPP estimator is trained on speech in additive noise, the SI predictor does well for non-linearly processed speech. We conclude that this way of predicting SI, through data-driven estimation of SPP, is accurate and generalizable to a wide variety of noise/processing conditions.

4.2 Summary of Contributions

This section will briefly summarize and contextualize the contributions of the collection of research papers presented in Part II. The research presented in these papers includes identification of the problem of listening test data scarcity for data-driven SI prediction and proposed solutions that constitute a first effort to enable data-driven SI prediction in spite of this scarcity.

In papers A and B, the possibilities of intrusive SI prediction using deep neural networks trained on listening test data are investigated. In paper A it is found that such a network trained on four distinct listening tests performs well when tested on a subset of the same listening tests withheld from the training set. In paper B, larger and fully data-driven architectures are trained on a bigger collection of listening tests and tested on other listening test conditions which were not used for training. It is found that when tested on unseen conditions, from different listening tests, the data-driven methods fail to meet the baseline set by non-data-driven state-of-the-art SI predictors.

In paper C the problem of listening test data scarcity for the purpose of training data-driven SI predictors is identified and addressed. Paper C addresses the scarcity in two ways:

- 1) One way of reducing the impact of limited amounts of training data is to limit the number of trainable parameters in the data-driven SI predictor, cf. Section 2.2. The architecture of the SI predictor proposed in this paper is a hybrid of a trainable CNN front-end and a fixed ESTOI back-end, which facilitates a significant reduction in the number of trainable parameters relative to paper B.

- 2) paradigm-specific sigmoidal mapping functions are used in the training phase to prevent the SI predictor from specializing to the paradigms underlying the training data. This SI predictor is demonstrated to exceed state-of-the-art performance in condition it has been trained while matching state-of-the-art performance for unseen conditions.

In paper D, a data-driven, non-intrusive SI predictor is proposed that uses per-time-frequency-tile SPP's, which we show can be estimated by a neural network trained on automatically labelled speech in noise data, which is available in abundance, and hence more easily obtainable than listening test data. With a neural network trained to estimate SPP, we demonstrate that a relatively simple post processing stage can accurately map the estimated SPP's to measured SI in a wide variety of noise and non-linear processing conditions.

4.3 Directions of Future Research

4.3.1 Crowdsourcing SI Data

If the scarcity of listening test data is to be truly resolved, more listening test data needs to be collected. Perhaps a different paradigm of listening test data collection is required. It may be feasible to crowd source SI labels by way of small-scale listening tests involving potentially thousands of listeners or more as demonstrated by [138] and [22]. Distributing a listening test to many people each using their own equipment is likely to result in many different paradigm differences, with potentially significantly lower quality and fewer data samples per paradigm, as compared to performing the listening test in controlled lab conditions. Further research would be required to understand whether paradigm specific mapping functions, as introduced in paper C, can be adapted to this more challenging scenario.

4.3.2 Applying SI Predictors to Speech Enhancement

SI prediction is a valuable tool for evaluating speech enhancement systems [124]. Going beyond the use of SIP for evaluation, and using SI predictors to guide, e.g., speech enhancement systems during their employment, would bring great potential benefits [98]. However, as shown in [78] it may be more complicated than simply optimizing a speech enhancement system for an SI predictor like STOI [133]. It appears that data-driven speech enhancement systems are able to exploit loopholes where, e.g. STOI does not reflect the actual SI. It may be possible to solve this problem by jointly training an SI predictor along with a speech enhancement system, as this would facilitate the removal of such loopholes in the SI predictor. The current listening test data scarcity will undoubtedly make this a challenging task, however.

4.3.3 Applying Data-Driven SIP in Portable Devices

SI-aware hearing assistive devices, such as hearing aids, that predict the SI of their own processed speech, and use the estimated SI to adjust or improve the processing are a desired application of the field of SIP. The requirements for SI predictors to be used in SI-aware hearing assistive devices are:

- 1) non-intrusivity since clean reference signals are not available.
- 2) low computational complexity as the algorithm must be executed on small, battery-driven, low complexity devices.

In contrast to the machines and servers used in research to develop data-driven SI predictors, portable hearing assistive devices are limited in terms of memory and computational power. As such, a substantial reduction in complexity of the current state-of-the-art is required to reach the point of porta-

bility and wearability. Besides the limitations, portable devices also introduce the possibility of utilizing biological modalities like electroencephalograms, which have proven useful in assisting with SI prediction [58, 83].

References

- [1] S. Alhanbali, P. Dawes, R. E. Millman, and K. J. Munro, "Measures of listening effort are multidimensional," *Ear and Hearing*, vol. 40, no. 5, p. 1084, 2019.
- [2] J. B. Allen, "The articulation index is a shannon channel capacity," in *Auditory Signal Processing*. Springer, 2005, pp. 313–319.
- [3] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Predicting the Intelligibility of Noisy and Nonlinearly Processed Binaural Speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.
- [4] —, "A non-intrusive short-time objective intelligibility measure," Mar. 2017, pp. 5085–5089.
- [5] —, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.
- [6] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.
- [7] K. Arai, S. Araki, A. Ogawa, K. Kinoshita, T. Nakatani, K. Yamamoto, and T. Irino, "Predicting speech intelligibility of enhanced speech using phone accuracy of dnn-based asr system." in *Interspeech*, 2019, pp. 4275–4279.
- [8] L. Atlas, T. Homma, and R. Marks, "An artificial neural network for spatio-temporal bipolar patterns: Application to phoneme classification," in *Neural Information Processing Systems*, 1987.
- [9] S. P. Bacon, J. M. Opie, and D. Y. Montoya, "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 3, pp. 549–563, 1998.
- [10] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.
- [11] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [12] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension, and evaluation of a binaural speech intelligibility model," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2479–2497, 2010.
- [13] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acust. United Ac.*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

References

- [14] A. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, 1988.
- [15] —, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3132–3139, 1992.
- [16] A. Chabot-Leclerc, E. N. MacDonald, and T. Dau, "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 192–205, 2016.
- [17] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical signal processing and control*, vol. 8, no. 3, pp. 311–314, 2013.
- [18] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded speech," *Ear and hearing*, vol. 32, no. 3, p. 331, 2011.
- [19] —, "Impact of snr and gain-function over-and under-estimation on speech intelligibility," *Speech Communication*, vol. 54, no. 2, pp. 272–281, 2012.
- [20] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2719–2732, Oct. 1999.
- [21] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, Feb. 2006.
- [22] M. Cooke and M. L. García Lecumberri, "How reliable are online speech intelligibility studies with known listener cohorts?" *The Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 1390–1401, 2021.
- [23] J. De Laat and R. Plomp, "The reception threshold of interrupted speech for hearing-impaired listeners," in *Hearing—Physiological bases and psychophysics*. Springer, 1983, pp. 359–363.
- [24] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Signal Processing of Speech Signals*. John Wiley & Sons, 2000.
- [25] D. D. Dirks and R. H. Wilson, "The effect of spatially separated sound sources on speech intelligibility," *Journal of Speech and Hearing Research*, vol. 12, no. 1, pp. 5–38, 1969.
- [26] J. R. Dubno, A. R. Horwitz, and J. B. Ahlstrom, "Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2897–2907, 2002.
- [27] A. Duquesnoy, "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 739–743, 1983.
- [28] A. Duquesnoy and R. Plomp, "Effect of reverberation and noise on the intelligibility of sentences in cases of presbycusis," *The Journal of the Acoustical Society of America*, vol. 68, no. 2, pp. 537–544, 1980.

References

- [29] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *The Journal of the Acoustical Society of America*, vol. 35, no. 8, pp. 1206–1218, 1963.
- [30] A. Edraki, W. Y. Chan, J. Jensen, and D. Fogerty, "Speech intelligibility prediction using spectro-temporal modulation analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 210–225, 2020.
- [31] —, "A spectro-temporal glimpsing index (stgi) for speech intelligibility prediction," in *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. International Speech Communication Association, 2021, pp. 2738–2742.
- [32] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, no. 2-3, pp. 331–348, Oct. 2003.
- [33] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.
- [34] T. H. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Aug. 2010.
- [35] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, p. 103204, 2022.
- [36] M. Fernández-Díaz and A. Gallardo-Antolín, "An attention long short-term memory based system for automatic classification of speech intelligibility," *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103976, 2020.
- [37] J. Festen, "Speech-reception threshold in a fluctuating background sound and its possible relation to temporal auditory resolution," in *The psychophysics of speech perception*. Springer, 1987, pp. 461–466.
- [38] J. M. Festen, "Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1295–1300, 1993.
- [39] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1725–1736, 1990.
- [40] J. L. Flanagan, *Speech analysis synthesis and perception*. Springer Science & Business Media, 2013, vol. 3.
- [41] H. Fletcher and J. Steinberg, "Articulation testing methods," *The Bell System Technical Journal*, vol. 8, no. 4, pp. 806–854, 1929.
- [42] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.
- [43] 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.

References

- [44] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [45] H. Å. Gustafsson and S. D. Arlinger, "Masking of speech by amplitude-modulated noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 1, pp. 518–529, 1994.
- [46] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scandinavian audiology*, vol. 11, no. 2, pp. 79–87, 1982.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun. 2016, pp. 770–778.
- [48] D. Hepsiba and J. Justin, "Role of deep neural network in speech enhancement: A review," in *International Conference of the Sri Lanka Association for Artificial Intelligence*. Springer, 2018, pp. 103–112.
- [49] K. Hermus, P. Wambacq, and H. Van Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP journal on advances in signal processing*, vol. 2007, pp. 1–15, 2006.
- [50] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [51] E. Holmes and I. S. Johnsrude, "Speech spoken by familiar people is more resistant to interference by linguistically similar speech," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 46, no. 8, p. 1465, 2020.
- [52] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.
- [53] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [54] T. Houtgast and H. J. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acta Acust. united Ac.*, vol. 25, no. 6, pp. 355–367, Dec. 1971.
- [55] P. A. Howard-Jones and S. Rosen, "Uncomodulated glimpsing in "checker-board" noise," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2915–2922, 1993.
- [56] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.
- [57] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.
- [58] I. Iotzov and L. C. Parra, "Eeg can predict speech intelligibility," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036008, 2019.
- [59] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [60] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 92–102, 2011.

References

- [61] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, Jan. 2014.
- [62] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, no. 5, pp. 1016–1025, 2015.
- [63] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1475–1487, 2011.
- [64] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–446, 2013.
- [65] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," *ICASSP*, pp. 624–628, Mar. 2016.
- [66] M. Karbasi, S. Bleack, and D. Kolossa, "Non-intrusive speech intelligibility prediction using automatic speech recognition derived measures," *arXiv preprint arXiv:2010.08574*, 2020.
- [67] M. Karbasi and D. Kolossa, "Asr-based measures for microscopic speech intelligibility prediction," Aug. 2017.
- [68] M. Karbasi, A. H. Abdelaziz, H. Meutznier, and D. Kolossa, "Blind non-intrusive speech intelligibility prediction using twin-hmms." in *INTERSPEECH*, 2016, pp. 625–629.
- [69] M. Karbasi and D. Kolossa, "A microscopic approach to speech intelligibility prediction using auditory models," in *Proc. Annual Meeting of the German Acoustical Society (DAGA)*. German Acoustical Society Berlin, 2015, pp. 16–19.
- [70] —, "Asr-based speech intelligibility prediction: A review," *Hearing Research*, p. 108606, 2022.
- [71] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.
- [72] —, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, Nov. 2014.
- [73] S. Kay, *Intuitive probability and random processes using MATLAB®*. Springer Science & Business Media, 2006.
- [74] G. Kim and P. C. Loizou, "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1581–1596, 2011.
- [75] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.
- [76] Y. Kim, D. Yun, H. Lee, and S. H. Choi, "A non-intrusive speech intelligibility estimation method based on deep learning using autoencoder features," *IEICE Transactions on Information and Systems*, vol. 103, no. 3, pp. 714–715, 2020.

References

- [77] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.
- [78] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5059–5063.
- [79] K. Kondo, K. Taira, and Y. Kobayashi, "Binaural speech intelligibility estimation using deep neural networks," *Interspeech*, pp. 1858–1862, Sep. 2018.
- [80] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [81] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.
- [82] A. Leijon, "Articulation index and shannon mutual information," in *Hearing—From Sensory Processing to Perception*. Springer, 2007, pp. 525–532.
- [83] D. Lesenfants, J. Vanthornhout, E. Verschueren, L. Decruy, and T. Francart, "Predicting individual speech intelligibility from the cortical tracking of acoustic and phonetic-level speech representations," *Hearing research*, vol. 380, pp. 1–9, 2019.
- [84] H. Levitt and L. R. Rabiner, "Use of a sequential strategy in intelligibility testing," *The Journal of the Acoustical Society of America*, vol. 42, no. 3, pp. 609–612, 1967.
- [85] N. Li and P. C. Loizou, "Factors influencing glimpsing of speech in noise," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1165–1172, 2007.
- [86] —, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.
- [87] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.
- [88] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 47–56, 2010.
- [89] A. MacPherson and M. A. Akeroyd, "Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey," *Trends in Hearing*, vol. 18, p. 2331216514537722, Jun. 2014.
- [90] A. MacPherson, "The factors affecting the psychometric function for speech intelligibility," 2013.
- [91] G. A. Miller, "The masking of speech." *Psychological bulletin*, vol. 44, no. 2, p. 105, 1947.

References

- [92] G. A. Miller and J. C. Licklider, "The intelligibility of interrupted speech," *The Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 167–173, 1950.
- [93] J. P. Moncur and D. Dirks, "Binaural and monaural speech intelligibility in reverberation," *Journal of speech and hearing research*, vol. 10, no. 2, pp. 186–195, 1967.
- [94] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.
- [95] A. K. Nábělek and J. Pickett, "Reception of consonants in a classroom as affected by monaural and binaural listening, noise, reverberation, and hearing aids," *The Journal of the Acoustical Society of America*, vol. 56, no. 2, pp. 628–639, 1974.
- [96] P. B. Nelson, S.-H. Jin, A. E. Carney, and D. A. Nelson, "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 961–968, 2003.
- [97] J. B. Nielsen and T. Dau, "Development of a Danish Speech Intelligibility Test," *Int. J. Audiol.*, vol. 48, no. 10, pp. 729–741, 2009.
- [98] R. Nuthakki, P. Masanta, and T. Yukta, "A literature survey on speech enhancement based on deep neural network technique," *ICCCE 2021*, pp. 7–16, 2022.
- [99] D. O'Shaughnessy, *Speech communications: Human and machine*. Addison-Wesley, New York, 1987.
- [100] K. L. Payton and L. D. Braida, "A method to determine the speech transmission index from speech waveforms," *The Journal of the Acoustical Society of America*, vol. 106, no. 6, pp. 3637–3648, 1999.
- [101] K. L. Payton, L. D. Braida, S. Chen, P. Rosengard, and R. Goldsworthy, "Computing the sti using speech as a probe stimulus," *Past, present and future of the speech transmission index*, pp. 125–138, 2002.
- [102] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A Neural Network for Monaural Intrusive Speech Intelligibility Prediction," *ICASSP*, pp. 336–340, May 2020.
- [103] M. B. Pedersen, A. H. Andersen, S. H. Jensen, Z. H. Tan, and J. Jensen, "Training data-driven speech intelligibility predictors on heterogeneous listening test data," *IEEE Access*, vol. 10, pp. 66 175–66 189, Jun. 2022.
- [104] M. B. Pedersen, M. Kolbæk, A. H. Andersen, S. H. Jensen, and J. Jensen, "End-to-end Speech Intelligibility Prediction Using Time-Domain Fully Convolutional Neural Networks," *INTERSPEECH*, Oct. 2020.
- [105] R. W. Peters, B. C. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 577–587, 1998.
- [106] J. Pickles, "An introduction to the physiology of hearing," in *An Introduction to the Physiology of Hearing*. Brill, 1998.
- [107] C. J. Plack, *The sense of hearing*. Routledge, 2018.

