Non-Intrusive Speech Intelligibility Prediction

Sørensen, Charlotte

Publication date:
2019

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

NON-INTRUSIVE SPEECH INTELLIGIBILITY PREDICTION

BY
CHARLOTTE SØRENSEN

DISSERTATION SUBMITTED 2019

AALBORG UNIVERSITY
DENMARK

Non-Intrusive Speech Intelligibility Prediction

Ph.D. Dissertation

Charlotte Sørensen

Dissertation submitted December, 2019

Dissertation submitted: December, 2019

University PhD Supervisor: Prof. Mads Græsbøll Christensen
Aalborg University

Industrial PhD Supervisor: Ph.D. Jesper Bünsow Boldt
GN Hearing

PhD committee: Associate Professor Sofia Dahl (chair)
Aalborg University

Tao Zhang, PhD Director
Starkey Hearing Technologies

Associate Professor Tiago Henrique Falk
Director, MuSAE Lab

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

Department: Department of Architecture, Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-560-4

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2020

About the Author

Charlotte Sørensen

Charlotte Sørensen received her B.S. and M.Sc. degree in Biomedical Engineering and Informatics from Aalborg University, Aalborg, Denmark, in 2012 and 2014, respectively. She is currently employed at GN Hearing A/S and affiliated with Audio Analysis Lab, Department of Architecture, Design and Media Technology, Aalborg University, while pursuing the Ph.D. degree at Aalborg University in collaboration with GN Hearing. Her main research interests include speech signal processing with emphasis on speech intelligibility prediction, measurement of speech intelligibility and speech enhancement for hearing aid applications.

Abstract

The ability to communicate through speech is important for social interaction. We rely on the ability to communicate with each other even in noisy conditions. Ideally, the speech is easy to understand but this is not always the case, if the speech is degraded, e.g., due to background noise, distortion or hearing impairment. One of the most important factors to consider in relation to such degradations is speech intelligibility, which is a measure of how easy or difficult it is to understand the speech. In this thesis, the focus is on the topic of speech intelligibility prediction.

The thesis consists of an introduction to the field of speech intelligibility prediction and a collection of scientific papers. The introduction provides a background to the challenges with speech communication in noisy conditions, followed by an introduction to how speech is produced and perceived by the listener. After this, the topic of speech intelligibility and the factors governing speech intelligibility is covered. Finally, the concept of speech intelligibility prediction is introduced and a background to existing intrusive and non-intrusive speech intelligibility prediction measures is provided.

The primary contribution of the thesis is the collection of papers, which propose objective measures for non-intrusive speech intelligibility prediction. The measures are based on the same approach in which an existing intrusive speech intelligibility measure is extended such that it can predict speech intelligibility non-intrusively without access to a clean reference signal. The principle is to estimate a reference signal from its degraded counterpart and use this as input to an intrusive measure. The difference between them lies in how the reference signal is estimated, where they can broadly be divided into two approaches to the problem; Paper A, B and F propose a multichannel solution to the problem, where the spatial content of the desired source is used to extract the signal, while paper C-E propose a single-channel solution, where the reference signal is estimated by finding a combination of signals from a model of the speech production system, which best fits the data. The measures are shown to be well correlated with both the intrusive scores and data from subjective listening tests.

Resumé

Evnen til at kommunikere gennem tale er vigtig for den sociale interaktion. Vi er afhængige af evnen til at kunne kommunikere med hinanden i selv støjfyldte omgivelser. Ideelt er talen let at forstå, men dette er ikke altid tilfældet, hvis talen er forringet f.eks. grundet meget baggrundsstøj, forvrængning eller høretab. En af de vigtigste faktorer man skal tage højde for i forhold til sådanne forringelser er taleforståeligheden. I denne afhandling er fokus på emnet prædiktion af taleforståelighed.

Afhandlingen består af en introduktion til området indenfor prædiktion af taleforståelighed og en samling af videnskabelige artikler. Introduktionen giver en baggrund til udfordringerne med kommunication i støjfyldte omgivelser efterfulgt af en introduktion til, hvordan tale bliver genereret og modtaget. Efter dette bliver emnet omkring taleforståelighed, samt de faktorer, der påvirker taleforståeligheden dækket. Til slut introduceres emnet omkring prædiktion af taleforståelighed, både med og uden adgang til et rent referencesignal.

Det primære bidrag af afhandlingen er samlingen af artikler, hvor der foreslås objektive mål til reference-fri prædiktion af taleforståelighed. Taleforståelighedsmålene er baseret på den samme tilgang, hvor et eksisterende taleforståelighedsmål, der kræver adgang til et rent referencesignal, videreudvikles, således de kan prædiktere taleforståelighed uden adgang til referencesignalet. Princippet er at estimere et referencesignal ud fra det støjfyldte signal og bruge dette som input til et taleforståelighedsmål, som kræver et referencesignal. Forskellen mellem metoderne består i, hvordan reference signalet er estimeret. Metoderne kan generelt inddeles i to tilgange; Artikel A, B og F foreslår en multi-kanalsløsning til problemet, hvor signalets spatiale karakteristik bruges til at estimere referencesignalet, mens artikel C-E foreslår en én-kanalsløsning, hvor referencesignalet estimeres ved at finde en kombination af signaler fra en model af taleproduktionssystemet, der afspejler dataet bedst. De foreslåede metoder er godt korreleret med både taleforståelighedsmålene med adgang til referencesignalet og data fra subjektive lytteforsøg.

List of Publications

The main body of this thesis consists of the following papers:

[A] C. Sørensen, J. B. Boldt, F. Gran and M. G. Christensen, "Semi-Non-Intrusive Objective Intelligibility Measure using Spatial Filtering in Hearing Aids", in *Proceedings of the 24th European Signal Processing Conference (EUSIPCO)*, pp. 1358–1362, Budapest, Hungary, 2016.

[B] C. Sørensen, A. Xenaki, J. B. Boldt and M. G. Christensen, "Pitch-Based Non-Intrusive Objective Intelligibility Prediction", in *Proceedings of the 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 386–390, New Orleans, United States, 2017.