References

- [108] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice-Hall, 1988.
- [109] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.
- [110] H. Relañó-Iborra, A. Chabot-Leclerc, C. Scheidiger, J. Zaar, and T. Dau, "The speech-based envelope power spectrum model (sepsm) family: Development, achievements, and current challenges," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3970–3970, 2017.
- [111] J. Rennies, T. Brand, and B. Kollmeier, "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," *The Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2999–3012, 2011.
- [112] J. Rennies, A. Warzybok, T. Brand, and B. Kollmeier, "Modeling the effects of a single reflection on binaural speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1556–1567, 2014.
- [113] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.
- [114] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3988–3997, 2006.
- [115] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
- [116] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [117] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological review*, vol. 65, no. 6, p. 386, 1958.
- [118] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.
- [119] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [120] A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter, "Objective measures of listening effort: Effects of background noise and noise reduction," 2009.
- [121] M. R. Schädler, A. Warzybok, S. Hochmuth, and B. Kollmeier, "Matrix sentence intelligibility prediction using an automatic speech recognition system," *International Journal of Audiology*, vol. 54, no. sup2, pp. 100–107, 2015.

References

- [122] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind Estimation of the Speech Transmission Index for Speech Quality Prediction," *ICASSP*, pp. 591–595, Apr. 2018.
- [123] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, pp. 84–94, Jun. 2016.
- [124] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1363–1374, 2014.
- [125] M. Sondhi, C. Schmidt, and L. Rabiner, "Improving the quality of a noisy speech signal," *Bell System Technical Journal*, vol. 60, no. 8, pp. 1847–1859, 1981.
- [126] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1358–1362.
- [127] C. Sørensen, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 386–390.
- [128] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51–66, 2018.
- [129] H. J. Steeneken and T. Houtgast, "Mutual dependence of the octave-band weights in predicting speech intelligibility," *Speech communication*, vol. 28, no. 2, pp. 109–123, 1999.
- [130] —, "Basics of the sti measuring method," in *Past, Present, and Future of the Speech Transmission Index, International Symposium on STI, The Netherlands, 2002*, pp. 13–44.
- [131] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.
- [132] V. Sunnydayal, N. Sivaprasad, and T. K. Kumar, "A survey on statistical based single channel speech enhancement techniques," *International Journal of Intelligent Systems and Applications*, vol. 6, no. 12, p. 69, 2014.
- [133] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
- [134] 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)*. IEEE, Mar. 2012, pp. 65–68.
- [135] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, Sep. 2013.

References

- [136] S. Tamura and A. Waibel, "Noise reduction using connectionist models," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1988, pp. 553–556.
- [137] S. J. van Wijngaarden and R. Drullman, "Binaural intelligibility prediction based on the speech transmission index," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4514–4523, 2008.
- [138] S. D. Voran, "A crowdsourced speech intelligibility test that agrees with, has higher repeatability than, lab tests," Institute for Telecommunication Sciences, Tech. Rep., 2017.
- [139] M. B. Winn, D. Wendt, T. Koelewijn, and S. E. Kuchinsky, "Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started," *Trends in hearing*, vol. 22, p. 2331216518800869, 2018.
- [140] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*. Springer, 2016, vol. 1.
- [141] D. Yun, H. Lee, and S. H. Choi, "A deep learning-based approach to non-intrusive objective speech intelligibility estimation," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 4, pp. 1207–1208, 2018.
- [142] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *arXiv preprint arXiv:2111.02363*, 2021.
- [143] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "Stoi-net: A deep learning based non-intrusive speech intelligibility assessment model," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Dec. 2020, pp. 482–486.

References

Part II

Papers

Paper A

A Neural Network for Monaural Intrusive Speech Intelligibility Prediction

Mathias Bach Pedersen, Asger Heidemann Andersen, Søren
Holdt Jensen and Jesper Jensen

The paper has been published in
*Proceedings IEEE International Conference on Acoustics, Speech, and Signal
Processing*, pp. 336–340, 2020.

© 2020 IEEE

The layout has been revised.

Abstract

Monaural intrusive speech intelligibility prediction (SIP) methods aim to predict the speech intelligibility (SI) of a single-microphone noisy and/or processed speech signal using the underlying clean speech signal. In the present work, we propose a neural network for monaural intrusive SIP. The proposed network is trained on data from multiple listening tests to predict SI. In the interest of using the available listening test data as efficiently as possible and to facilitate SI prediction of short duration speech signals, training is based on a local-time intelligibility curve derived from the listening test data. The trained neural network is evaluated, in terms of rank order correlation, against the classical monaural intrusive predictors STOI and ESTOI. The network is found to perform the best overall with a Kendall's tau of 0.825 measured over long duration, i.e. speech signals up to several minutes in duration. For short-term prediction using short speech signals of 1 - 10 seconds the network also shows better performance and smaller prediction variance.

1 Introduction

In recent years, there has been an increasing interest in using Neural Networks (NN) and other data-driven methods to predict the speech intelligibility (SI) of noisy, processed speech signals [1–4]. SI is usually defined as the percentage of intelligible phonemes, words or sentences in a given noisy or processed speech signal. This makes SI a highly relevant aspect of speech signals intended for human listeners. SI is measured by way of listening tests, which require test subjects and are time consuming.

Speech intelligibility prediction (SIP) is concerned with estimating the SI of speech signals algorithmically, i.e. without performing an actual listening test. Classically, SIP methods have been based on measures of similarity between the noisy or processed test speech signal and the underlying clean speech reference signal. This approach is seen as early as in the Articulation Index (AI) [5] and Speech Intelligibility Index (SII) [6], and is still prevalent in modern SI-predictors like the Short Time Objective Intelligibility (STOI) and Extended STOI (ESTOI) [7, 8], the Hearing-Aid Speech Perception Index (HASPI) [9], Speech Intelligibility In Bits (SIIB) [10] and the Spectro-Temporal Modulation Index (STMI) [11]. These methods have all been demonstrated to correlate with measured intelligibility under various conditions.

More recent methods, such as the Binaural SI Model (BiSIM) [12] and binaural STOI, (D)BSTOI [13, 14] have focused on binaural SIP. Binaural methods attempt to explain the improvement in SI observed under conditions, where noise sources are spatially separated from the target talker [15]. Other recent studies including the Speech to Reverberation Modulation energy Ratio (SRMR) [16], Non-Intrusive STOI [17] and convolutional neural networks

for non-intrusive SIP [2] focus on non-intrusive SIP. Non-intrusive methods attempt to predict SI exclusively from the noisy test signal. As such non-intrusive methods can be applied even when no reference signal is available and binaural methods can be applied to a wider variety of conditions than the monaural.

Given the advancements enabled by machine learning methods in areas such as speech enhancement and recognition, see e.g. [18–20], it seems reasonable to expect the application of machine learning in SIP to lead to improvements. Some studies [1–4], have already taken steps in this direction. These studies, however, are all binaural or non-intrusive, and to the best of our knowledge no study exists of a data-driven, monaural, intrusive SI-predictor. Even though binaural, non-intrusive SI-predictors are more versatile, the classical monaural intrusive methods are still the most widely used by far, likely due to their simplicity and tried and true performance. Notably, they are used as evaluation metrics for developing speech enhancement systems, e.g. [21, Part III]. The existing monaural intrusive methods have been studied extensively, so their performance is well known under many conditions. The precision of intrusive SI-predictors can also be expected to be higher than that of non-intrusive SI-predictors, since the latter rely on a subset of the information available to the former. For these reasons performance improvements in monaural intrusive SIP are still very valuable.

Data-driven SIP faces the challenge that little listening test data is available for training, part of the reason being that listening tests are very time consuming. Combining data from different listening tests can also introduce problems: tests may, for instance, differ in scoring, e.g. based on phonemes, words or sentences. Furthermore they may or may not allow repeated listening to the same signal. Finally, the redundancy of the speech material used in the test can also influence SI, e.g. if otherwise unintelligible words can be inferred from context, the SI will be higher. Such factors mean that SI cannot readily be compared between different listening tests without methodology-dependent calibration. Perhaps for these reasons some current studies use data from only a single listening test, [3, 4], or label data with classical SI-predictors, [1]. Ideally, however, to be of practical use, a data-driven SI-predictor should be trained on a larger quantity and greater variety of data than a single listening test offers. Furthermore a data-driven SI-predictor trained on data labelled by another SI-predictor [1] can only be as good as that predictor. The desirable approach is thus ostensibly to train a predictor on a large variety of listening test data.

The present work proposes a monaural, intrusive, data-driven SI-predictor in the form of a neural network. The architecture is inspired in part by the work in [2], which presented a NN for non-intrusive SIP. The proposed SI-predictor is novel in that it is data-driven, monaural and intrusive, and that the ground truth used in training is computed locally in time, rather than

globally from the listening test results. This is done in order to make more efficient use of the limited quantity of listening test data by giving the NN access to more information than a single average within each condition. Classical SI-predictors, e.g. ESTOI [8] and SIIB [10], are evaluated based on their long-term performance, i.e. over many seconds or even several minutes of speech. By training the proposed NN on locally computed SI, we aim to also achieve good performance for shorter speech signals. This would for instance be useful for reducing the computational load of speech processing system evaluation schemes, such as parameter sweeping. Compared to existing data-driven SI-predictors, the proposed predictor is trained on more listening test data than [4], and rather than using STOI predictions as labels like [1] the labels come from listening tests. We demonstrate that the proposed SI-predictor performs better than STOI and ESTOI for both long and short duration speech signals.

2 Neural Network SI-Predictor

2.1 Preprocessing

A neural network for monaural intrusive SIP is proposed. The network uses a preprocessing scheme very similar to the time frequency decomposition used in STOI [7] and ESTOI [8]. This preprocessing is used in part because STOI and ESTOI have been demonstrated to work well, and because it has been successfully used in the data-driven non-intrusive predictor proposed by [2]. This preprocessing consists firstly of a Short-Time discrete Fourier Transform (STFT) of the test, $x[t]$, and reference, $s[t]$, inputs. The STFT uses a 50% overlapping, 25.6 ms Hamming window, $w[t]$. Each window is zero padded to 51.2 ms before the Fourier transform is applied. The short-time discrete Fourier transformed inputs are denoted $X[t, f]$ and $S[t, f]$ respectively. An ideal voice activity detector using the clean reference signal, $S[t, f]$, is then applied to $X[t, f]$ and $S[t, f]$, such that all time steps t_{silent} for which the energy of $S[t_{\text{silent}}, f]$ is less than -40 dB w.r.t. the time step with the highest energy, are removed from both $X[t, f]$ and $S[t, f]$. Finally, a 1/3 octave band transform, [7, eq. (1)], is applied to $X[t, f]$ and $S[t, f]$ leading to the time-frequency representations $\tilde{X}[t, k] \in \mathcal{R}^{Q \times T}$ and $\tilde{S}[t, k] \in \mathcal{R}^{Q \times T}$, respectively. Here Q denotes the number of 1/3 octave bands and T , the number of time instances in these time frequency representations. For more details we refer to [7].

2.2 Architecture

The proposed NN architecture is shown in Fig. A.1. The preprocessed test and reference signals, \tilde{X} and \tilde{S} , with dimension $(Q \times T)$, are fed to the net-

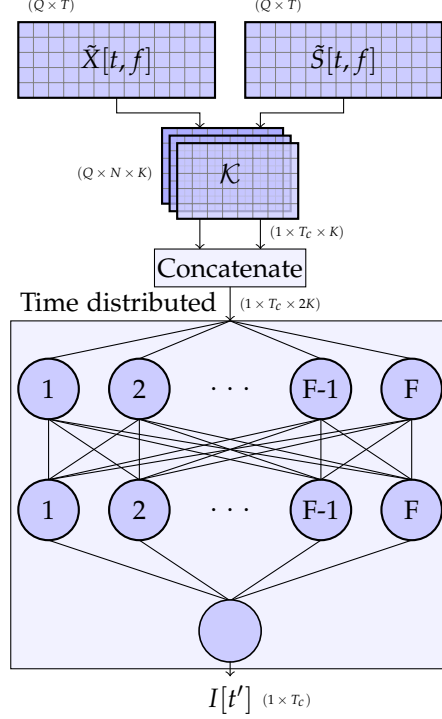


Fig. A.1: Architecture of the proposed network. Preprocessed time-frequency inputs \tilde{X} and \tilde{S} are passed through the convolution layer resulting essentially in two vector time-series. The two signals are concatenated and finally passed through the FC layers, where “Time distributed” means that one time sample at a time is passed through the FC layers. The output, $I[t']$, represents the predicted time-varying SI of the input $x[t]$.

work input. Both are first independently passed through the same convolutional layer. This layer consists of a set, \mathcal{K} , of K kernels of dimension $(Q \times N)$, uses a stride of s and a \tanh activation function. The outputs of the convolutional layer are both of dimension $(1 \times T_c \times K)$ where $T_c = \left\lfloor \frac{T-N+1}{s} \right\rfloor$ and $\lfloor \cdot \rfloor$ denotes the floor function. These outputs can be thought of as vector time series, where each time instance is a vector of kernel activations. \tilde{X} and \tilde{S} are then concatenated along the third axis, the axis of kernel activations, resulting in a signal of dimension $(1 \times T_c \times 2K)$. The concatenated signal is passed through 3 time-distributed Fully Connected (FC) layers. Time distributed means that the FC layers are applied independently for each $(1 \times 1 \times 2K)$ dimensional time instance of the series. The 2 first FC layers have F units and use ReLU activations. The last FC layer has a single unit and uses a sigmoid activation, with the intent of limiting the output range to the interval $(0, 1)$. At the output of the time distributed FC layers the signal can now be

thought of as a scalar time series, denoted by $I[t']$, with dimension $(1 \times T_c)$. $I[t']$ should be interpreted as the predicted time-varying SI throughout $x[t]$.

3 Network Training

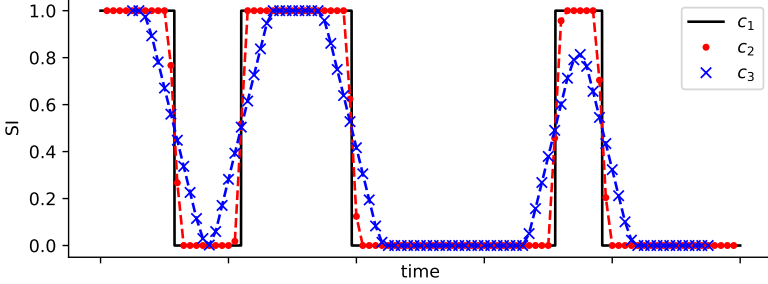


Fig. A.2: An example illustration of the temporal resolution and smoothing of the training target. c_1 is the underlying ground truth word by word binary indicator, c_2 corresponds to the temporal frames of the preprocessed data and c_3 is the local intelligibility curve used as training target.

Given input 1/3 octave band spectrograms $\tilde{X}[t, k]$ and $\tilde{S}[t, k]$ of dimension $(Q \times T)$, the network produces an output, $I[t']$ of dimension T_c . The target during training is a temporal curve of local intelligibility derived from listening test data. This local intelligibility curve is derived from a binary indicator at time sample, t , of whether the current word was correctly identified or not. An example of this curve is illustrated in Fig. A.2, c_1 . To derive the target intelligibility, i.e. the desired output of the network, this binary curve is smoothed by a 50% overlapping rectangular window with the same length as $w[t]$ used in the preprocessing. This reduces the target curve to T time steps matching those of the preprocessed data as illustrated in Fig. A.2, c_2 . A second windowing step follows, once again with a rectangular window, now with length and stride matching the kernels of the convolutional layer. This results in a smoothed curve, e.g. c_3 in Fig. A.2, that can take on values between 1 and 0 and has the same sampling frequency as the network output. As a result of the normalized rectangular windows, a point on this curve corresponds to the proportion of intelligible speech in the underlying time segment. The motivation behind using this curve as the training target, rather than the average intelligibility measured in each test condition, is that more information is preserved. Classically SI is measured by this condition-wide average, but there is no reason to believe that this should be the optimal choice for training the network. Instead, with the proposed scheme, the network is allowed to learn how to use this information on its own.

4 Listening Test Data

The Network is trained using data sets denoted D1 through D4, from four listening tests all using the Dantale II speech corpus [22], with different noise types and processing schemes. The Dantale sentences each consist of five contextually independent words. Listeners in each test were presented with one sentence at a time and asked to identify the words, either choosing from lists of candidates in a software interface, in the cases of D1, D2 and D4, or verbally repeating to a test operator, in the case of D3. All data sets are scored by words correct. The data sets have been chosen because, ESTOI and STOI have been demonstrated to perform well on D2 and D3 respectively, and to cover a wider variety of noise and processing conditions than a single listening test typically covers. Since all the tests use Dantale and similar scoring, it seems reasonable to assume that the measured intelligibility is comparable across data sets.

D1: the first data set is described in [13], as *Experiment 3*. The listening test is binaural with directional sound sources, but contains diotic conditions, 3.2, 3.5 and 3.8 [13, Table II]. The conditions in this data set encompass additive bottling hall noise (BHN) and speech shaped noise (SSN) processed with Ideal Time-Frequency Segregation (ITFS). The SNR before the applied ITFS is in the range -30 to -5 dB. D1 consists of 756 sentences of audio in total.