[C] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt and M. G. Christensen, "Non-Intrusive Intelligibility Prediction using a Codebook-Based Approach", in *Proceedings of the 25th European Signal Processing Conference (EUSIPCO)*, pp. 216–220, Kos, Greece, 2017.

[D] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt and M. G. Christensen, "Non-Intrusive Codebook-Based Intelligibility Prediction", in *Speech Communication* Vol. 101, pp. 85–93, 2018.

[E] C. Sørensen, J. B. Boldt and M. G. Christensen, "Validation of The Non-Intrusive Codebook-Based Short Time Objective Intelligibility Metric for Processed Speech", in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 4270–4274, Graz, Austria, 2019.

[F] C. Sørensen, J. B. Boldt and M. G. Christensen, "Harmonic Beamformers for Non-Intrusive Speech Intelligibility Prediction", in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 4260–4264, Graz, Austria, 2019.

In addition, the following academic presentations have been given during the project:

C. Sørensen, "Semi-Non-Intrusive Objective Intelligibility Measure using Spatial Filtering in Hearing Aids," presentation at the *Audio Analysis Workshop*, Aalborg, Denmark, 2016.

C. Sørensen, "Non-Intrusive Intelligibility Prediction using a Codebook-Based Approach," presentation at the *Audio Analysis Workshop*, Aalborg, Denmark, 2017.

C. Sørensen, J. B. Boldt and M. G. Christensen, "Methods for Non-Intrusive Objective Intelligibility Prediction," poster presentation at the *International Symposium on Auditory and Audiological Research*, Nyborg, Denmark, 2017.

Furthermore, the following patents have been filed in relation to the project:

C. Sørensen, J. B. Boldt, A. Xenaki, M. S. Kavalekalam and M. G. Christensen, "Hearing Device and Method with Non-Intrusive Speech Intelligibility," European patent application, EP17181107.8, 2017.

C. Sørensen, J. B. Boldt, A. Xenaki and M. G. Christensen, "Hearing Device, Method and Hearing System," European patent application, EP17158989.8, 2017.

J. B. Boldt, C. Sørensen and R. B. Johannesson, "Speech Intelligibility-Based Hearing Devices and Associated Methods," European patent application, EP17170175.8, 2017.

Contents

Preface

This thesis is written as documentation of the work underlying the Ph.D. project "Non-Intrusive Speech Intelligibility Prediction" at the Technical Doctoral School of IT and Design, Aalborg University. The project was carried out in the period between February 2015 and December 2019 as a joint collaboration between GN Hearing A/S and Aalborg University, supported by the Danish Innovation Foundation as part of the "Cocktail Party Project". My workplace was mainly with the research group in GN Advanced Science at GN Hearing's headquarter in Copenhagen, as well as the Audio Analysis Lab at the Department of Architecture, Design and Media Technology at Aalborg University.

The thesis is divided into two parts: An introduction to the field of speech intelligibility prediction and a collection of scientific papers published during the course of the Ph.D. The project has both professionally and personally been a memorable and enriching experience.

I would like to take this opportunity to express my sincere appreciation and gratitude to all the people, who have crossed by path during this process. In truth, I could not have accomplished this work without the unwavering support from my supervisors, colleagues, family, and friends. Especially, I owe a special thanks to my highly skilled and passionate supervisors Mads Græsbøll Christensen and Jesper Bünsow Boldt who provided patient advice and guidance throughout my entire doctoral studies. Mads always gave me competent and insightful comments as well as a valuable insight into the academic world and a broader perspective on possible solutions and ideas. Jesper has supported me immensely both professionally and personally, which has been absolutely invaluable. He has provided me with a unique insight to industrial research and applied science. Through his pedagogical way of patiently explaining complicated concepts and empathetic leadership he has become a role model who I hope to follow one day. I am grateful for the magnificent support and commitment from both of you throughout the entire project and, especially, when things were difficult and you helped me to move on. I also wish to thank Fredrik Gran and Angeliki Xenaki, who both contributed with valuable supervision in the initial stages of the project. Hav-

ing spent almost five years at the Research Group in GN Hearing, I have had the privilege to work with some wonderful people to whom I wish to thank for your sense of humor as well as many interesting and fruitful discussions and I am happy for being able to stay with you afterward.

Above all, I wish to thank my family and friends, who supported me with love and understanding, but most of all, Michael for his endless love, patience and encouragement when I was in doubt during the process. Thank you for always being there for me.

Charlotte Sørensen
Copenhagen, December 11, 2019

Part I

Introduction

Introduction

1 Speech Communication

Speech plays a central role in human interaction impacting how we understand and communicate with the world around us [98]. Humans rely on the ability to speak with each other in order to exchange valuable information such as knowledge, ideas, opinions and feelings.

Ideally, the speech is easy to understand, i.e. intelligible, without any degradation of the speech. However, if the speech is unintelligible, e.g., due to hearing impairment, distortion in telecommunication systems or background noise, it can given the importance of speech have detrimental effects on the ability to communicate and interact with each other and, thus, lead to social isolation. In order to overcome this challenge, research into the development of speech enhancement algorithms have been of great interest in many applications [50, 70], e.g., hearing aids [36], telecommunication systems [54, 96], and architectural acoustics [47]. Such algorithms can be helpful in difficult situations with high background noise in order to increase intelligibility. However, in less noisy conditions, the same algorithms may have a negative impact on the speech quality or the naturalness of the sound [70, 99].