D2: the second data set is described in [8, Section IV] as *Additive Noise Set I*. This data set consists of various additive noise types modified using sinusoidal intensity modulation (SIM) with modulation frequencies ranging from 4 to 16 Hz. The noise includes SSN, babble (BBL), intensity modulated BBL and machine-gun and destroyer operations room noise from the Noisex database [23]. The SNR is in the range -30 to -5 dB. D2 consists of 2160 sentences of audio in total.

D3: the third data set is described in [24, Section II]. This data set consists of ITFS processed noisy speech. The noise types include SSN, bottling hall noise, café noise and car cabin noise. The SNR before the applied ITFS is in the range -23 to -7 dB. In addition to this range, a -60 dB SNR condition is included for each noise type. It was not possible to associate the test scores with audio word by word in this data set. Instead the training target is defined as the average measured intelligibility. D3 consists of 25200 sentences of audio in total.

D4: the fourth data set is a subset of the listening test data described in [25, Section VI]. Much like D1, this listening test is binaural, with some diotic conditions. The diotic conditions are those that use binaural beamformers and where both target and noise are exactly frontal. These are the conditions included in D4. The noise type is BBL. The SNR ranges from -17 to -8 dB. D4 consists of 880 sentences of audio in total.

5. Results

Table A.1: Data set summary.

	Noise	Proc.	SNR [dB]	Sentences	Cf.
D1	SSN, BHN	ITFS	$(-30, -5)$	756	[13]
D2	NOISEX, SSN, BBL	none	$(-30, -5)$	2160	[8]
D3	SSN, BHN, café, car	ITFS	$(-60, -7)$	25200	[24]
D4	BBL	beamform.	$(-17, -8)$	880	[25]
D5	The union of the data sets, D1 through D4, listed above				

D5: Finally denote by D5 the union of D1 through D4.

5 Results

5.1 Training Details and SIP Performance

The NN was trained on D5 using $K = 200$ kernels of width $N = 30$ with a stride of $s = 1$, and $F = 200$ nodes in the FC layers. For the preprocessing, $Q = 17$ bands were used in the 1/3 octave band transform. Approximately 15% of D5 is set aside as the test set, in such a way that it contains an equal amount of data from each listening test condition. The remaining 85% of D5 is used as the training set. This means that both the test and training sets contain all of the noise and processing conditions available in the listening tests, but that the noise realizations are different. To compensate for differences in audio duration available per condition, the different listening tests are individually weighted in the loss function during training. The weights are computed such that each noise/processing condition is of equal importance. The NN is trained using the ADAM optimization algorithm, [26], to minimize the binary cross entropy cost function between the network output and target. For the purpose of efficient training, D5 was arranged into matrices of dimension (17×512) . Each training batch was made up of 256 pairs of these matrices, i.e. $\tilde{X}[t, k]$ and $\tilde{S}[t, k]$. The proposed NN is compared against the STOI [7] and ESTOI [8] SI-predictors. This comparison is made both within individual data sets, D1-D4, and across the combined data set, D5. The performance is evaluated by Kendall's rank correlation coefficient, τ , [27]. This coefficient takes on values in the interval $[-1, 1]$, and indicates the degree of monotonicity in the relation between measurements and predictions. As such, higher values correspond to better performance. This coefficient is used, because SI-predictors should be able to rank speech signals according to SI.

5.2 Long-Term SIP Performance

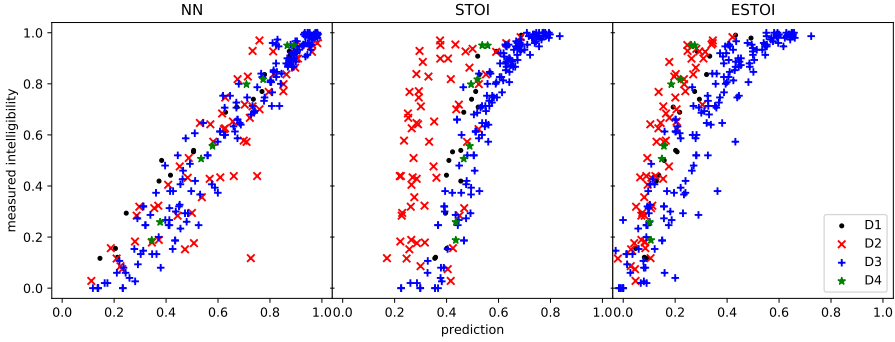


Fig. A.3: Intelligibility as measured over each listening test condition is plotted against predicted intelligibility by the NN, STOI and ESTOI.

Table A.2 shows the τ scores for each SI-predictor on each data set. The NN achieves the highest τ overall, and is only outperformed by ESTOI on D2. Fig. A.3 provides scatter plots of the averaged predictions within each test condition, for each SI-predictor. Each point corresponding to the average over one noise/processing condition. Comparing the plots in Fig. A.3, the neural network seems to be better at predicting absolute intelligibility. This is likely due to the fact that training and testing is based on data from the same listening tests, which has enabled the network to learn the corresponding mapping to absolute intelligibility. For different speech material this mapping will be different; hence the predictions should not be interpreted as absolute intelligibility, but rather treated as an index, i.e. expected to have a monotonous relation to absolute intelligibility. The NN also makes a noticeable error on a particular condition from D2, predicting an SI of .7, where measured SI is .1. This condition contains machine gun noise from NOISEX, a rather unique noise type in D5, which could explain why the NN has not acquired good performance for it. STOI can be seen, by Table A.2 and by the distinct cluster of red x's in Fig. A.3 to perform poorly on D2, which contains speech in modulated noise. It is well known that STOI may perform poorly for such noise types [8, 28].

5.3 Short-Term SIP Performance

The short term performance of the NN SI-predictor is evaluated by sampling short audio clips from each test condition and computing Kendall's τ . Fig. A.4 shows the performance in terms of τ as a function of audio length for the proposed NN, STOI and ESTOI. Each curve shows the average of Kendall's τ along with the 2.5 and 97.5 percentiles based on 250 trials. Each trial consists

6. Conclusion

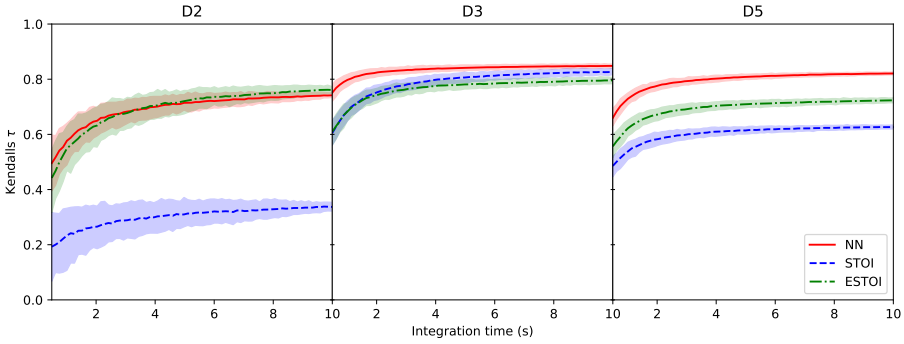


Fig. A.4: To evaluate the short-term performance Kendall’s τ versus audio length for the data sets D2, D3 and D5. D1 and D4 are omitted here since they are only subsets and thus too short to produce meaningful plots of this kind on their own.

Table A.2: Kendall’s τ computed within each data set.

	D1	D2	D3	D4	D5
NN	.948	.745	.854	1.00	.825
STOI	.856	.344	.838	.857	.635
ESTOI	.817	.762	.812	.929	.731

of randomly selecting an audio segment from each condition in the data set, for which Kendall’s τ is computed. The horizontal axis shows the duration of the sampled audio segments. These curves show that in terms of monotonicity, the predictors all reach a stable performance within 10 seconds of audio. Notably, the NN shows better performance at short audio lengths than STOI and ESTOI even when testing with D2, where ESTOI is seen to have slightly better long-term performance. The local ground truth intelligibility varies significantly in this data set, as a result of the low-frequency intensity modulations of the noise. For short inputs this leads to larger prediction errors w.r.t. the measured long term intelligibility, as is evident by the comparatively large percentiles in this figure. Notice also that for D3 the proposed SI predictor achieves smaller prediction spread as compared to STOI and ESTOI. Finally for the combined data set D5, the proposed predictor reaches the long term performance of STOI and ESTOI for speech signals as short as one second.

6 Conclusion

A neural network for monaural, intrusive speech intelligibility prediction was proposed. The network was trained on existing listening test data, to output

a curve of local intelligibility predictions. The performance of the network was evaluated and compared to that of the STOI and ESTOI predictions. The performance of the network, evaluated through Kendall's τ , for long speech signals was demonstrated to be better than that of STOI and ESTOI for most of the data sets, and significantly better on average. The network was also shown to perform better than STOI and ESTOI for short speech signals.

Acknowledgements

This work is funded by the Independent Research Fund Denmark. Project ID. DFF - 7017-00017.

References

- [1] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, pp. 84–94, Jun. 2016.
- [2] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.
- [3] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," *ICASSP*, pp. 624–628, Mar. 2016.
- [4] K. Kondo, K. Taira, and Y. Kobayashi, "Binaural speech intelligibility estimation using deep neural networks," *Interspeech*, pp. 1858–1862, Sep. 2018.
- [5] A. S3.5-1969, "Methods for the calculation of the articulation index," *American National Standards Institute, New York*, 1969.
- [6] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.
- [7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
- [8] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

References

- [9] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, Nov. 2014.
- [10] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.
- [11] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2719–2732, Oct. 1999.
- [12] S. Cosentino, T. Marquardt, D. McAlpine, J. F. Culling, and T. H. Falk, "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals," *J. Acoust. Soc. Am.*, vol. 135, no. 2, pp. 796–807, Feb. 2014.
- [13] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Predicting the Intelligibility of Noisy and Nonlinearly Processed Binaural Speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.
- [14] —, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Commun.*, vol. 102, pp. 1–13, Sep. 2018.
- [15] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acust. United Ac.*, vol. 86, no. 1, pp. 117–128, Jan. 2000.
- [16] T. H. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Aug. 2010.
- [17] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," Mar. 2017, pp. 5085–5089.
- [18] M. Kolbæk, Z. H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.
- [19] L. Lightburn and M. Brookes, "SOBM - a binary mask for noisy speech that optimises an objective intelligibility metric," Apr. 2015, pp. 5078–5082.

References

- [20] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2604–2612, May 2016.
- [21] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.
- [22] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a danish sentence test in noise," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, Jan. 2003.
- [23] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
- [24] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.
- [25] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for bin-aural hearing aids," *J. Acoust. Soc. Am.*, vol. 145, May 2019.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, Dec. 2014.
- [27] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun. 1938.
- [28] S. Jørgensen, R. Decorsière, and T. Dau, "Effects of manipulating the signal-to-noise envelope power ratio on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 137, no. 3, pp. 1401–1410, Mar. 2015.

Paper B

End-to-end Speech Intelligibility Prediction Using Time-Domain Fully Convolutional Neural Networks

Mathias Bach Pedersen, Morten Kolbæk, Asger Heidemann
Andersen, Søren Holdt Jensen and Jesper Jensen

The paper has been published in
Proceedings of Interspeech, pp. 1151–1155, 2020.

© 2020 ISCA

The layout has been revised.

Abstract

Data-driven speech intelligibility prediction has been slow to take off. Datasets of measured speech intelligibility are scarce, and so current models are relatively small and rely on hand-picked features. Classical predictors based on psychoacoustic models and heuristics are still the state-of-the-art. This work proposes a U-Net inspired fully convolutional neural network architecture, NSIP, trained and tested on ten datasets to predict intelligibility of time-domain speech. The architecture is compared to a frequency domain data-driven predictor and to the classical state-of-the-art predictors STOI, ESTOI, HASPI and SIIB. The performance of NSIP is found to be superior for datasets seen in the training phase. On unseen datasets NSIP reaches performance comparable to classical predictors.

1 Introduction

Data-driven speech enhancement has garnered huge interest in the last decade with studies such as [1–6]. A more recent trend has been towards end-to-end solutions like [7–10], working fully in the time-domain. Most of these speech enhancement studies aim at enhancing speech intelligibility (SI), either in the evaluation or even as part of the objective. SI is a very relevant aspect of processed speech intended for human listeners, e.g. telecommunication systems and hearing assistive devices. Unfortunately, SI is time consuming to measure and hence speech intelligibility prediction (SIP) is of great importance to the field of speech enhancement in particular, and to the broader area of speech processing in general. SIP as a field however, has not seen the same rapid advancement in terms of data-driven methods as other fields in speech processing.

Presently, data driven SIP has only been attempted with relatively small datasets, and partially data-driven models using hand-engineered features [11–16]. Why is this? One of the main reasons is certainly that data-driven SIP is limited by data scarcity. In most other speech processing fields ground truth data is simply clean speech signals, which are relatively easily obtainable in bulk. Obtaining training data for SIP, however, requires time-consuming measurements of speech intelligibility through listening tests of individual noise/processing conditions. Thus the availability of speech data accompanied by subjectively measured SI is rather low.

Most state of the art SI-predictors like STOI [17], ESTOI [18], SIIB [19] and HASPI [20], are still not based on machine learning, but rather on psychoacoustic models and heuristics, and validated empirically using relatively small datasets with measured intelligibility. In spite of their non-data-driven design, these predictors have demonstrated excellent performance in a variety of noise and processing conditions, and remain among the most widely

used. An overview of classical predictors is presented in [21]. It is, however, not fully understood exactly under which conditions these predictors perform well.

Some *data-driven* SI-predictors have been proposed, but they are all limited in one way or another. In [11–13] existing non-data-driven intelligibility predictors are used to either label the training data or as part of the architecture respectively. The systems in [14–16] are trained with measured intelligibility, though [14] uses data from a single listening test. These systems all rely on hand determined features, i.e. Mel frequency bands in [14], and 1/3-octave bands in [15, 16].

In this paper we propose and analyse the performance of an intrusive end-to-end speech deep neural network (DNN) intelligibility predictor. The network is a fully convolutional architecture inspired by U-Net [22] and resembles that used in a large body of literature including works involving speech enhancement (e.g. [7, 23, 24]). This network is trained and tested on speech and SI measurements of a wide variety of conditions from a range of listening tests. The network takes time-domain speech signals along with the corresponding clean speech as input and outputs SI-predictions as a function of time, and is thus an end-to-end data-driven SI-predictor. The architecture is explained in greater detail in Section 2 and the data and simulations are described in Section 3. The predictor is tested in a comparison with ESTOI, SIIB and HASPI, using the Pearson and Spearman correlation within each listening test. The results are presented in Section 4, and the conclusion in Section 5.

2 Data-driven Intelligibility Prediction

In this study we use a data-driven approach for speech intelligibility prediction. Specifically, we propose the neural speech intelligibility predictor (NSIP) model given by Fig. B.1, which shows the architecture of an end-to-end intrusive speech intelligibility predictor based on fully convolutional neural networks.

2.1 Intrusive Speech Intelligibility Prediction

Intrusive SIP refers to the problem of estimating the SI of a noisy/processed speech signal, $x[t]$, using $x[t]$ itself and the corresponding clean speech signal, $s[t]$. Intrusive SI-predictors are classically more successful than their non-intrusive counterparts, which only rely on $x[t]$. Intrusive prediction can use $s[t]$ as a reference to measure how dissimilar $x[t]$ is to clean speech, while non-intrusive prediction requires a built-in model of generic clean speech in order to make such a comparison. This makes the classical intrusive predic-

2. Data-driven Intelligibility Prediction

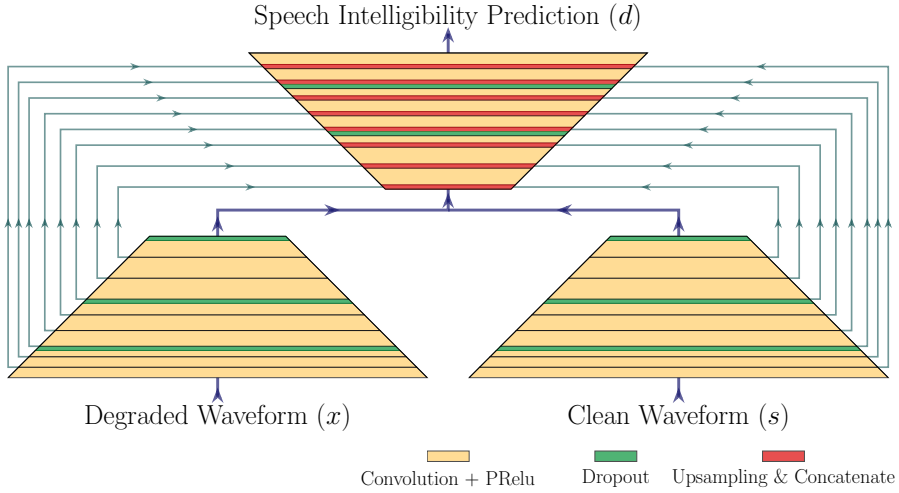


Fig. B.1: Architecture of an intrusive neural speech intelligibility predictor based on fully convolutional neural networks. The predictor is trained end-to-end to estimate the sample-level speech intelligibility of a degraded speech waveform.

tors simpler and more robust. In transitioning to DNN's, the argument of simplicity changes, because DNN's rely on their great parametric complexity in the first place. This makes non-intrusive architectures somewhat simpler, because they only need to work with one input rather than two. Intrusive architectures still have the potential to be more robust though, and because of the data scarcity, the extra clean speech input might be valuable.