In order to push the limits for the performance and possibilities of the speech enhancement algorithms it is increasingly important to understand the underlying factors governing speech intelligibility and communication. A lot is known about the physiology of the human speech production system [83], i.e., the vocal tract, and the hearing system, i.e., the ear. On the other hand, very little is known about how the brain processes speech [77]. For example, humans have a remarkable ability to perceive speech even in adverse listening conditions with background noise, reverberation and competing speakers [12]. This ability was coined by Edward Colin Cherry in 1953 with the question: "*How do we recognize what one person is saying when others are speaking at the same time?*" [18], but was already considered by Hermann von Helmholtz in 1870 [101]. Helmholtz described how "*...the ear is able to distinguish all the separate constituent parts of the confused whole*" even in the presence of a mixture of sound that is "*complicated beyond conception*" in

the adverse listening environment of the ball-room with competing speakers, music, clinking glasses, rustling garments etc. This ability to attend to one particular speaker, while filtering away interfering speakers and background noise has been termed the "cocktail party phenomenon" [11].

Speech communication in difficult listening environments such as the cocktail party scenario can be studied in various ways with emphasis on, e.g., listening effort [87], speech quality [70] or speech intelligibility [36]. This thesis has focused on the field of speech intelligibility as this is the foundation for even being able to communicate through speech. In order to understand and model the factors governing speech intelligibility it is necessary first to understand how speech is produced and perceived by the listener, which will be covered in the following parts.

1.1 Speech Production

An advantageous anatomy of the speech production system, i.e. the vocal tract and lungs, and large portions of the brain make it possible for humans to produce sounds in ways no other animal can [83]. The formulation of sentences are carried out in the brain, which produces a series of motor commands controlling the movement of the muscles in the vocal tract and lungs in order to produce the intended sound wave [66].

In order to get an understanding of the properties of the speech signal, it is necessary to briefly consider the details of the mechanical system of the speech production system. The ways of producing speech can generally be separated into three different categories depending on the involvement of the lungs and the different parts of the vocal tract: 1) voiced speech that includes vowels, and unvoiced speech that includes 2) fricatives (e.g. [s] and [f]), and 3) plosives (e.g. [p] and [k]) [83].

In the first case of the vowel sound, the lungs build up air pressure, which sets the vocal folds into a vibratory motion. The vocal folds then convert the steady air flow into a series of periodical bursts at a rate of 60 to 300 Hz [20] determining the fundamental frequency of the speech, i.e., the pitch, which can be considered as the excitation signal of the speech signal [83]. The pitch of the speech signal can be changed by changing the shape of the vocal folds. The excitation signal is then transformed by resonances of the vocal tract, containing the pharyngeal, oral and nasal cavities, which can be described as a time-varying filter of the excitation signal as illustrated in Figure 1. Different resonance characteristics are created depending on the shape and size of these cavities determining the formants, which facilitates the production of different vowel sounds.

In the second case of the fricative, air from the lungs passes through the open vocal cords into the vocal tract, where constrictions in the tract produce turbulence giving rise a noise-like sound [83]. In the third case of the plo-

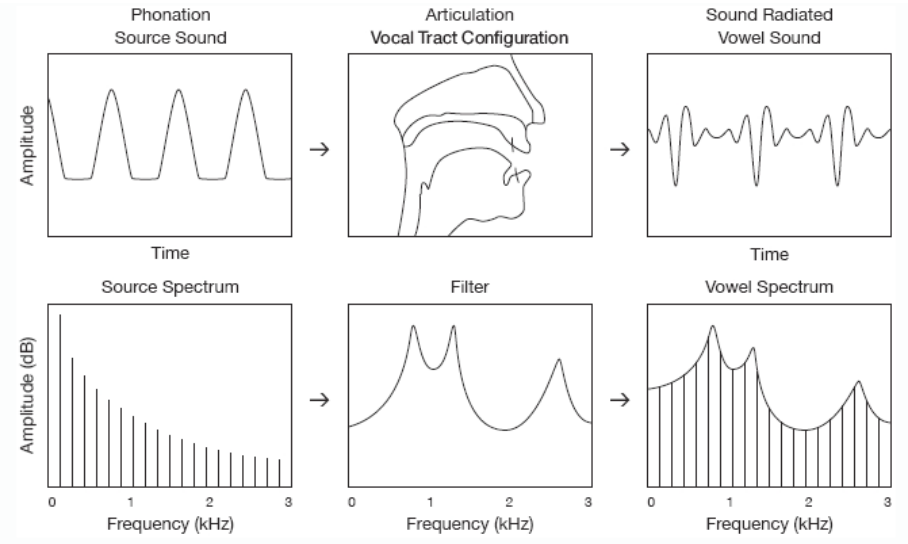


Fig. 1: A schematic drawing of the production of the vowel adapted from [73]. The vocal tract tract acts a time-varying filter on a periodical excitation signal.

sive sound, also known as the stop, the sound is produced by either a rapid release of a constriction in the vocal tract or a sudden constriction of the air flow giving rise to trancient clicks and pops [64]. Opposite to the vowel and fricative, the stop sound does not require air pressure from the lungs.

A Source Filter Speech Model

The above described process for production of speech can be explained in a simplified manner by the "source-filter theory of speech production" based on the experiments by Johannes Müller in 1848 [83]. In this model an acoustic excitation signal, representing the source signal from the lungs, is modulated by a time-varying filter function, representing the vocal tract, which results in a shaped spectrum with broadband energy peaks. The excitation signal for the unvoiced speech signal can be modeled by white Gaussian noise, whereas the voiced speech signal can be modeled by a periodic signal.

Given the source-filter speech model, the speech signal can then be modeled as a stochastic auto-regressive (AR) process in which the excitation signal of the speech is given by white Gaussian noise and the AR parameters determine the filter coefficients:

$$s(n) = -\sum_{i=1}^{P} a_s(i)s(n-i) + u(n), \tag{1}$$

which can also be expressed in vector notation as

$$u(n) = \mathbf{a}_s^T \mathbf{s}(n) \tag{2}$$

where P is the order of the AR process, $\mathbf{s}(n) = [s(n), s(n-1), \ldots, s(n-P)]^T$ is a vector collecting the P past speech samples, $\mathbf{a}_s = [1, a_s(1), a_s(2), \ldots, a_s(P)]^T$ is a vector containing the speech AR parameters with $a_s(0) = 1$, and $u(n)$ modeling the excitation signal.

The AR model has been widely used for modeling the speech production system [42]. However, it should be noted that it is a very simplistic model that does not account for the nasal cavity. Furthermore, using white Gaussian noise as excitation signal is only a suitable model for unvoiced speech and less representative for voiced speech [24]. Nevertheless, it can be considered appropriate for low-dimensional representations of the speech spectrum.

A Harmonic Speech Model

The voiced speech signal can more appropriately be modeled by a harmonic speech model [19]. For example the all-voiced utterance "Why where you away a year, Roy?" shown in Figure 2 can be appropriately modeled by a harmonic speech model.

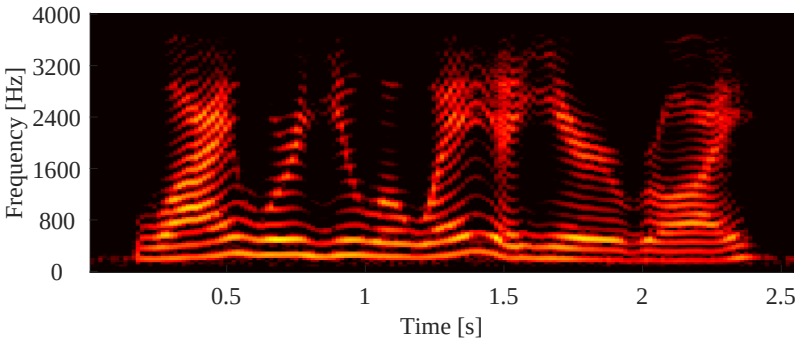


Fig. 2: A spectrogram of the voiced sentence: "Why where you away a year, Roy?" from the corpus in [21].

In the harmonic speech model, the speech signal is modeled as a sum of complex sinusoids with frequencies that are multiples of the fundamental frequency, i.e. the pitch, also known as harmonics. As such, the speech signal can be modeled as:

$$s(n) = \sum_{l=1}^{L} a_l e^{j\omega l n}, \tag{3}$$

where L is the number of harmonics, i.e., model order, the complex amplitude of the lth harmonic of the speech signal is given by $a_l = A_l e^{j\phi_l}$ and the phase is given by ϕ_l.

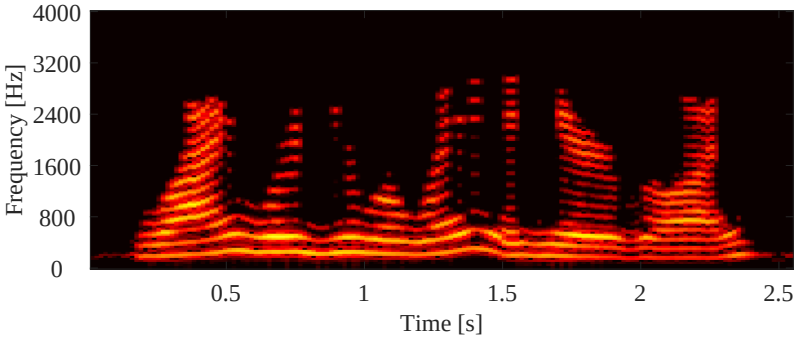


Fig. 3: A spectrogram of the reconstruction of the voiced sentence: "Why where you away a year, Roy?" using the harmonic speech model.

The speech signal is highly non-stationary and, thus, the pitch of the speech signal varies over time. However, over a sufficiently short duration of time of approximately 10-30 ms the spectral characteristics are fairly stationary such that the non-stationary nature of the speech signal can be accounted for by assuming the speech signal to be quasi-stationary over short time windows [70]. During such a time window the parameters of the harmonic speech model, i.e., the complex amplitude, the model order and the pitch, are assumed to be constant and can, thus, be estimated in order to model the voiced speech signal. A reconstruction of the utterance in Figure 2 using the harmonic speech model is depicted in Figure 3.

1.2 Speech Perception

An advantageous anatomy of the auditory system make it remarkable in its ability to sense acoustic stimuli across a wide dynamic range (0-140 dB) and a wide range of frequencies (16 Hz to 20 kHz). When the acoustic signal reaches the ear of the listener, it first hits the pinna, i.e., the external part of the ear [78]. The pinna causes multiple reflections of the sound signal, which can give a cue about the direction of arrival of the signal. The acoustic sound wave travels through the ear canal until the eardrum, where it is converted into mechanical vibrations. The mechanical vibrations are transferred into the inner ear through the ossicular chain consisting of the malleus, incus and stapes. Inside the inner ear, the cochlea converts the mechanical vibrations into neural activity [78]. The basilar membrane within the cochlea is mechanically tuned to resonate at different frequencies along its length ranging from

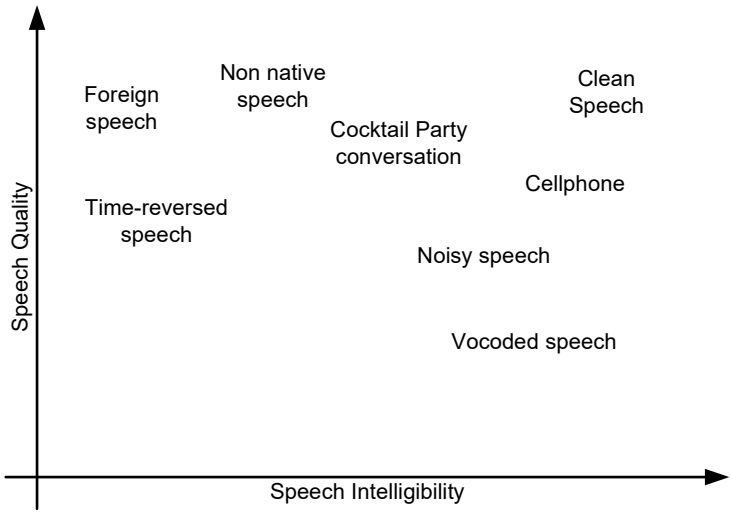


Fig. 4: It is important to differentiate between speech intelligibility and quality, since they are not necessarily related. As such, a high speech quality does not per default result in a high speech intelligibility and vice versa.

high frequencies at the base to low frequencies at the apex [77]. As such, the basilar membrane functions as spatially distributed bandpass filters [75]. In the basilar membrane the mechanical vibrations are converted into electrical signals through the displacement of hair cells located on the organ of Corti. The electrical signals are then sent to the primary cortex in the brain through the auditory nerve and auditory pathway [78]. A deep review of the functioning of the human brain is beyond the scope of this thesis but a more detailed description of the anatomy ans physiology of the ear and brain can be found in [77, 78].