The network architecture used in this paper is intrusive, since it receives the inputs, $s[t]$ and $x[t]$, which in this context are time-domain clean and noisy/distorted speech signals. The desired output is defined as a time domain piece-wise constant curve, $d[t]$, corresponding to measured SI of the input $x[t]$, as it is also done in [16]. The network output can then be integrated over time to produce an SI prediction for a particular span of time.

2.2 Neural Speech Intelligibility Prediction

The NSIP model depicted in Fig. B.1 is based on a fully convolutional neural network architecture with 18 convolutional layers utilizing parameterized ReLU (PReLU) activation functions between the layers [25]. The model is inspired by U-Net [22] and follows an encoder-decoder methodology where skip-connections are applied between corresponding layers to allow data at various sample rates to flow between the encoder and decoder.

Differently from a standard U-net, the proposed model has two encoders, as shown in Fig. B.1, one for the clean and one for the degraded speech wave-

Model	#filters in encoder layers 1 – 9				#filters in decoder layers 10 – 18				#Params (millions)
	1 – 3	4 – 6	7 – 8	9	10 – 11	12 – 14	15 – 17	18	
NSIP1	6	12	16	32	32	16	12	1	0.122M
NSIP2	8	16	24	64	64	24	16	1	0.349M
NSIP3	12	18	36	80	80	36	18	1	0.603M
NSIP4	12	24	48	96	96	48	24	1	0.946M
NSIP5	16	32	64	128	128	64	32	1	1.68M

Table B.1: Number of output filters in each layer of the NSIP-model given by Fig. B.1 for five different configurations. All filters are 11 samples long.

forms, since intrusive speech intelligibility prediction can make use of both of these. Specifically, the two encoders each contain eight convolutional layers and the output of the two encoders, which contain compressed information about the clean and degraded speech signals, are concatenated and propagated to a joint decoder that performs the final SI prediction. The encoders both use a stride of two in each layer, except for the first layer where a stride of one is used. This drives the final dimension at the outputs of the encoders to be compressed with a factor of 256. Similarly, all layers in the decoder, except for the last layer, use upsampling with a factor of two, such that the final output has the same dimension as the inputs, which allows sample-level SI prediction.

To study how the number of parameters influence the SI performance of the proposed architecture, five NSIP models are trained and evaluated with a varying number of filters. The configurations of the individual NSIP systems are shown in Table B.1. The number of parameters for the five models vary from 0.122×10^6 to 1.68×10^6 , which is comparable to the 0.224×10^6 parameters of a recently published frequency-domain technique [16] that will serve as an NSIP baseline in Sec. 4. Finally, all filters have a size of 11 samples.

The SIP-systems are trained to minimize the binary cross entropy between estimated and measured intelligibility using the ADAM optimizer [26] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and an initial learning rate of 0.0005, which is controlled by a learning rate schedule that reduces the learning rate with a factor of two, if the validation loss has not decreased for two epochs. Finally, during training, 20% dropout is applied for every third layer, and a batch size of 16 is used. Training is stopped, if the validation loss has not decreased for five epochs or a maximum of 200 epochs has elapsed.

The SIP-systems have been implemented using Keras¹ with a TensorFlow²

¹<https://keras.io/>

²<https://tensorflow.org/>

backend and the python implementation of the trained NSIP-models, are available online³, to allow interested readers, to use and evaluate the models further.

3 Experimental Design

To establish the potential of the proposed architecture in terms of predicting speech intelligibility of noisy/distorted speech, a series of experiments are conducted. In the following, the datasets used for training, validation, and test are presented.

3.1 Training, Validation and Test Data

Table B.2 summarizes the ten datasets used for training, validating and testing the NSIP-models. The data consist of clean and noisy/distorted speech signals and measured SI scores, which are used as labels. Due to the number of datasets, space limitations make it impractical to give a detailed description of each listening test here. Since they are all well described in other works, we instead refer the interested reader to the respective sources. The datasets contain multiple talkers, languages, noise types and processing schemes. Classical predictors have shown varying performance on different subsets of these datasets, which is also verified in Section 4. There are significant differences in the size of these datasets, and Table B.2 contains a breakdown of the size (#files) and number of different acoustic conditions (#cond.) in each dataset. Because of the limited amount of data, we do not attempt to balance the datasets by excluding data from the bigger datasets.

3.2 Cross Validation

Datasets 0 – 6 have been split randomly into training, validation and test comprised of approximately 80, 10, and 10 % of the data, respectively. Each listening test condition has been split in this way, such that every condition is represented in the test set. Furthermore, due to the limited amount of test data available, 10-fold cross validation has been performed and for each split of the data into training, validation, and test, ten differently initialized sets of NN-weights have been trained. In other words, 100 models of each architecture have been trained. Finally, to demonstrate the performance in unseen conditions datasets 7 – 9 have been left out of the training and validation sets, and are used exclusively for testing. As such we distinguish between *seen* conditions, i.e. belonging to 0 – 6 and *unseen* conditions belonging to 7 – 9.

³https://git.its.aau.dk/mok/neural_sip.git

Dataset		Training		Validation		Test	
No.	Ref.	#files	#cond.	#files	#cond.	#files	#cond.
0	[18]	564	60	60	58	60	58
1	[27]	6295	168	673	168	840	168
2	[17]	320	34	35	32	35	32
3	[15]	1744	327	77	76	318	299
4	[28]	784	24	96	24	96	24
5	[29]	439	18	54	18	54	18
6	[18]	3460	20	436	20	437	20
7	[30]	0	0	0	0	278	9
8	[31]	0	0	0	0	241	20
9	[32, 33]	0	0	0	0	64	52

Table B.2: Datasets used for training, validation and test. Each file corresponds to approx. 6.6s of speech. See references for further details regarding the general design of the datasets.

4 Experimental Results

4.1 End-to-end Data-driven Intelligibility Prediction

The NSIP-models defined in Table B.1 have been evaluated using Spearman and Pearson correlation. The models were given the clean references and corresponding noisy/processed test data signals, and the predictions were integrated over each acoustic condition. Examples of these integrated predictions can be seen, compared to measured SI, in Figure B.2. The Spearman and Pearson scores were then computed and are presented in Tables B.3 and B.4 with standard deviations from the cross-validation reported in parentheses. Spearman is a rank correlation and measures monotonicity between predictions and measurements, whereas Pearson correlation measures the linearity of their relationship. For each dataset the Spearman and Pearson correlation of the NSIP predictions are measured.

From Tables B.3 and B.4 it is seen that NSIP5 with 1.68×10^6 parameters reaches an average Spearman of .91 across seen conditions and .85 across unseen conditions, with corresponding average Pearson correlations of .91 across seen conditions and .85 across unseen conditions. The performance of NSIP5 is visualized for a few datasets in Figure B.2.

4. Experimental Results

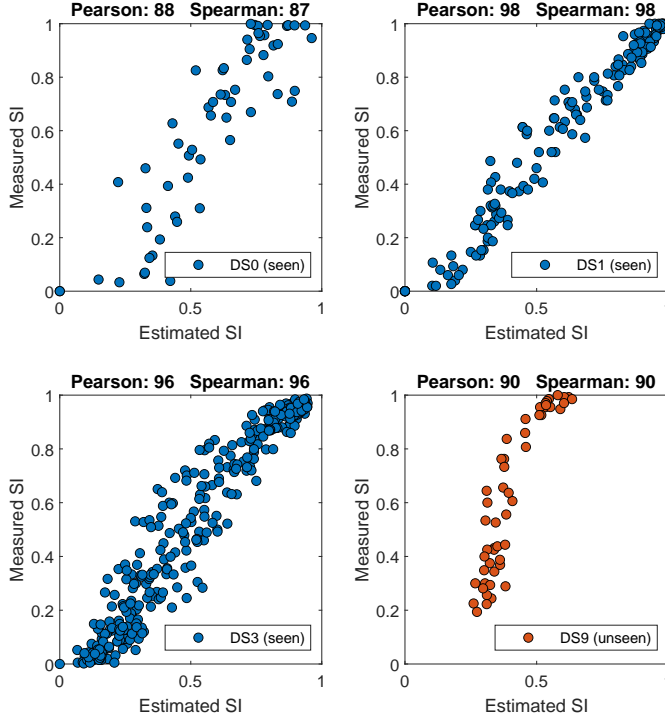


Fig. B.2: Scatter plots showing relation between measured SI and estimated SI, estimated by the NSIP5 system, for seen datasets DS0, DS1 and DS3, as well as the unseen dataset DS9. The Pearson and Spearman correlations are scaled a factor of 100.

4.2 Data-driven vs. Non-data-driven SIP

We compare the results from the NSIP-models on the test data with the classical predictors STOI, ESTOI, HASPI and SIIB, and a retrained network with the architecture of [16]. Similar to STOI and ESTOI, this architecture takes 1/3-octave band representations of s and x as inputs and outputs SI-predictions, and as such can be used as a frequency-domain benchmark. Tables B.3 and B.4 show the dataset-wise results in terms of Spearman and Pearson correlation respectively, for the NSIP-models and the classical predictors. We distinguish between the conditions which have and have not been seen by the NSIP-models during training, and report the average of the performance measures across these subsets as well. We stress that “seen” conditions are not training data, but distinct test data signals belonging to listening test conditions that also appear in the training set. In the case of Pearson correlation, a dataset dependent logistic curve is often fitted to the predictions before computing the correlation. This function has been used to

Paper B.

Spearman $\times 100$													
Predictor	Mean	Mean	Seen Data						Unseen Data			#Params (millions)	
	(seen)	(unseen)	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8		DS9
NSIP1 (time) :	82 (2.9)	85 (7.1)	76 (5.2)	96 (0.4)	78 (2.3)	93 (1.1)	57 (3.4)	74 (7.1)	98 (0.6)	97 (1.6)	77 (10.7)	80 (9.0)	0.122M
NSIP2 (time) :	85 (2.3)	82 (5.1)	84 (2.8)	97 (0.2)	81 (2.1)	95 (0.6)	64 (4.9)	76 (5.2)	98 (0.4)	98 (1.1)	64 (9.9)	85 (4.3)	0.349M
NSIP3 (time) :	88 (2.2)	83 (4.6)	87 (1.8)	98 (0.1)	82 (1.7)	96 (0.4)	73 (6.1)	80 (4.9)	99 (0.3)	97 (1.1)	64 (10.0)	87 (2.6)	0.603M
NSIP4 (time) :	89 (2.2)	85 (3.8)	87 (1.7)	98 (0.1)	83 (1.8)	96 (0.4)	81 (6.2)	81 (5.0)	99 (0.2)	98 (1.1)	69 (7.6)	87 (2.7)	0.946M
NSIP5 (time) :	91 (2.1)	85 (3.5)	88 (1.7)	98 (0.1)	84 (1.8)	96 (0.4)	87 (5.9)	83 (4.7)	99 (0.3)	97 (1.0)	70 (7.3)	89 (2.2)	1.68M
NSIP6 (freq) :	88 (1.9)	74 (4.7)	79 (3.7)	97 (0.1)	81 (1.4)	96 (0.6)	82 (4.1)	83 (3.0)	97 (0.4)	96 (1.9)	70 (5.1)	56 (7.2)	0.224M
STOI:	74	93	47	96	60	81	57	83	98	95	96	87	–
ESTOI:	78	92	82	96	49	84	56	86	96	98	95	85	–
HASPI:	71	88	62	78	50	93	64	65	84	98	96	70	–
SIIB:	80	96	73	91	39	93	75	94	98	98	97	94	–

Table B.3: Spearman correlation for NSIP models and classical non-data-driven SIP techniques. NSIP1-5 are time-domain models configured according to Fig. B.1 and Table B.1 and NSIP6 are an frequency-domain baseline model from [16]. All models are trained with data according to Table B.2. The score are mean scores computed based on 10-fold cross validation and the scores in parenthesis are standard deviations.

Pearson Correlation $\times 100$													
Predictor	Mean	Mean	Seen Data							Unseen Data			#Params (millions)
	(seen)	(unseen)	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	
NSIP1 (time) :	84 (2.7)	83 (7.0)	75 (4.5)	96 (0.4)	77 (2.5)	93 (1.1)	77 (3.7)	76 (5.9)	97 (0.6)	95 (2.1)	76 (10.3)	77 (8.7)	0.122M
NSIP2 (time) :	88 (1.7)	80 (6.1)	83 (2.8)	97 (0.3)	80 (1.7)	95 (0.6)	87 (1.6)	79 (4.7)	98 (0.4)	97 (1.1)	62 (13.0)	83 (4.2)	0.349M
NSIP3 (time) :	90 (1.4)	81 (5.7)	86 (2.1)	98 (0.2)	81 (1.4)	96 (0.4)	89 (1.0)	82 (4.2)	98 (0.2)	96 (1.4)	62 (12.8)	85 (2.8)	0.603M
NSIP4 (time) :	91 (1.2)	84 (4.1)	87 (1.9)	98 (0.2)	82 (1.4)	96 (0.4)	90 (0.8)	83 (3.7)	99 (0.2)	97 (1.3)	69 (8.5)	86 (2.7)	0.946M
NSIP5 (time) :	91 (1.1)	85 (3.7)	89 (1.6)	98 (0.1)	83 (1.2)	96 (0.4)	91 (0.8)	85 (3.5)	99 (0.2)	96 (1.3)	71 (8.0)	87 (1.7)	1.68M
NSIP6 (freq) :	89 (1.2)	73 (5.2)	77 (3.8)	97 (0.1)	79 (1.1)	96 (0.6)	91 (0.7)	86 (2.1)	98 (0.2)	93 (2.0)	70 (7.1)	57 (6.5)	0.224M
STOI:	77	92	51	91	56	78	80	85	98	98	89	90	–
ESTOI:	79	92	77	93	44	80	81	86	95	97	93	86	–
HASPI:	62	80	42	77	45	85	37	69	81	91	74	76	–
SIIB:	77	88	62	85	32	80	89	95	94	96	77	90	–
STOI (fitted):	78	96	51	96	58	80	76	85	99	99	96	91	–
ESTOI (fitted):	81	94	83	95	45	82	78	87	97	100	95	88	–
HASPI (fitted):	65	89	61	77	45	88	36	70	80	97	93	78	–
SIIB (fitted):	82	97	74	90	33	92	92	95	98	99	95	96	–

Table B.4: As Table B.3 but for Pearson correlation.

map SI-predictions to measurements by [17, 19]. We do this for the classical predictors, and the Pearson correlations denoted by (fitted) in Table B.4 thus measure the correlation in a logistic rather than linear sense. This increases their average Pearson correlation, but in the seen conditions, even with the added dataset-specific knowledge, they are still outperformed by the NSIP architectures, which has been given no such dataset-specific mapping.

The NSIP-models achieve better average performance, in terms of Spearman and Pearson correlation in seen conditions as compared to the classical predictors. Comparing the measures for the unseen datasets, NSIP is on par with the classical methods for datasets 7 and 9, but not dataset 8. Consequently, the average NSIP performance on the unseen datasets is lower than average performance of the classical predictors on the same datasets.