2 Speech Intelligibility

A helpful way to study the perception of speech is to consider the fundamentals of speech intelligibility. Generally, intelligibility can be viewed as a measure of how comprehensible speech is in a given acoustic environment. Intelligibility is measured by how much of the speech is correctly identified and not by how much of the speech that is correctly understood, since some listening test material includes nonsense words, which can only be identified correctly but not understood [1].

When considering the topic of speech intelligibility, it is also important to differentiate between speech intelligibility and speech quality. Speech quality is a measure of the naturalness, clarity and distortion of the speech signal.

Both measures are of high relevance when trying to improve the perception of speech with, e.g., speech enhancement algorithms. However, increasing both speech intelligibility and speech quality can sometimes be two conflicting goals as shown in Figure 4. An improvement in speech intelligibility does not necessarily result in an improvement in speech quality or vice versa [70]. As such, it is important to have a clear focus on the objective, when employing or developing speech enhancement algorithms.

After having gained an understanding of how speech is produced and perceived, it is of interest to consider how different factors influence speech intelligibility. The objective of the following sections is not to provide a comprehensive analysis of the factors influencing speech intelligibility, but to identify the most important elements that might influence the intelligibility. The most important factors influencing the speech intelligibility are the characteristics of the speech or the degradation of the speech (from the acoustic environment, telecommunication systems or hearing aids), which will be covered in the following sections.

2.1 Speech Cues Influencing Speech Intelligibility

Whether the goal is to improve or predict the intelligibility of speech, it is important to obtain an understanding of which cues of the speech that make it intelligible. A number of studies have investigated the relationship between specific speech cues and speech intelligibility [12, 31], e.g. the fundamental frequency [8, 26] and temporal envelope [30, 89]. However, despite the extensive research for many decades on speech intelligibility it has not been possible to completely answer which speech cues make speech intelligible. The remainder of this section will not provide an extensive review but shortly summarize some of the most important speech characteristics contributing to speech intelligibility.

Temporal Envelope

The temporal envelope of the speech is characterized by the slow variation in the overall amplitude of the speech signal over time (see Figure 5). The envelope may be an important carrier of the semantic information of the speech signal [31, 39, 92] with the most important amplitude modulations for speech intelligibility being in the range between 4 and 16 Hz [32]. Only presenting the temporal envelope of the speech signal without the temporal fine structure (the instantaneous variation in the sound pressure) still preserves a high speech intelligibility in noise free conditions, as experiments with vocoding [30, 89] and chimaeric speech [92] show.

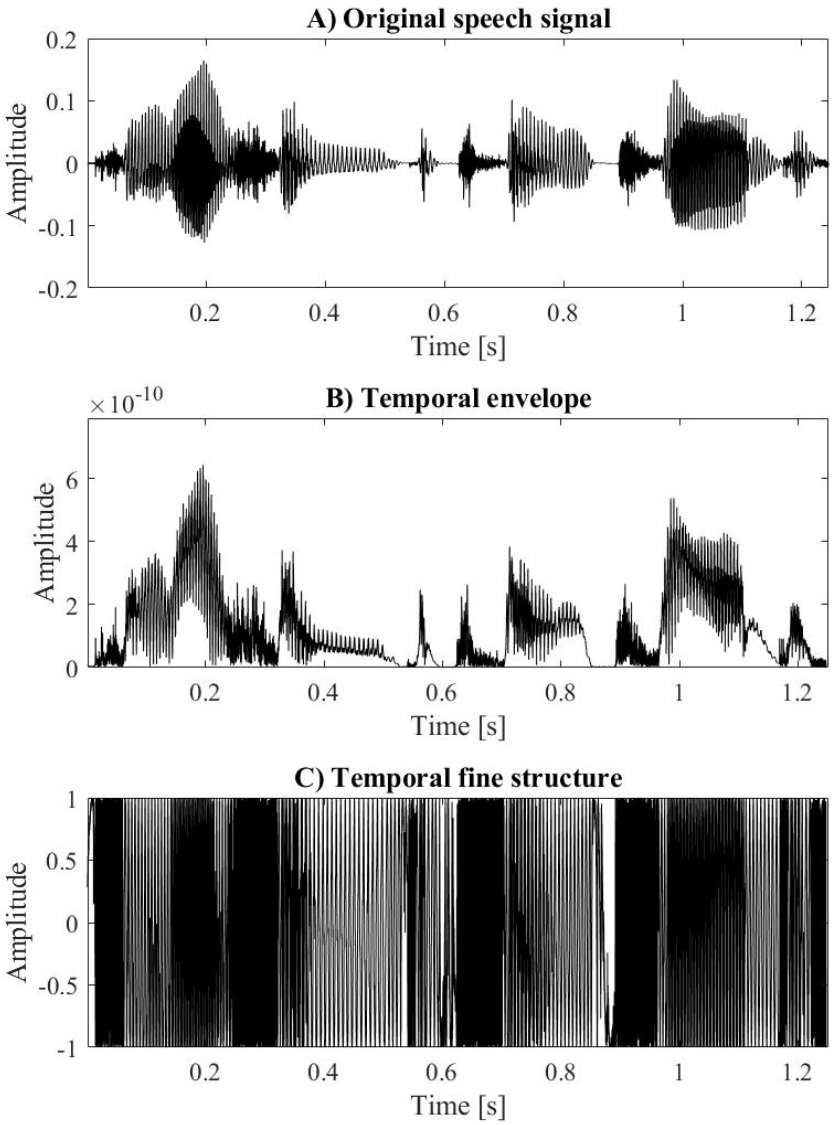


Fig. 5: A) The original speech signal. B) The temporal envelope of the speech signal, which shows slow fluctuations over time. C) The temporal fine structure of the speech signal, which shows no fluctuation over time.

Pitch

Pitch is a strong cue for grouping speech signals aiding the ability to follow a single speaker in noisy conditions [8, 14, 26]. Particularly, speech intelligibility is significantly higher when the target and interferer are uttered by different genders instead of being uttered by a talker of the same gender [11]. The harmonicity of the pitch, i.e. integer multiples of the frequency of the pitch, is a helpful cue for grouping the signal across frequencies in order to determine whether a sound segment belongs to the same speaker [51]. However, except for tonal languages the pitch is not a fundamental cue for speech intelligibility in noise free conditions, since the temporal envelope is sufficient for carrying the semantic meaning of the speech signal as already mentioned [30, 89, 92].

Formant Frequencies

Formant frequencies are the spectral peaks of the short-term spectrum, which can be seen in Figure 1 on page 5 in last panel in the bottom row. The vowels consist primarily of the two lowest formants between 300 and 2500 Hz [29]. The formant contour over the course of a vowel has been shown to be perceptually important and significantly correlated with intelligibility [43, 44].

Pauses

Onsets and offsets during pauses indicate the beginning and ending of words and sentences, which is helpful in segregating the speech and, thus, positively influences the speech intelligibility.

Duration

The duration of the words together with the length of the pauses determine the speaking rate, which is an important component to speech intelligibility [93]. Lowering the speaking rate positively influences the speech intelligibility, where the duration especially is an important component for vowel identification [61].

Energy

The overall energy, i.e. intensity, of the speech signal is an important component that significantly affects speech intelligibility [40]. The speech energy is concentrated in spectro-temporal regions such that it will be the dominant contributor in many regions even when mixed with interfering noise [12, 22]. This effect is furthermore enhanced by the logarithmic intensity transformation performed by the auditory system in which a stronger signal will be

dominant when added to a weaker one [12, 25]. The glimpsing model of speech perception in noise has described how a few glimpses in time and frequency of the target speech is sufficient to obtain a high intelligibility [22], which is also utilized in binary masks where all spectro-temporal regions that are not dominated by the target speech are removed [104, 105].

Spatial information

The spatial information about the location of the speaker is useful to segregate the target from interference, especially in noisy conditions [11, 13]. The auditory system benefits from the fact that humans have two ears with the head acting as an acoustic "shadow" such that the signals arriving at the two ears have different interaural time difference (ITD) and interaural level difference (ILD) [13]. The ITD and ILD are helpful cues for unmasking interfering sounds and, thus, an important component for speech intelligibility.

2.2 Speech Degradation

The speech intelligibility can be negatively affected if the speech signal is degraded before it is received by the listener. Numerous factors can contribute to the degradation of the speech signal. Some of the most common contributors to degradation of the speech signal can be attributed to either the effect of the acoustic environment, including additive noise and reverberation, or the effect of a communication channel, e.g. hearing aids.

For example, when two people are having a conversation close to one-another, the speech signal can be degraded by the surrounding acoustic environment such as interfering speakers, background noise and reverberation. If one of the participants in the conversation is also hearing impaired and wearing a hearing aid, this additional communication channel might distort the speech signal due to signal processing. These degradation types, i.e., additive noise, reverberation and distortion, are shortly explained in the following part.

Additive Noise

Noise is present wherever we go, for instance, in restaurants with interfering speech from people talking in nearby tables, the office with noise from PC fans and air ducts or the street with wind noise and cars passing by [70]. Generally, the different noise types are characterized by their temporal and spectral properties. The temporal properties characterize whether the noise is stationary, i.e., remains constant over time (e.g. noise from the PC fan), or the noise is fluctuating, i.e., changes over time (e.g. multiple people speaking in a restaurant). The spectral properties characterize the shape of the spectrum, i.e., the distribution of the energy of the noise in the frequency domain.