4.3 Frequency-domain Data-driven SIP

In order to judge the potential advantage of an end-to-end architecture, we compare NSIP to the architecture of [16], which takes 1/3-octave band transformed speech signals as inputs, similar to STOI and ESTOI. This architecture has been retrained on the same data as the proposed time-domain NSIP architecture. This is done to gauge the advantage of NSIP's access to the full information in the time-domain. As was the case for the time-domain architecture, the frequency-domain architecture is trained and tested on the ten cross validation data-splits. The test results are shown in the rows labelled NSIP6 (freq) in Tables B.3 and B.4. It appears that the time-domain architectures of similar parameter size perform slightly better on average in terms of Spearman and Pearson on the unseen Datasets 7 and 8, and significantly better on Dataset 9. This could be due to the loss of information in the 1/3-octave band transform employed in NSIP6. On the seen datasets the frequency-domain architecture performs as well as NSIP3 and 4.

5 Conclusion

We proposed a time-domain neural speech intelligibility predictor (NSIP) based on a fully convolutional neural network architecture, for intrusive speech intelligibility prediction. This network was trained on seven listening test datasets and tested on ten. Performance was evaluated in terms of Spearman and Pearson correlation, and compared to the classical predictors STOI, ESTOI, HASPI and SIIB, and a retrained frequency-domain architecture, [16]. The NSIP architectures showed the best performance on the seven seen datasets, but were outperformed by the classical predictors on one of the unseen datasets. The frequency-domain architecture was found to reach performance similar to that of larger, in terms of parameters, time-domain architectures, with much fewer parameters.

References

- [1] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.
- [2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.
- [3] E. W. Healy, S. E. Yoho, J. Chen, Y. Wang, and D. Wang, "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel

References

- segments of the same noise type," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1660–1669, 2015.
- [4] M. Kolbæk, Z. H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.
- [5] M. Kolbæk, Z. Tan, and J. Jensen, "On the Relationship Between Short-Time Objective Intelligibility and Short-Time Spectral-Amplitude Mean-Square Error for Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 283–295, 2019.
- [6] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [7] A. Pandey and D. Wang, "A New Framework for CNN-Based Speech Enhancement in the Time Domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.
- [8] S. W. Fu, T. W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 570 – 1584, 2018.
- [9] S. R. Park and J. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," in *Proc. Interspeech*, 2017, pp. 1993–1997.
- [10] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 825–838, 2020.
- [11] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, pp. 84–94, Jun. 2016.
- [12] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," *ICASSP*, pp. 624–628, Mar. 2016.
- [13] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind Estimation of the Speech Transmission Index for Speech Quality Prediction," *ICASSP*, pp. 591–595, Apr. 2018.

- [14] K. Kondo, K. Taira, and Y. Kobayashi, "Binaural speech intelligibility estimation using deep neural networks," *Interspeech*, pp. 1858–1862, Sep. 2018.
- [15] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.
- [16] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A Neural Network for Monaural Intrusive Speech Intelligibility Prediction," *ICASSP*, pp. 336–340, May 2020.
- [17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
- [18] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [19] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.
- [20] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, Nov. 2014.
- [21] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Processing Magazine*, pp. 114–124, Feb. 2015.
- [22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015, pp. 234–241.
- [23] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.
- [24] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," 2019. [Online]. Available: <http://arxiv.org/abs/1909.01019>

References

- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proc. ICCV*, 2015, pp. 1026–1034.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, Dec. 2014.
- [27] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.
- [28] T. Bentsen, A. A. Kressner, T. Dau, and T. May, "The impact of exploiting spectro-temporal context in computational speech segregation," *J. Acoust. Soc. Am.*, vol. 143, no. 1, pp. 248–259, Jan. 2018.
- [29] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.
- [30] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 145, May 2019.
- [31] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Predicting the Intelligibility of Noisy and Nonlinearly Processed Binaural Speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.
- [32] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, Mar. 2014.
- [33] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time sii," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 5, pp. 851–862, May 2015.

Paper C

Training Data-Driven Speech Intelligibility Predictors on Heterogeneous Listening Test Data

Mathias Bach Pedersen, Asger Heidemann Andersen, Søren
Holdt Jensen, Zheng-Hua Tan and Jesper Jensen

The paper has been published in
IEEE Access Vol. 10, pp. 66175–66189, 2022.

© 2022 IEEE

The layout has been revised.

Abstract

Prediction of Speech Intelligibility (SI) is a topic of interest for most speech processing applications, where intelligibility is of any importance, e.g., speech coding, transmission and enhancement. Traditionally, SI predictors have been based on signal processing methods and heuristics, but more recently, an increasing number of data-driven SI-predictors have been proposed. Data-driven prediction of SI requires large quantities of labelled data, ideally from many listening tests. Listening tests differ in factors such as vocabulary, talker, listener’s task, etc. collectively referred to as the paradigm. A naïve strategy of training SI-predictors directly on stimuli, pooled from different listening tests, is futile because the exact map from the stimulus to SI is determined, not only by the stimulus, but also by the paradigm. Data-driven SI-predictors trained in this way become specialized to the paradigms of the training data by erroneously attributing all paradigm influences on SI to the stimulus. The problem is fundamental and persists even in the idealized situation where training data is abundant. We propose a strategy for training data-driven SI-predictors that is independent of the paradigms, underlying the training data. The proposed strategy is to concatenate an SI-predictor and a layer of trainable dataset-specific mapping functions, each corresponding to a single paradigm in the training data. These mapping functions are trained jointly with the SI-predictor and serve to efficiently approximate the psychometric functions implied by each paradigm. The mapping functions prevent the predictor from specializing to these paradigms during training. We present an SI-predictor with a novel architecture that incorporates a convolutional network and an ESTOI back-end, train it with this strategy, compare it to naïve training and a range of existing non-data-driven predictors. The proposed training strategy and architecture results in higher performance overall and increased robustness to unseen paradigms.

1 Introduction

Speech Intelligibility (SI) is an important concept for speech communication devices, such as hearing aid systems or devices for communicating under extreme acoustic conditions, such as aeroplane cockpits or emergency response situations. Because of this, SI is repeatedly measured during the development of these devices. The most reliable measurements of SI come from listening tests, where human listeners respond to examples of the noisy or processed speech in question. Since many human listeners need to be involved, these listening tests are significant time-sinks, slowing down iterative development of speech processing methods, concerned with SI.

To speed up this development, SI-prediction has become a popular and valuable tool. SI-prediction refers to algorithms or models, designed to predict the SI of noisy or processed speech signals, as it would be rated by a

panel of human listeners. SI-prediction offers fast and reproducible results and can significantly increase the speed of development for speech processing systems, when used in place of listening tests. A potential disadvantage is that SI-predictors, like any predictor or estimator, may exhibit variable accuracy, depending on variations in the signals under study. These variations include, for instance, the type and intensity of noise or distortions deteriorating the signals and the type of processing applied, if any. We refer to a particular combination of these variations as a *listening condition*. Applying an SI-predictor to listening conditions, on which it has not been validated by a listening test, can give misleading results. Robustness to a wide variety of listening conditions is thus an important quality in SI-prediction.

Data driven SI-predictors are designed using machine learning methods, such as neural networks. These predictors usually have a large number of parameters, which are optimized through training on labelled speech in different listening conditions. When data-driven SI-predictors are trained on speech data from a set of listening test conditions, it makes sense to refer to these listening conditions as *seen* conditions for that predictor. This is in contrast to *unseen* conditions, which refers to conditions not represented in the training set. Data-driven SI-predictors have demonstrated performance improvements over state-of-the-art classical predictors in seen conditions, but not in unseen conditions.

Listening test paradigms are important to consider, when dealing with SI prediction. The paradigm of a listening test refers to factors other than the physical stimuli, such as different talkers, languages, vocabulary, sentence structure, lexical redundancy, test scoring methods, listening equipment and more that have an impact on the measured SI. The effects of a given paradigm can be approximated well by an s-shaped curve, which maps predictions of SI to absolute measured SI. This curve is called a psychometric function, a type of function that relates human responses on a test to some physical quantity, e.g., the SI experienced by the listener vs. signal to noise ratio of the stimulus. Psychometric functions for SI are typically modelled by a sigmoid function, where the parameters depend on the paradigm and the SI-predictor [1]. Note that a difference of slope between the psychometric functions of two listening tests implies that a similar change in a physical quantity, such as SNR, results in different changes to SI.

We use the term “pooling” to refer to constructing a dataset that contains speech stimuli and SI labels from multiple listening tests. However, naïve pooling of listening test data may be a questionable approach, because different listening tests have different underlying paradigms and psychometric functions associated with them. The speech stimuli alone do not completely account for the specific SI measurements of a listening test. For instance, the loudspeakers or headphones, used in two different listening tests, could make a difference in the subjective scores of the test subjects. Furthermore,

1. Introduction

some languages might be easier or harder to understand, under certain noise types. Similarly, coherent sentences allow some words to be inferred by context, which leads to higher SI scores than randomly constructed sentences, devoid of context, in the same listening conditions. These influences on the SI scores of different listening tests result in different parameters of the psychometric function.

Many studies of classical SI-predictors apply listening test specific mapping functions to convert the predictor output to absolute SI in performance tests, e.g., STOI [2], SIIB [3] and SII [4]. This is done in order to take the psychometric functions specific to each listening test into account, when evaluating predictor performance, and thus facilitate comparisons of predictions and performance across different listening tests. The predictions prior to these mappings are typically called *SI indices*, since they are, ideally, related monotonically to the subjectively measured SI, or *absolute SI*. SI indices can be meaningfully compared within the same paradigm, with a higher index corresponding to a higher absolute SI, but indices from different paradigms can not, since the psychometric function, and thus the map from SI index to absolute SI, changes with the paradigm.

When a data-driven SI-predictor is trained on a dataset of pooled listening tests, a fundamental problem arises. The input signals, used to train the SI-predictor, i.e., the speech stimuli, do not contain the complete information that determines the shape of the psychometric functions. With the information available in the training inputs and labels, the predictor can learn the specific psychometric functions underlying the training data, but it can not learn how to adapt to new unseen psychometric functions. This means that the predictor specializes in the paradigms underlying the training data.

We propose and investigate a method for training data-driven predictors, which allows the use of pooled listening test data from different paradigms, by taking the differences in psychometric functions in the training data into account. In particular, the method introduces trainable mapping functions with dataset dependent parameters. These mapping functions, which we call *Dataset-Specific Mapping Functions* (DSMF's), serve to model the psychometric functions specific to each individual listening test in the training data. We apply the training strategy to an SI-predictor¹ consisting of a Convolutional Neural Network (CNN) with a back-end inspired by ESTOI. This CNN is trained with pooled data consisting of speech datasets with SI-labels from different listening tests. The parameters of the trainable mapping functions are learned independently for each dataset. Their purpose is to approximate the psychometric function of each dataset, separately from the SI-predictor. After the training is complete, the trained DSMF's are discarded, because

¹The implementation of this SI-predictor can be found at https://github.com/Mapede/DSMF_SI_Predictor

the information they contain, namely an approximation of the psychometric functions of the training sets, is generally not useful, when predicting the SI of unseen datasets and paradigms. The trained SI index predictor is simply the remaining CNN-ESTOI network depicted in Figure C.1.

We show that training a data-driven SI-predictor with this strategy prevents it from learning an internal representation of the psychometric functions, inherent in the training data. It is demonstrated that this enables the proposed data-driven predictor to reach higher performance for seen conditions, and also to be more robust to new unseen test-paradigms. First, two SI-predictors are trained using the same architecture and pooled data, one using the proposed strategy, the other trained naïvely. This experiment shows that the proposed strategy leads to higher performance on average. Secondly, a series of hold-one-out cross validation experiments are conducted, where SI predictors are trained according to the proposed strategy, using all the available datasets except for one. The dataset, held out of training, is instead used for testing. In these experiments, the average performance of the trained predictors, on their respective unseen datasets, is higher than that of the classical predictors used for comparison.

The paper is organized as follows. Section 2 goes into detail on existing SI-predictors, both classical and data-driven. Section 3 describes the architecture of the proposed SI predictor and details of the proposed training procedure. Section 4 describes the datasets used to train and test the proposed SI predictor, as well as the training procedure and hyper parameters. Section 5 describes the experiments, and presents a performance evaluation of the proposed SI predictor. Finally, Section 6 contains the conclusions of the work.

2 Related work

SI-predictors may be roughly divided into classical, or data-driven methods. Classical SI-predictors, e.g., the Articulation Index (AI) [5], the Extended Speech Intelligibility Index (ESII) [6], the Speech-to-Reverberation Modulation energy Ratio (SRMR) [7], the Short-Time Objective Intelligibility (STOI) [2], the Spectro-Temporal Modulation Index (STMI) [8], the Extended Short-Time Objective Intelligibility (ESTOI) [9], the Speech Intelligibility In Bits (SIIB) [3] and the Hearing Aid Speech Perception Index (HASPI) [10], are hand-crafted models, often inspired by models of auditory perception, with only few parameters optimized for listening data. Data-driven SI-predictors, e.g., Non-Intrusive Speech Assessment (NISA) [11], a twin hidden Markov model [12], the data-driven STI estimator proposed by [13], the neural network proposed by [14], the convolutional neural network proposed by [15], and the convolutional neural network proposed by [16], learn a prediction

2. Related work

model primarily, or in full, by a process of optimization on a dataset of speech with labels of measured, or in some cases predicted, SI.

Another mode of classification for SI-predictors is, whether they are intrusive or non-intrusive. Intrusive predictors use both the clean reference signal and the noisy/processed test signal, whereas non-intrusive predictors only require the noisy/processed test signal. The advantages of intrusive SI predictors is that they are given more information than their non-intrusive counterparts, and can, in principle, reach a higher accuracy. The advantage of non-intrusive predictors is that they can be used when the clean reference is unavailable.

The AI [5] is perhaps the first classical method, and has served as inspiration for many following predictors. The AI performs a frequency weighted comparison of the long-term intensities of the underlying clean speech and the noise to estimate SI. The primary focus of the AI was speech in additive noise, and it was also designed for calculation by hand. The Speech Transmission Index (STI) [17] analyzes a set of probe signals passed through the transmission channel or processing algorithm of interest. In particular, the preservation of the probe signal modulations are measured, and used to quantify SI. Assuming that the channel is known, the STI supports non-additive distortions, such as clipping, filtering and reverberation.

The Speech Intelligibility Index (SII) [4] and Extended SII (ESII) [6] compute a weighted average of Signal to Noise Ratios (SNR) of specific frequency bands. The SII was proposed as an updated version of the AI, suitable for calculation by computer. In ESII, the SNR is computed in short time frame averages, rather than the long-term average used in SII. This improves its performance for speech signals in fluctuating noise [6]. The STMI [8] decomposes the signal under study into spectro-temporal components, and makes a comparison to the clean reference via cross correlation.

STOI [2] and Extended STOI (ESTOI) [9] use averages of sample correlations between the test signal and clean reference in short time segments in the 1/3 octave band magnitude domain. These sample correlations predict SI well when the time-frequency tiles are independent of each other. Since this is not generally the case, STOI and ESTOI normalize the signal segments before the sample correlations are computed. In STOI each segment is normalized across time, whereas in ESTOI they are also normalized across the 1/3 octave bands. This allows ESTOI to better handle temporally fluctuating noise, compared to STOI, [9]. SIIB [3] provides an estimate of SI via an estimate of the mutual information between the clean speech and noisy/processed speech. The idea of using mutual information to predict SI has been used earlier, see e.g., Speech Intelligibility using Mutual Information (SIMI) [18], the AI [19, 20] and Mutual Information Variational Bayes (MI-VB), MI K Nearest Neighbours (MI-KNN) and MI Expectation Maximization (MI-EM) [21].

HASPI [10] computes an intelligibility score based on an auditory model,

including both spectral envelope features and coherence. HASPI is also able to account for hearing impairment.

Data driven SI-predictors can be categorized by the type of labels used for training. The predictors proposed in [14], [15], [16] and [22], which are all different types of neural networks, are trained to estimate actual listening test results. Other data-driven SI-predictors are trained to emulate existing classical predictors in circumstances, where the classical predictor in question can not be used. In these cases, the labels are SI predictions produced by the classical predictor. For instance, the Non-Intrusive Speech Assessment (NISA) method [11] is trained to predict the outcome of STOI, without the clean reference that STOI normally requires. The important distinction is that NISA is trained using labels generated by STOI rather than a listening test. This circumvents the limitations imposed by the scarcity of listening test data, but also imposes the performance of STOI as an upper bound on the performance of NISA. Other examples include the predictors described in [13], a convolutional neural network emulating the STI, and [12] a hidden Markov model emulating STOI.

The data-driven methods proposed by [12], [14], and [16] are not evaluated on unseen conditions. The methods proposed by [11] and [13] are tested on unseen conditions, though these conditions are in the same category as the seen data, additive noise for [11], and reverberation from convolution with room impulse responses for [13]. Furthermore, these methods were trained using labels generated by classical predictors, STOI and STI for [11] and [13] respectively, rather than measured SI. Finally, the methods proposed in [15] and [22] were tested on unseen datasets, revealing highly dataset dependent performance.

3 Architecture and mapping functions

The data-driven SI-predictor proposed in this paper is a Convolutional Neural Network (CNN) with inspirations from ESTOI [9]. The architecture is shown in Figure C.1. The model takes two inputs: a potentially noisy and/or processed speech signal, $X[t, f]$, and the corresponding time-aligned clean speech signal, $S[t, f]$. In the training phase the model is also given a third input, the paradigm selector vector, d , which is a vector with a 1 in the entry corresponding to the listening test from which the training sample, i.e., $X[t, f]$ and $S[t, f]$, was drawn, out of a total set of D listening tests used for training. This vector is used to select the appropriate mapping function. Spectrograms $X[t, f]$ and $S[t, f]$ are 1/3 octave band representations of the time-domain speech signals $x[\tau]$ and $s[\tau]$, respectively. To obtain $X[t, f]$ and $S[t, f]$, both $x[\tau]$ and $s[\tau]$ are resampled to 20 kHz. Then a Short-Time Fourier Transform (STFT) is performed, followed by a 1/3 octave band transform, similar to

that of ESTOI, yielding $X[t, f]$ and $S[t, f]$. For the STFT a 50% overlapping Hann window, W samples in length, and zero padding to $2W$ samples is used. The input signals are processed in a number of CNN layers, followed by an ESTOI back-end, which performs the comparison between the signal under study and the clean reference. During network training, the output of the ESTOI back-end is mapped to absolute SI by the mapping function corresponding to the listening test from which the inputs and SI label were obtained.

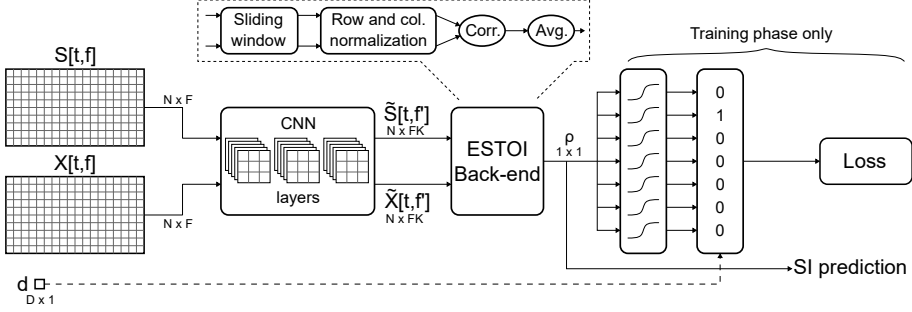


Fig. C.1: Proposed SI prediction architecture. From left to right are the 1/3 octave band inputs $S[t, f]$ of clean speech and $X[t, f]$ of noisy/processed speech, the dataset selection vector, d , used to choose the mapping function matching the listening test, the CNN layers which are applied to both $S[t, f]$ and $X[t, f]$ using the same kernels, yielding $\tilde{S}[t, f]$ and $\tilde{X}[t, f]$, the ESTOI back-end consisting of normalization and correlation performed on a sliding window, and an average across frames resulting in the SI index prediction, ρ . In the training phase, the SI index, ρ , i.e., the output of the ESTOI back-end, is mapped to a prediction of absolute SI using the logistic function indicated by d .

3.1 Network design

The goal in designing the network is to increase robustness to unseen datasets. The architecture is designed to be relatively small, in order to mitigate overfitting to the seen datasets. The proposed architecture has fewer than 10^4 trainable parameters, whereas network sizes used in [22] range from 10^5 to 10^6 parameters. These large models showed signs of overfitting, as the performance was drastically lower for certain unseen datasets. This is also the reason why we have chosen to incorporate part of ESTOI into the network, i.e., to reduce the required number of trainable parameters.

We choose ESTOI, specifically, because of its simplicity and performance. The ESTOI back-end provides an anchor point of performance, in that the network should be able to perform at least as well as ESTOI on the training set. Hence, with this network design we expect performance on par with or better than ESTOI for seen conditions. The trainable part of the network, i.e., the CNN layers, is placed before the ESTOI normalization for a number of

reasons. First, it guarantees that when the studied signal is in fact clean, i.e., $x[\tau] = s[\tau]$, the predicted SI is maximized. This is due to the fact that $x[\tau]$ and $s[\tau]$ are subject to the exact same mathematical operations, because they share the same CNN layers.

A CNN architecture was chosen, because it allows processing of variable lengths of input signals [15], and because CNN's have proven efficient for speech processing tasks in general, see e.g., [15], [22], [23], [24] or [25]. In a preliminary experiment we tested other architectures, particularly including a trainable weighted averaging across frequency bands in the back-end. This weighted average was found to have no significant impact on performance. The proposed training procedure is not limited to the architecture described here. It can be applied to the training of any data driven SI-predictor.

3.2 1/3 Octave band transform

The 1/3 octave band transform is applied as presented in [2]. First, the STFT, given by:

$$\hat{X}[t, k] = \frac{1}{\sqrt{2\pi}} \sum_{s=0}^{W-1} x \left[\frac{tW}{2} + s \right] w[s] e^{-jks}, \quad (\text{C.1})$$

is applied, where $\hat{X}[t, k]$ is the STFT of x at time t , and frequency k , $w[\cdot]$ is a Hann analysis window of length W , and j denotes the imaginary unit. Then, the magnitudes of each 1/3 octave band are computed as follows:

$$X[t, f] = \sqrt{\sum_{s=k_l[f]}^{k_h[f]} |\hat{X}[t, s]|^2}, \quad (\text{C.2})$$

where $X[t, f]$ is the 1/3 octave band representation of x at time t , and 1/3 octave band f , and where $k_l[f]$ and $k_h[f]$ are the indices of the lowest and highest frequency bands of \hat{X} within the f 'th 1/3 octave band. Similar operations are applied to $s[\tau]$ to obtain $S[t, f]$. For more details we refer to [2].

3.3 CNN layers

The 1/3 octave band transformed signals, $X[t, f]$ and $S[t, f]$, are now run independently through the same CNN layers, cf. Figure C.1. We use L CNN layers of K kernels with Rectified Linear Unit (ReLU) activation functions. The signals are zero padded to preserve their size after each convolution.

3.4 ESTOI back-end

The CNN layers produce K outputs, $\tilde{X}[t, f, 0], \dots, \tilde{X}[t, f, K-1]$, each corresponding to one kernel in the final layer. These K outputs are concatenated

3. Architecture and mapping functions

along the frequency-axis:

$$\tilde{X}[t, f'] = [\tilde{X}[t, f, 0] \dots \tilde{X}[t, f, K-1]], \quad (\text{C.3})$$

where f' is used to index the new concatenated frequency axis. This concatenation results in a computationally convenient representation of \tilde{X} for the next step. Following the CNN layers are a series of operations from ESTOI, as illustrated in the details of the "ESTOI Back-end" in Figure C.1 [9]. A sliding rectangular window, N samples wide, is applied along the temporal axis, splitting the input spectrograms into short overlapping matrices. The n 'th of these matrices is given by:

$$\tilde{X}_n[t, f'] = [\tilde{X}[n, f']^\top \dots \tilde{X}[n+N-1, f']^\top]^\top. \quad (\text{C.4})$$

For each n , $\tilde{X}_n[t, f']$ is normalized across time and frequency as follows. First, the mean is subtracted across time:

$$\tilde{X}_{n,2}[t, f'] = \tilde{X}_n[t, f'] - \frac{1}{N} \sum_{s=0}^{N-1} \tilde{X}_n[s, f']. \quad (\text{C.5})$$

Then, the variance is normalized across time:

$$\tilde{X}_{n,3}[t, f'] = \tilde{X}_{n,2}[t, f'] / \sqrt{\sum_{s=0}^{N-1} \tilde{X}_{n,2}^2[s, f']}. \quad (\text{C.6})$$

Now, the mean across frequency is subtracted:

$$\tilde{X}_{n,4}[t, f'] = \tilde{X}_{n,3}[t, f'] - \frac{1}{N} \sum_{s=0}^{F-1} \tilde{X}_{n,3}[t, s]. \quad (\text{C.7})$$

Finally, the variance is normalized across frequency:

$$\tilde{X}_{n,5}[t, f'] = \tilde{X}_{n,4}[t, f'] / \sqrt{\sum_{s=0}^{F-1} \tilde{X}_{n,4}^2[t, s]}. \quad (\text{C.8})$$

$\tilde{S}_{n,5}[t, f']$ is computed similarly. The correlation coefficient between each corresponding matrix of the noisy/processed and clean speech signals is now given by:

$$\rho_n = \frac{1}{N} \sum_{t=0}^{N-1} \sum_{f=0}^{F-1} \tilde{X}_{n,5}[t, f'] \tilde{S}_{n,5}[t, f']. \quad (\text{C.9})$$

The average across frames, ρ , of these correlation coefficients is the output of the network.

3.5 Dataset-specific mapping functions

Ideally, the trained SI index predictor should be independent of the paradigms specific to the listening tests included in the training data. To achieve this, we append a number of DSMF's to the architecture used exclusively for the network training and validation phases. This is marked as "Training phase only" in Figure C.1. During network training, an additional input is given. This input, d , is a vector with a 1 in the entry corresponding to the index of the dataset from which the inputs $s[t]$ and $x[t]$ originate, and 0's in all other entries. The DSMF's used in this study are logistic functions, defined as:

$$\sigma(x) = \frac{1}{1 + e^{-(ax+b)}}, \quad (\text{C.10})$$

where x is the input, and a and b are the trainable parameters. Conveniently, computing these functions corresponds to a single fully connected layer, with a number of nodes equal to the number of listening tests, followed by a sigmoid activation function. The parameters a and b then, respectively, correspond to the weights and biases of the fully connected layer. This layer is designed to apply all the DSMF's to the network output in parallel during training, which is represented by the block filled with s-shaped curves in Figure C.1. The inner product is now taken between the outputs of the fully connected layer and the selector vector d , in order to select the relevant DSMF. Thus, only the DSMF corresponding to the dataset indicated by d is passed through this operation. This particular implementation was chosen because it is differentiable, which allows for back propagation. In this way, the network can be simultaneously trained on multiple pooled datasets, while the mapping functions absorb the different psychometric functions, which the network could otherwise only account for by over-fitting. The choice to use logistic functions as DSMF's is inspired by the fact that logistic functions are often used to model psychometric functions for classical SI-predictors, see e.g., STOI [2], ESTOI [9], SIIB [3], CNN [15], SII [4] or the survey of psychometric functions for SI in [1]. Importantly, because of the choice to train with logistic DSMF's it can be expected that the network outputs SI-indices that are logistically related to absolute SI.

The DSMF training procedure is designed to give the network a parameter efficient way to represent the psychometric functions that arise from the training data. The psychometric functions are thus learned separately from the CNN, which means that the internal parameters of the network can be utilized more efficiently, leading to better SI-prediction performance even though the DSMF's themselves are discarded in the end.

The result of the proposed DSMF training procedure is a network that outputs an unmapped SI-index, ρ , which correlates highly with absolute SI. In practice, for unseen data, ρ would be used as the SI-prediction. In general, SI-indices produced by this network are not predictions of absolute SI,

4. Dataset description and training procedure

Table C.1: Overview of the datasets used for training and testing of the proposed SI predictor. The datasets have been split into files of equal duration of approximately 6.6s of speech. The column labelled #subj. list the number of participating listeners, and the column labelled #cond. lists the number of different listening conditions resulting from the various noise types and SNR’s as well as processing types and settings.

Dataset		Size			Content	
No.	Ref.	#subj.	#files	#cond.	Speech material	Noise & processing types
DS0	[26, Sec. VI]	11	278	9	Dantale II (closed)	BBL, Beamforming
DS1	[27, Sec. III-C]	14	241	20	Dantale II (closed)	BFN, ITFS
DS2	[9, Sec. IV-1]	12	684	60	Dantale II (closed)	Noisex, SSN, BBL, Temporal modulation
DS3	[28, Sec. II]	15	7808	168	Dantale II (open)	SSN, BBL, café, car, ITFS
DS4	[29, Sec. IV]	9	64	52	Dutch Hagerman test (closed)	SSN, pre-noise enhancement
DS5	[2, Sec. III-C]	7	390	35	IEEE database (open)	SSN, BBL, ITFS w. artificial errors
DS6	[15, Sec. III-D ₄]	8	2139	327	ADD (open)	SSN, Low- and high-pass filtering
DS7	[30, Sec. III]	15	976	24	CLUE database (open)	ICRA, Speech segregation
DS8	[31, Sec. III-B]	16	547	18	Dutch Hagerman test (closed)	SSN, BBL, pre-noise filtering
DS9	[9, Sec. IV-5]	13	4333	20	Dutch Hagerman test (closed)	SSN, Single channel noise reduction

because the network does not account for psychometric functions. In special cases, however, where the listening test paradigm is known, i.e., when the data comes from a known listening test, the corresponding DSMF could be appended to produce predictions of absolute SI. In the interest of facilitating a fair comparison with competing predictors, however, we will not be using the trained DSMF’s in the test phase.

4 Dataset description and training procedure

4.1 Datasets

The experiments described in this paper are based on a pooled dataset consisting of the results from ten listening tests. Table C.1 describes the datasets with a few keywords pertaining to the speech material, noise types and processing in each listening test. The noisy/processed speech stimuli, $x[t]$, and the clean reference signals, $s[t]$, from each noise/processing condition in each listening test were extracted. The label for each pair of signals was taken to be the average fraction of correct words across all listeners within the given condition. It would have been desirable to use more granular SI-labels, e.g., binary labels indicating whether each individual word was correctly identified in the corresponding listening test. However, for the vast majority of the datasets we use, particularly DS3 through DS9, only the average SI is available. For the sake of consistency, we use the average SI labels for all datasets.

All ten listening tests were conducted with normal hearing native speakers. The listening tests were either conducted with a closed set, which allowed participants to select each word from a list, or an open set, which required the participants to either write down or repeat each word without a

list of candidate words. For more detailed descriptions of the datasets and listening tests, we refer to the respective sources listed in Table C.1. In Table C.1, Dantale II refers to the Danish matrix test speech corpus described in [32]. ADD refers to Akustiske Databaser for Dansk², which contains meaningful Danish sentences. CLUE refers to the Danish speech corpus described in [33]. The Dutch Hagerman matrix test speech corpus is described in [34], and the IEEE database contains English speech. These speech datasets each contain speech signals from a single talker, apart from ADD, which contains speech signals from multiple talkers. The Noisex database is described in [35], and contains various recorded noise types. Speech shaped noise (SSN) refers to white Gaussian noise, filtered to match the long term spectral envelope of speech. Babble (BBL) noise refers to the mixture of a number of competing talkers. The number of competing talkers varies from 2 to 20 depending on the dataset. Bottle factory noise (BFN) refers to recorded noise of bottles clinking against each other on a conveyor belt. ICRA is a database of noise signals, constructed to mimic the short term modulations of speech [36]. Ideal time-frequency segregation (ITFS) is a method for enhancing a signal in the time-frequency domain by utilizing the true signal to noise ratio (SNR) for each time-frequency tile, in order to, for example, compute ideal gains or cut-off thresholds [37]. The signals in DS6 have been recreated using a different speech database than the one used in the original listening test, the full details and verification experiments can be found in [15, Sec. III].

4.2 Training

Each dataset was split randomly into 80% training, 10% validation and 10% test data. This was done to ensure that all datasets would be represented in the test set. The data was partitioned into training samples of equal duration, to enable the construction of mini batches. The duration of 512 frames, corresponding to approximately 6.6 seconds, which is long enough to accommodate one to two sentences, was chosen. This fixed duration resulted in some training samples spanning two listening test conditions. The labels for these samples were computed as the weighted average of the measured SI for those two conditions, with weights equal to the number of frames from each condition in that training sample. A batch size of 32 was found to give the best compromise between GPU-memory, training speed and end-performance. The network was trained on batches from the training dataset using the Adam optimizer [38], and the Mean Squared Error (MSE) loss function. An early stopping scheme was used, where the learning rate was halved for every 25 epochs without a new global minimum in validation cost, and the training was stopped early if this continued for 35 epochs. Training was allowed to proceed for a maximum of 300 epochs. Training of the models

²http://www.nb.no/sbfil/dok/nst_taledat_dk.pdf

5. Performance evaluation

with DSMF involves forward passing training samples, i.e., triplets of $X[t, f]$, $S[t, f]$ and d , through the CNN layers, the ESTOI back-end and finally the DSMF's, after which the loss function is evaluated. For the test phase, the trained DSMF's are discarded. To take the psychometric functions into account for the evaluation, logistic functions are fitted for each listening test in the test data by least squares for all the evaluated predictors. This is done in order to facilitate a fair comparison between the DSMF trained networks and the classical predictors. This also allows the DSMF trained networks to be tested on unseen datasets. The architecture was implemented using Tensorflow 2.1 [39].