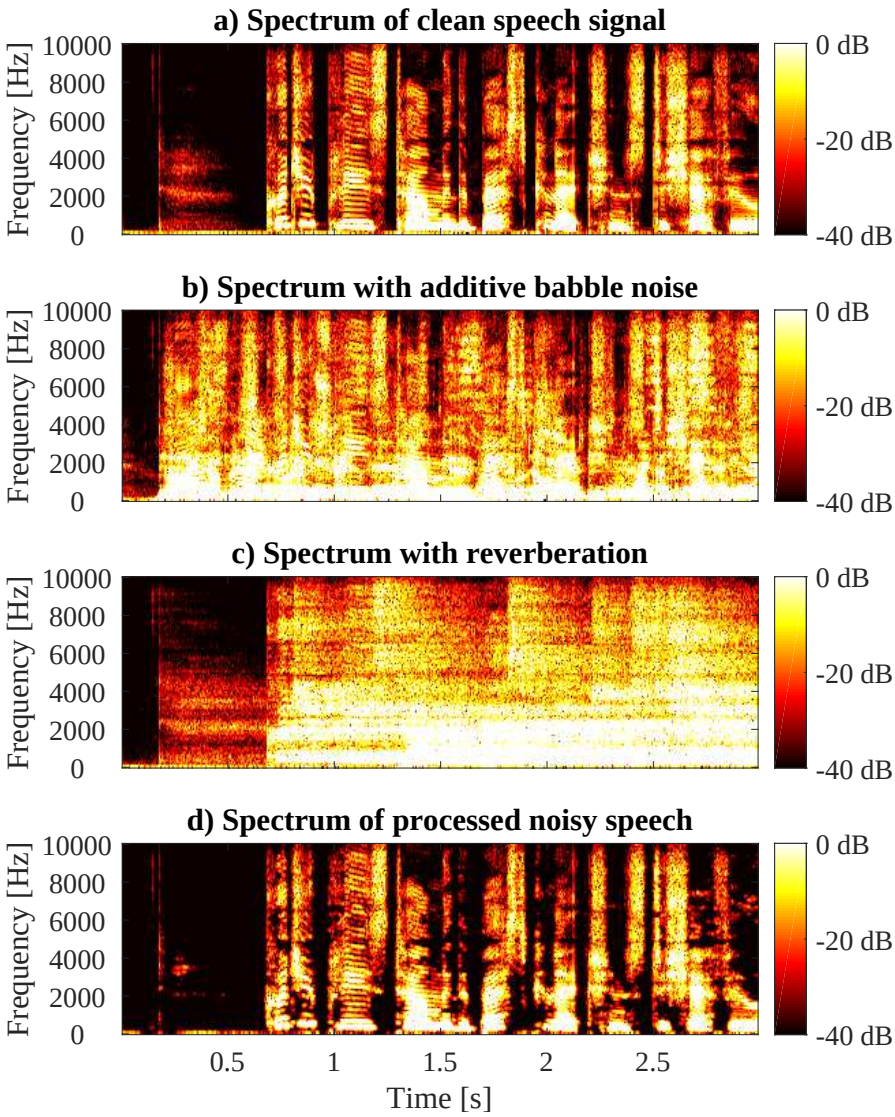


Fig. 6: Examples of the effect of the different degradation types on the speech signal. A) Spectrum of the original clean speech signal. B) Spectrum of the clean signal interfered with additive babble noise. The interfering noise overlap with the clean signal in time and frequency but is highly fluctuating making it possible to listen in the dips. C) The clean spectrum with high reverberation. The spectrum is highly smeared by the reverberation. D) Spectrum of the noisy signal in B) processed with Ideal Binary Masking (IBM) [104, 105]. The processing succeeds in restoring the spectrum of the clean signal but do also introduce distortions such that parts of the time-frequency regions of the clean signal is missing and some of the noise is still evident.

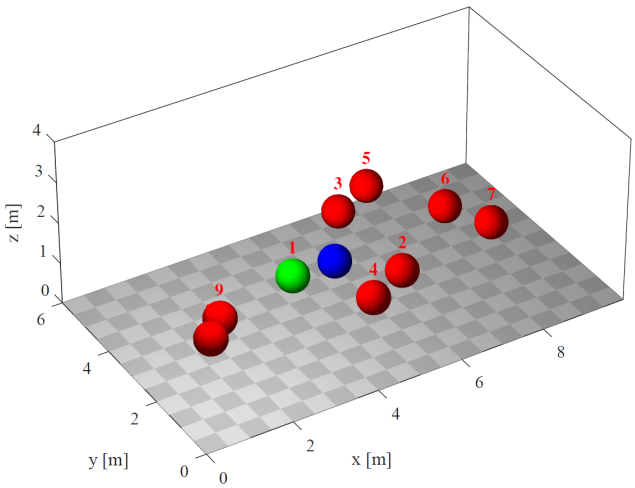


Fig. 7: An example of a simulated cocktail party scenario, where the target speech signal (indicated by the green ball) received at the listener (indicated by the blue ball) is interfered with additive babble noise (indicated by the red balls). Figure adapted from Paper B [94].

Regarding the temporal properties, humans are able to take advantage of glimpses of spectro-temporal regions with high SNR, i.e. listen in the dips, as already mentioned in Section 2.1 such that speech intelligibility is generally higher for strongly fluctuating noises compared to stationary noises [22, 76]. Regarding the spectral properties, the most challenging condition is when the noise has a similar spectrum to speech, i.e., contain energy in the same frequency regions as speech [40]. As such, Speech Shaped Noise (SSN), i.e. stationary Gaussian noise filtered to match the long-term spectrum of speech, is often used as interfering noise in experiments.

Noise from interfering speakers, i.e., babble noise, is especially challenging, since the babble noise is spectrally similar to the target speech, but is also too stationary to allow enough spectro-temporal glimpses to listen in the dips [22, 76]. This thesis primarily focuses on the condition with babble noise, since it is the most challenging scenario.

Reverberation

Reverberation is the effect of the acoustic signal being reflected by floor, walls, ceilings and other surfaces as illustrated in Figure 8. Reverberation can cause temporal smearing of the energy of the signal into higher modulations of the temporal envelope, which can negatively impact the speech intelligibility [11, 37]. The severity of the reverberation is measured by the reverberation time, T60, which is the time it takes for the sound signal to decay 60 dB [63].

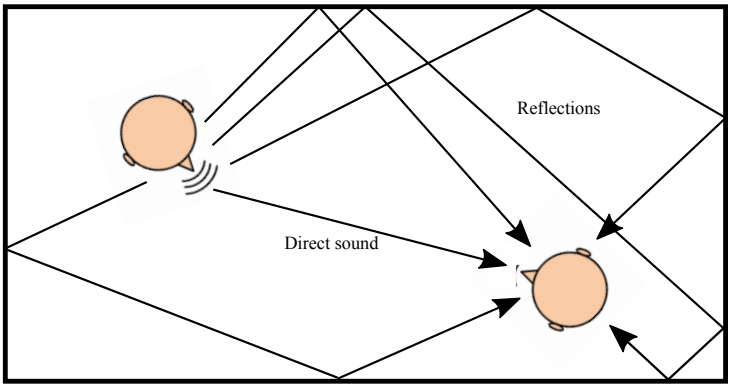


Fig. 8: Reverberation is caused by the acoustic signal being reflected by walls or other surfaces.