4.3 Network parameters

We trained the networks with the following parameters. The window length of the 1/3 octave band transform is $W = 512$. A preliminary experiment showed that $L = 3$ CNN layers with $K = 20$, 3×3 kernels resulted in the best performance. Networks with 1, 2, 3 and 4 CNN layers and 5, 10, 15 and 20 kernels per layer were tested. Due to memory constraints, we were unable to test higher numbers of kernels. The window length of the ESTOI back-end is $N = 30$, cf. [9]. The lowest 1/3 octave band is centred around 150 Hz, and the highest around 6050 Hz, for a total of $F = 17$ bands. This is an increase from the conventional ESTOI, which uses 15 bands. According to the band importance function of the SII, [4], this frequency range accounts for most of the intelligibility of speech. In total the architecture has 7,460 trainable parameters.

5 Performance evaluation

Two experiments, A and B, are performed to investigate the properties of the proposed DSMF training strategy and the resulting SI predictor. In Experiment A the goal is to validate that the DSMF's absorb the information related to the different psychometric functions, and result in improved prediction performance over plain pooling with no DSMF's. In Experiment B the goal is to investigate the robustness of the network and training method to new or unseen listening conditions and test paradigms. The Spearman and Pearson correlation coefficients, along with the Mean Squared Error (MSE) values, are used as evaluation metrics.

In order to evaluate the efficacy of the DSMF training procedure, models were trained both with and without DSMF. Both models have the same number of parameters in the CNN layers, but since the DSMF's should be able to represent the psychometric functions of each dataset, we expect the DSMF trained model to utilize these parameters more efficiently. As a result, the

DSMF trained model is expected to reach higher performance than the one without DSMF. These models are both tested against each other, and against a variety of classical predictors, i.e., ESTOI, SIIB, HASPI, STOI and SI-SDR. We remind the reader that the trained DSMF's are not used in the test phase. Instead, as part of the evaluation of the performance of each predictor on the test data, logistic psychometric functions are fitted to the test data using least squares, and used to transform the outputs of the predictors to absolute SI, before computing the Spearman and Pearson correlations as well as the MSE values. These logistic functions should not be confused with the trained DSMF's, and we stress that they are solely used as part of the evaluation of the SI-predictors, facilitating the comparison between predictions and measured absolute SI. This has no impact on the Spearman coefficient, because it is invariant under monotonically increasing transforms, i.e., the fitted logistic functions. It does affect the Pearson correlation and MSE, however, since the logistic fitting attempts to map predictions onto a straight line, which should increase the Pearson correlation, and reduce the MSE. This facilitates a fair comparison between the trained and classical predictors, and better reflects the performance that can be expected in practice. Specifically, if - hypothetically - the trained DSMF's were used in the test phase, the proposed network might have an advantage specific to the datasets used in this work, but this advantage would not generalize, since trained DSMF's only exist for seen datasets.

Table C.2 shows the Spearman correlations, as computed dataset-wise, for each predictor. The predictors trained in this experiment are marked with (seen) in Table C.2. The prediction for each listening test condition was made by concatenating all the speech signals available in the test set for that particular condition, resulting in one pair of inputs for each condition. These pairs were given to the predictors as inputs yielding one scalar SI-prediction per condition as output. The correlations between the predictions of all conditions within each dataset and the corresponding measured SI from the listening tests were then computed. The performance in terms of Pearson correlations is computed in a similar way and seen in Table C.3. Additionally, the mean squared error of each predictor is reported in Table C.4. We noticed no loss in performance as a consequence of the relatively longer test signals. This is likely because the architecture has a very small receptive field because of the small kernels in the CNN layers.

As expected, the DSMF trained network reaches a higher performance than the non-DSMF trained network in terms of both Spearman and Pearson correlation. Since the only difference between these two networks is the presence of DSMF's during training, it is clear that training with DSMF's has a positive effect on the final performance of the SI-predictor, indicating that the DSMF's are working as intended. In particular, the performance average across datasets is higher with DSMF. The exceptions, where the non-DSMF

model performs better are DS2 and DS7. A possible explanation for this could be that networks sacrifice performance for some datasets in order to perform better on average. DS2 is a difficult dataset for many SI-predictors, because it contains temporally modulated noise [9]. Note also that STOI performs poorly, whereas ESTOI does particularly well on this dataset. This is an expected result, as ESTOI was proposed in order to improve STOI’s performance on speech in temporally modulated noise, and evaluated using DS2 [9]. Figures C.2 and C.3 show scatter plots of measured SI and predictions, each point representing one listening test condition. Figure C.2 shows the raw predictions, or SI indices, i.e., before logistic functions are fitted, vs. measured absolute SI for the various SI predictors. From Figure C.2, the proposed SI predictor, ESTOI and SIIB manage to produce fairly concentrated clusters of predictions, whereas STOI and HASPI struggle to do so. This is also reflected in Tables C.2 and C.3. For DS2 specifically, the predictions show a much wider spread at the high end of the SI-spectrum. This is consistent with the observations in [9] that many SI-predictors tend to underestimate the SI in this dataset. In DS7 there are very few conditions at the extreme ends of the measured SI spectrum, i.e., 0 and 1, where prediction errors are generally smaller. This could explain why many of the SI predictors score relatively low on this dataset. In Figure C.2 it can be seen that the network trained without DSMF produces indices with an approximately linear relation to absolute SI, whereas the network trained with DSMF produces indices with separate, approximately logistic relations to absolute SI. This clearly illustrates the difference between training with and without DSMF; the non-DSMF trained network must necessarily be dedicating internal parameters to recognizing and mapping each of the datasets to absolute SI, i.e., the network has specialized to the training data. Recall that the psychometric functions cannot generally be determined from the network inputs alone. The network trained with DSMF, however, does not appear to have any internal representation of the psychometric functions of the datasets, since each dataset forms a separate s-shaped cluster, indicating that the DSMF’s were able to absorb the different psychometric functions of the training data.

5.1 Experiment A

Among the classical predictors, ESTOI and SIIB have the best performance, which is in accordance with existing studies, see e.g., SIIB [3] or ESTOI [9]. While the classical predictors are not primarily data-driven, some of the datasets we test on, were used in the development of the classical predictors. Specifically DS3, DS5 and DS9 were used in the development of STOI, DS2, DS3 and DS9 in the development of ESTOI [9], and DS3, DS4 and DS9 in the development of SIIB [3]. This is reflected in the performance of these predictors on those respective datasets, as seen in Tables C.2, C.3 and C.4,

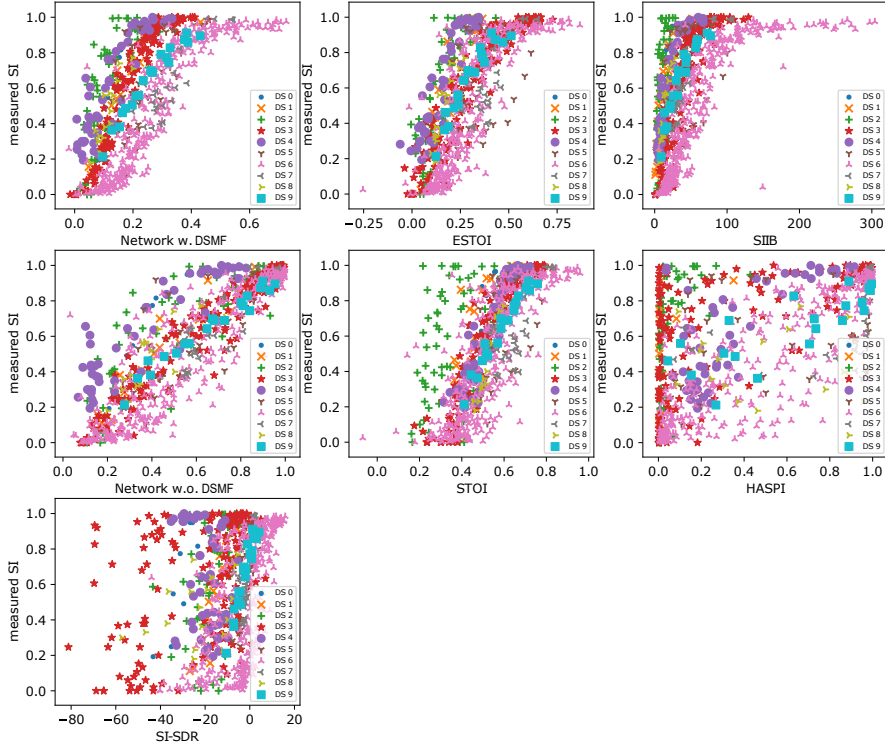


Fig. C.2: Experiment A: Absolute measured SI vs. raw SI indices output by the networks and classical predictors, i.e., with no fitted logistic functions.

where e.g., STOI reaches a Spearman correlation of 0.54 on DS5. These observations are well in line with conclusions drawn in [40] that SI-predictors tend to perform better on datasets used during their development. HASPI and SI-SDR show the lowest performance on average. SI-SDR shows drastic variation in performance from one dataset to the next, with high performance on DS1, DS5, DS7 and DS9, and low performance on DS0, DS2, DS3, DS4 and DS8. Note in particular the negative Spearman coefficients on DS0. This negative correlation could be due to the relatively few conditions in DS0, which means that fewer discordant pairs are necessary to significantly reduce the score. Note that high correlations with different signs may be detrimental to any SI-predictor: In order to be reliable in practice, it must be clear whether an increase in predictor output is indicative of an improvement or a decline in SI.

In the case of HASPI this can be attributed to slightly lower scores on most of the datasets and very low scores on DS0 and DS8 in particular. HASPI's very low score on DS8, might also be explained by the fact that DS8 has few

5. Performance evaluation

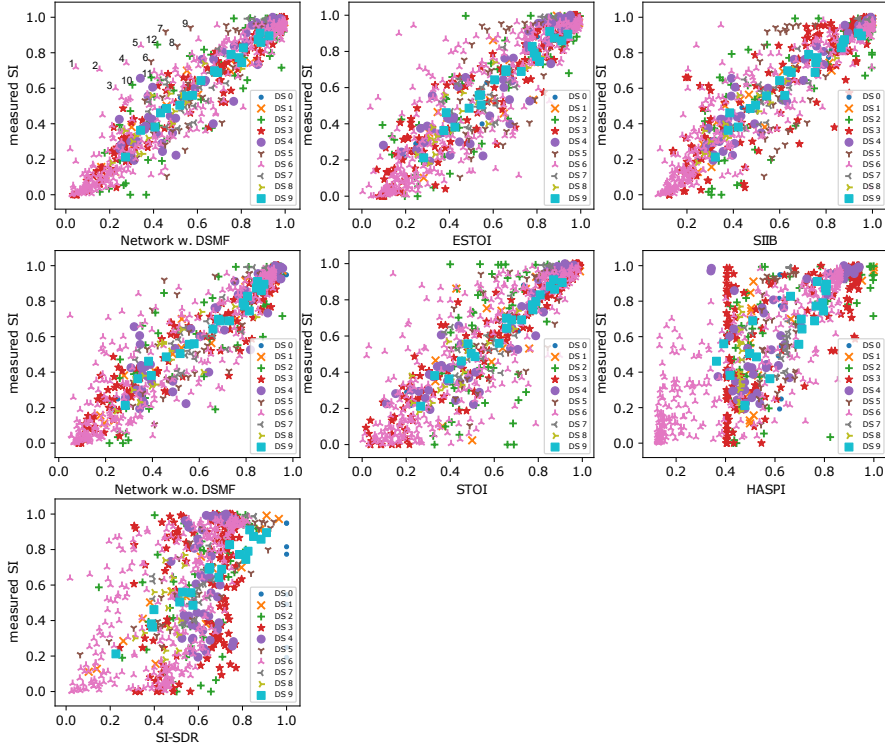


Fig. C.3: Experiment A: Absolute measured SI vs. SI-predictor output transformed by logistic functions fitted to each of the test datasets and predictors for both networks and classical predictors.

conditions.

Figure C.2 demonstrates the difference between training with and without DSMF's. In particular, the network trained without DSMF's attempts to force predictions from all the datasets onto the same line between (0,0) and (1,1). This is a clear indication that the network has learned an internal representation of the psychometric functions specific to the training datasets. As a consequence, the predictions show a substantial variance. When trained with DSMF's, however, the outputs related to different datasets form separate s-shaped clusters. The differences between these clusters are a result of the paradigm differences, meaning that the network has not learned an internal representation of the psychometric functions of the training data. It is evident that this has resulted in substantially reduced variance in the predictions. Note that the clusters corresponding to each dataset, appear similar for this DSMF trained network and for ESTOI, which could be a result of the similarities between the proposed architecture and ESTOI. These similarities

are further evidence that the different clusters represent different psychometric functions, since ESTOI does produce SI indices that map to absolute SI via a logistic psychometric function [9]. Given the similarities between the proposed architecture and ESTOI, it is not surprising to see similarities in their psychometric functions as well.

Figure C.3 shows the same results, but each dataset has now been mapped to absolute SI using logistic psychometric functions fitted by least squares. Note that these logistic functions are not the trained DSMF's. These logistic functions are fitted to the test data, as opposed to the DSMF's that are fitted to the training data. The DSMF's are also trained jointly with the network, whereas these logistic functions are only fitted after the network has been trained. Thus, the predictions now ideally cluster around the diagonal line from $(0,0)$ to $(1,1)$. From this figure it is easier to compare the performance across the different predictors because the predictions can now be considered absolute SI predictions, rather than SI indices. For instance, the DSMF network appears to be better at predicting low intelligibility than the non-DSMF network, as the clustering is tighter near $(0,0)$. This could be because DSMF allows training to focus on tightening each dataset cluster, tighter clusters being equivalent to higher precision in predictions, rather than spending degrees of freedom on bringing all the clusters together. In other words, the DSMF network learns to predict SI indices for each dataset, rather than absolute SI, and reaches better performance, because this task is simpler. As a result, the network trained with DSMF reaches the highest performance among the tested predictors.

The listening conditions associated with three sets of predictions with notable errors are listed in Table C.5. The conditions are labelled 1 - 12 in Figure C.3. These predictions come from datasets DS2, DS5 and DS6. In the case of DS2 there are three listening conditions, all with the noise type Sinusoidal Noise Amplitude Modulation (SNAM) at various modulation frequencies and low SNR. A possible explanation for why the network struggles with this noise type could be that it is similar to the stationary noise type SSN, which appears very frequently in the training set. However, speech in SNAM may be significantly more intelligible than speech in SSN [9], because the modulated noise allows the listener to "listen in the dips", see e.g., [41] for more details. Looking at the points from DS5, they all come from the same processing scheme involving Ideal Binary Masked (IBM) speech. In particular, this listening test investigated the effect of artificial errors in an IBM speech enhancement system. In this context the Type I error listening condition, cf. Table C.5, refers to IBM's where spectro-temporal gains of zero were converted to one, i.e., the enhancement system preserves too much of the noise. It is possible that the network overestimates the impact on SI of this extra noise, especially considering that this noise only appears in spectro-temporal regions which were noise dominated in the first place. For DS6 there does not

appear to be any pattern in the listening conditions. The errors here could be due to the fact that the stimuli in this dataset were recreated using a different speech corpus from the original listening test [15].

5.2 Experiment B

In general, SI predictor networks should ideally be applicable to other types of listening conditions than the ones used during the training phase. The generalizability of the network proposed in this study is tested in a cross-validation experiment. In this experiment we train the network with ten different initializations on ten different partitions of training, validation and test data, i.e., one hundred networks trained in total. More precisely, we move the training and validation data from one listening test at a time entirely to the test set. This means that each listening test is excluded from the training and validation phases of ten models, and that the dataset is unseen when testing those models. For each partition, the model with the lowest validation loss was selected for the test phase. As such, this experiment gives an indication as to how the networks will perform in unseen conditions, and how they react to unseen listening conditions and test paradigms. As in Experiment A, we expect that the DSMF trained models will reach higher performance than the non-DSMF trained models. This is because the non-DSMF trained models learn an internal representation of the psychometric functions related to the training datasets. Since the psychometric functions related to unseen test data may be completely different from those related to the seen training data, such internal representations are undesirable.

Tables C.2, C.3 and C.4 contain Spearman correlations, Pearson correlations and mean squared errors for the models. For any given dataset, the correlations and MSE values in the rows marked as Net (unseen) are computed for predictions made by a model with that dataset excluded from the training and validation sets. This means that each column describes a separate instance of the model, trained without access to the dataset corresponding to that column.

As expected, the performance for most of the datasets is lower when the dataset is unseen. The models experience the largest drops in performance on DS3, DS5 and DS6 as compared to when the datasets are seen. The reason for this could be that DS3 and DS6 are the largest and most diverse in terms of listening conditions. The exclusion of any of these datasets is a large reduction in the total amount of training data, which could result in the relatively larger loss of performance. Furthermore, large test sets also make it harder to produce a good ranking of a larger number of diverse conditions, as there are more opportunities for mistakes. As for DS5, judging by the relatively low scores, which the classical predictors achieve, it appears to be one of the hardest of these datasets for SI predictors in general. Despite the

Paper C.

Table C.2: Spearman correlations between the mapped predictions and the measured SI of the 10 datasets. In the (unseen) rows, each column represents a different permutation of training, validation and test data, where the corresponding dataset has been excluded from training and validation. The rightmost column shows the average of the Spearman coefficients across the datasets.

Spearman $\times 100$											
Predictor	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	Average
Net w. DSMF (seen)	97.62	97.94	86.37	97.51	89.66	66.78	93.46	68.43	88.85	98.61	88.52
Net w.o. DSMF (seen)	95.24	89.47	89.26	97.06	85.65	61.50	92.18	71.13	85.35	97.56	86.43
Net w. DSMF (unseen)	97.62	92.78	80.19	89.39	88.82	64.00	90.02	60.52	97.52	95.64	85.65
Net w.o. DSMF (unseen)	97.62	92.36	70.77	82.56	85.67	51.11	93.88	58.52	90.51	97.74	82.07
STOI	97.62	94.22	36.29	94.70	88.70	54.16	81.47	61.57	91.95	98.31	79.90
ESTOI	97.62	97.73	85.50	95.07	89.39	41.21	87.66	62.09	87.20	96.73	84.02
SIIB	100.00	96.70	79.28	91.05	93.78	35.52	92.33	76.43	92.78	97.93	85.58
HASPI	-2.38	63.88	68.13	68.79	62.02	38.03	85.65	65.91	2.37	82.21	53.46
SI-SDR	-83.83	91.74	54.83	43.73	29.28	84.75	66.43	66.43	33.54	95.75	48.27

Table C.3: Pearson correlations between the mapped predictions and the measured SI of the 10 datasets. In the (unseen) rows, each column represents a different permutation of training, validation and test data, where the corresponding dataset has been excluded from training and validation. The rightmost column shows the average of the Pearson coefficients across the datasets. The values marked with * are not significantly different compared to the best predictor on the given dataset.

Pearson $\times 100$											
Predictor	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	Average
Net w. DSMF (Seen)	99.68*	97.95*	87.10*	98.07*	93.67*	64.04	93.20*	84.06*	91.85	99.00*	90.86
Net w.o. DSMF (Seen)	99.21*	89.76	88.95*	96.65	91.15	60.56	91.93	82.84*	83.55	98.53*	88.31
Net w. DSMF (Unseen)	99.63*	93.49	78.32*	89.18	92.09	62.07	88.63	75.93	96.95*	96.57*	87.28
Net w.o. DSMF (unseen)	98.93*	93.30	69.01	82.82	90.48	47.36	90.83	55.58	92.07*	98.21*	81.86
STOI	99.20*	92.11	34.16	93.99	92.62	49.64	80.47	76.69	92.34*	99.01*	81.02
ESTOI	99.54*	97.79*	84.78*	94.61	90.77	36.05	86.47	76.34	83.73	97.52*	84.76
SIIB	99.18*	94.46	79.23*	89.94	95.86*	29.37	91.44	91.79*	94.91*	96.95*	86.31
HASPI	1.39	60.17	61.67	67.19	69.75	22.45	84.57	32.40	1.67	78.87	48.01
SI-SDR	-32.07	92.59	46.65	41.67	25.79	83.34*	60.37	73.14	34.48	97.15*	52.31

Table C.4: Mean squared error between the mapped predictions and the measured SI of the 10 datasets. In the (unseen) rows, each column represents a different permutation of training, validation and test data, where the corresponding dataset has been excluded from training and validation. The rightmost column shows the average mean squared error across the datasets. Note that all mean squared errors in this table have been scaled by a factor of 100 for better formatting.

Mean squared error $\times 100$											
Predictor	DS0	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9	Average
Net w. DSMF (Seen)	0.052	0.335	2.355	0.420	0.976	3.868	1.549	1.135	0.496	0.075	1.126
Net w.o. DSMF (Seen)	0.127	1.587	2.023	0.729	1.334	4.146	1.811	1.209	0.959	0.109	1.404
Net w. DSMF (Unseen)	0.060	1.028	3.218	2.247	1.204	4.026	2.509	1.634	0.191	0.253	1.637
Net w.o. DSMF (unseen)	0.171	1.060	4.365	3.438	1.439	5.081	2.059	2.631	0.483	0.133	2.086
STOI	0.127	1.238	6.430	0.827	1.143	4.361	1.673	1.539	0.182	0.088	1.761
ESTOI	0.071	0.361	0.987	1.071	1.401	5.370	1.959	1.663	0.566	0.162	1.361
SIIB	0.135	0.899	2.007	2.145	0.619	5.707	0.797	0.488	0.235	0.252	1.328
HASPI	7.755	5.212	6.046	5.995	4.072	6.217	3.334	3.399	3.168	1.416	4.661
SI-SDR	0.460	1.173	5.867	9.000	7.332	2.009	7.190	1.793	2.807	0.208	3.784

5. Performance evaluation

Table C.5: Marked points from the scatterplot "Network w. DSMF" in Figure C.3.

Point	Dataset	Condition
1	DS6	High-pass 1122 Hz, -8 dB SSN
2	DS6	Low-pass 3458 Hz, 2 dB SSN
3	DS6	High-pass 1122 Hz, -2 dB SSN
4	DS6	Low-pass 2239 Hz, -2 dB SSN
5	DS6	High-pass 178 Hz, 0 dB SSN
6	DS5	Type-I 20 talker, error rate 0.8
7	DS5	Type-I 20 talker, error rate 0.6
8	DS5	Type-I 20 talker, error rate 0.7
9	DS5	Type-I 20 talker, error rate 0.4
10	DS2	-27 dB SNAM 2 Hz
11	DS2	-19 dB SNAM 8 Hz
12	DS2	-21 dB SNAM 4 Hz

performance drop when this dataset is left out of training, the given model achieves higher performance than the classical predictors. Exceptionally, DS0, DS1 and DS8 have higher scores on the unseen models compared to the seen. These datasets all consist of few listening conditions, 20 or fewer. The explanation could be similar to the one for the large datasets, i.e., that the models simply perform better in general, when the training set is larger. Removing a small dataset from the training set, would then have only a small impact on performance.

Williams' t-test [42] was used to test for significant differences between the SI-predictors. This is a pairwise hypothesis test designed to detect significant differences in Pearson correlations. The null-hypothesis is that two different predictors have the same Pearson correlation with measured SI. Following the same procedure used in [9], we tested the highest performing predictor on each dataset against the others, and marked those not significantly different with * in Table C.3. A significance level of $\alpha = 0.05$ with Bonferroni correction, to account for multiple tests, was used. Note that DS0, DS1, DS8 and DS9 contain 20 or fewer datapoints, i.e., listening conditions, which means that the t-tests could be unreliable on these datasets, according to [42].

On average, the unseen models score slightly higher than the classical predictors, which suggests that the proposed architecture and training scheme generalizes well and produces predictors which perform on par with, or better than the existing classical predictors for listening conditions, on which it has not been trained. We attribute this robust performance to two main factors. First, the proposed network contains as few as 7,460 trainable parameters, which mitigates overfitting. Secondly, the use of DSMF during training

facilitates pooling of training data obtained from different listening tests, effectively increasing the amount of listening test data available for training.

Performance of the proposed SI predictors, when tested with signals from listening conditions similar to those used for training the SI-predictor, is substantially better than existing methods. This improved prediction performance may be advantageous for replacing some listening tests in iterative development of speech processing systems. Assuming that the processing scheme, or the stimuli, are not changed too drastically, then the SI-predictor network can be validated or even retrained in order to benefit from the high performance on seen conditions.

Looking at DS5 in Table C.2, there is a larger gap in performance between the (seen) and (unseen) models without DSMF's compared to the models with DSMF's. In particular, the difference in Spearman correlation is 0.0278, for the DSMF trained model and 0.1039, for the non-DSMF trained model. Noting that DS5 is the only dataset which contains English speech, this might be interpreted as the DSMF training successfully increasing the model's robustness to an unseen language. It should be noted, however, that language is not the only paradigm difference in DS5, so the drop in performance of the model without DSMF's might not only be due to the unseen language.

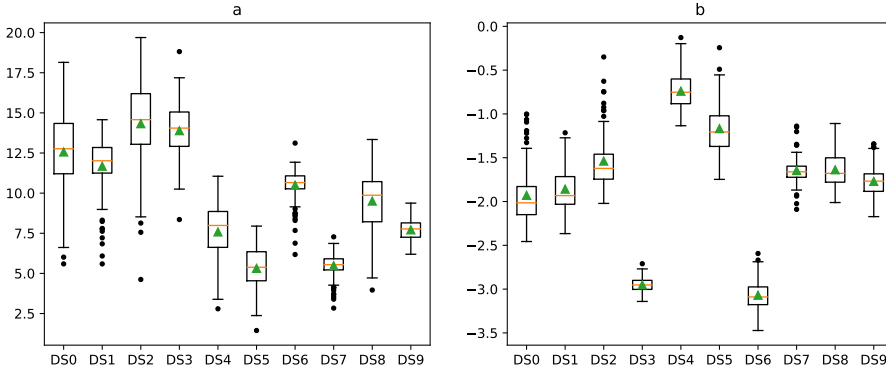


Fig. C.4: Boxplots of the trained DSMF parameters a and b described in eq. (C.10). The left figure shows a and right shows b , from the DSMF's of 90 differently initialized models. The green triangles denote the mean and the orange lines denote the median. The bottom and top of the boxes mark the 25% and 75% percentiles respectively. The black dots are outliers.

Figure C.4 contains box-plots of the trainable parameters, a and b described in eq. (C.10), of the DSMF's belonging to the seen datasets, i.e., 90 maps per dataset. While there are significant outliers, depending on the initialization, the majority of the DSMF's for each dataset are very similar. This is evident from the boxes which contain the parameters from 50% of the initializations. This is more evidence that the DSMF's are in fact consis-

tently used by the network to model specific information about each dataset. Since the DSMF's are trained jointly with their respective CNN's, variations in DSMF's can be compensated for by the CNN and vice versa, which means that a large spread of parameters, a and b , across initializations is not necessarily indicative of a similar spread in output predictions.

6 Conclusion

We proposed and investigated a training strategy for data-driven speech intelligibility predictors, using dataset-specific mapping functions. The proposed strategy allows the use of pooled listening test datasets during network training, without specializing to the paradigms of those listening tests. Solving this problem is important, because training of data-driven SI predictors almost inevitably involves the use of listening test data obtained from multiple listening tests employing different paradigms. Without these proposed dataset-specific mapping functions, data-driven SI-predictors trained on pooled listening test datasets undesirably learn an internal representation of the psychometric functions particular to the listening test paradigms included in the training data. This can cause the trained SI-predictor to perform poorly, or even fail, when employed on new unseen data. To demonstrate this, ten listening test datasets were used to train, validate and test instances of a data-driven SI predictor using this training strategy. The dataset-specific mapping functions consisted of trainable logistic functions at the output of the architecture, which were designed to absorb the different psychometric functions of the datasets, thus preventing an inefficient internal representation of these functions from being learned. Experiments were designed to test the efficacy of training with these dataset-specific mapping functions, along with the generalizability of the predictor. Using the dataset-specific mapping functions for training and validation improved the test performance of the network. A cross validation experiment, where each dataset was excluded from the training set one by one, demonstrated that the network generalized well to new listening conditions and test paradigms, with performance on par with state of the art classical speech intelligibility predictors, for datasets that were not seen during training, and improved performance for seen datasets.

References

- [1] A. MacPherson and M. A. Akeroyd, "Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey," *Trends in Hearing*, vol. 18, p. 2331216514537722, Jun. 2014.

- [2] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
- [3] S. V. Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, Jan. 2018.
- [4] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.
- [5] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.
- [6] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.
- [7] T. H. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Aug. 2010.
- [8] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, no. 2-3, pp. 331–348, Oct. 2003.
- [9] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [10] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, Nov. 2014.
- [11] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, pp. 84–94, Jun. 2016.
- [12] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," *ICASSP*, pp. 624–628, Mar. 2016.
- [13] P. Seetharaman, G. J. Mysore, P. Smaragdakis, and B. Pardo, "Blind Estimation of the Speech Transmission Index for Speech Quality Prediction," *ICASSP*, pp. 591–595, Apr. 2018.

- [14] K. Kondo, K. Taira, and Y. Kobayashi, "Binaural speech intelligibility estimation using deep neural networks," *Interspeech*, pp. 1858–1862, Sep. 2018.
- [15] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.
- [16] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A Neural Network for Monaural Intrusive Speech Intelligibility Prediction," *ICASSP*, pp. 336–340, May 2020.
- [17] T. Houtgast and H. J. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acta Acust. united Ac.*, vol. 25, no. 6, pp. 355–367, Dec. 1971.
- [18] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, Jan. 2014.
- [19] J. B. Allen, "The articulation index is a shannon channel capacity," in *Auditory Signal Processing*. Springer, 2005, pp. 313–319.
- [20] A. Leijon, "Articulation index and shannon mutual information," in *Hearing—From Sensory Processing to Perception*. Springer, 2007, pp. 525–532.
- [21] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, Sep. 2013.
- [22] M. B. Pedersen, M. Kolbæk, A. H. Andersen, S. H. Jensen, and J. Jensen, "End-to-end Speech Intelligibility Prediction Using Time-Domain Fully Convolutional Neural Networks," *INTERSPEECH*, Oct. 2020.
- [23] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.
- [24] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *INTERSPEECH*, pp. 1993–1997, Aug. 2017.
- [25] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *INTERSPEECH*, pp. 3642–3646, Aug. 2017.

References

- [26] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *J. Acoust. Soc. Am.*, vol. 145, May 2019.
- [27] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Predicting the Intelligibility of Noisy and Nonlinearly Processed Binaural Speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.
- [28] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.
- [29] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, Mar. 2014.
- [30] T. Bentsen, A. A. Kressner, T. Dau, and T. May, "The impact of exploiting spectro-temporal context in computational speech segregation," *J. Acoust. Soc. Am.*, vol. 143, no. 1, pp. 248–259, Jan. 2018.
- [31] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, Mar. 2013.
- [32] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a danish sentence test in noise," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, Jan. 2003.
- [33] J. B. Nielsen and T. Dau, "Development of a Danish Speech Intelligibility Test," *Int. J. Audiol.*, vol. 48, no. 10, pp. 729–741, 2009.
- [34] J. Koopman, R. Houben, W. A. Dreschler, and J. Verschuure, "Development of a speech in Noise Test (matrix)," *Proc. 8th EFAS Congr., 10th DGA Congr.*, 2007.
- [35] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
- [36] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial Noises with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.

References

- [37] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, Dec. 2014.
- [39] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: <https://www.tensorflow.org/>
- [40] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Evaluation of Intrusive Instrumental Intelligibility Metrics," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, Jul. 2018.
- [41] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, Feb. 2006.
- [42] E. J. Williams, "The comparison of regression variables," *J. Royal Stat. Society, Ser. B*, vol. 21, no. 2, pp. 396–399, 1959.

References

Paper D

Data-Driven Non-Intrusive Speech Intelligibility Prediction using Speech Presence Probability

Mathias Bach Pedersen, Søren Holdt Jensen, Zheng-Hua Tan
and Jesper Jensen

The paper has been submitted to
IEEE/ACM Transactions on Audio Speech and Language Processing.

This paper is currently under review.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-733-8

AALBORG UNIVERSITY PRESS