Distortion

In a scenario where the speech signal is electronically processed, e.g. in a hearing aid, mobile phone or other electronic device, the signal might be distorted, which can have a negative impact on speech intelligibility [54, 70]. The signal processing can for example include analog-to-digital conversion, compression, speech coding and enhancement, wireless transmission and digital-to-analog conversion [70]. Especially, the effect of speech enhancement algorithms is of interest when trying to improve speech intelligibility. While many speech enhancement algorithms might improve the speech quality, it does not guarantee an improvement in speech intelligibility [70, p. 564–567]. Generally, it has been found that many speech enhancement algorithms due to introduced distortion fail to improve speech intelligibility relative to the unprocessed noisy speech [49, 50, 70, 71]. As such, the effect of speech enhancement on speech intelligibility is an important factor to consider when developing such algorithms.

3 Subjective Evaluation of Speech Intelligibility

After having considered the factors influencing speech intelligibility, it is relevant to look into how intelligibility can be evaluated and measured. Basically, the intelligibility of a degraded, i.e., processed, distorted or noisy speech signal can be assessed by either a subjective or objective evaluation. The subjective evaluation of speech intelligibility is performed by experiments with human listeners, whereas the objective evaluation is performed by means of an algorithm. Generally, results obtained using subjective listening tests are more reliable, but are also more time-consuming, expensive and not applicable to real-time processing. As opposed to subjective listening tests, objective

Table 1: The fixed syntax from the GRID database [23] has a simple, fixed and semantically unpredictable structure consisting of a combination of a command, color, preposition, letter digit and adverb in that order. Table adapted from Paper D.

Command	Color	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	0-9	again
lay	green	by	(no W)		now
place	red	in			please
set	white	with			soon

speech intelligibility prediction algorithms are faster, cheaper and applicable for real-time processing, but less reliable.

Subjective speech intelligibility is evaluated by presenting a stimuli to a test subject in a controlled setting and recording how much was correctly identified. In order to ensure reproducibility and be able control the test setting, the stimuli is often presented via headphones or a loudspeaker array. The difficulty of the task can then be varied by, e.g., changing the Signal-to-Noise Ratio (SNR) or the level of signal processing in a controlled manner. The target stimuli can consist of e.g. short nonsense words [1], words [23], phonemes [74] and sentences [16, 103]. Generally, sentence corpora can be divided into either *syntactically fixed*, i.e., corpora containing sentences generated from a limited collection of words, or *syntactically open*, i.e., corpora containing sentences with no limitation in vocabulary. An example of a syntactically open corpus is the English sentences in the EUROM_1 database [16], which are used for the objective evaluation in Paper A, C and D. Examples of syntactically fixed sentence corpora are the GRID corpus [23] and the Dantale II corpus [103], which are used for the subjective evaluation in Paper D and E, respectively.

In the subjective listening tests in Paper D, the GRID corpus with fixed-syntax sentences was used. The sentences have a simple, syntactically fixed but semantically unpredictable structure, e.g. "set *green* by *C 8* now", which is shown in Table 1. The test subject has to identify the color, letter, and digit. Similarly, the Dantale II corpus, which is the Danish version of the matrix sentence test, has a syntactically fixed structure with *name*, *verb*, *number*, *adjective*, and *object*, e.g., "Michael bought ten pretty presents" [103]. The advantage of both the GRID and Dantale II corpora are that it is impossible to guess the correct response from the context. Furthermore, the test subjects can indicate their response using a Graphical User Interface (GUI) limiting the need for an experimenter to record the response.

From the response of the test subjects, the intelligibility can be measured either as the number of correctly identified words in percentage at a specific SNR or adaptively changed until a fixed percentage of words are correctly

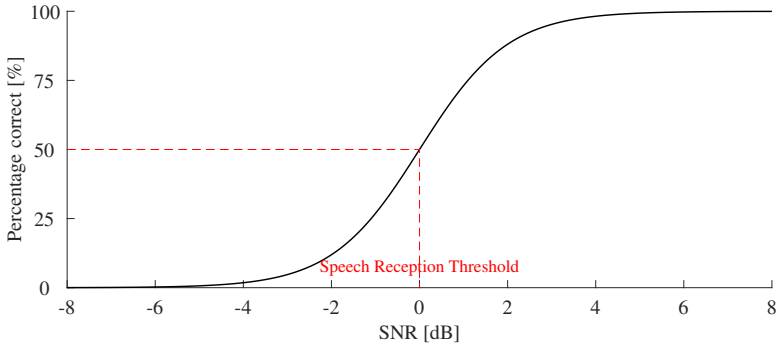


Fig. 9: Speech intelligibility as a function of SNR gives a typical psychometric function, where the Speech Reception Threshold (SRT) is the SNR at which 50 % are correctly identified, marked with the dashed red lines.

identified [10, 67, 68]. In the first case, the measured intelligibility scores vs. SNR result in a psychometric function, i.e, a S-shaped curve approaching 0% for low SNRs and 100% for high SNRs, as shown in Figure 9. In the latter case, the SNR is typically adapted to the level allowing 50% of the words to be correctly identified, also known as the Speech Reception Threshold (SRT) as indicated in Figure 9. It should be noted that a higher SRT corresponds to a lower performance, since it is necessary with a higher SNR before 50% of the words are correctly identified.

4 Objective Evaluation of Speech Intelligibility

Objective evaluations, i.e., objective intelligibility measures, predict the intelligibility scores of the corresponding subjective evaluation based on recordings of the degraded speech signal, i.e., noisy, processed or otherwise distorted signal. The intelligibility score is usually given as a number, e.g. between 0 and 1 [97], which can be calibrated into the percentage of correctly identified words using, e.g., a psychometric function that depends on various factors such as the test material used, the test conditions, etc.

Generally, objective measures are developed to predict the effects of interfering speakers, the acoustic environment (e.g. reverberation) or distortion from signal processing in the transmission channel (e.g. hearing aids or mobile phones) on speech intelligibility. They are based on various models of the human speech perception ranging from very simplified models [9, 40, 97] to more complex models of the auditory system [55, 56, 59]. However, the limited knowledge about the functioning of human speech perception and what makes speech intelligible also poses a limitation for how well the algorithms can model this. As such, it is important to be careful when applying

Table 2: Taxonomy and application conditions for existing intrusive and non-intrusive speech intelligibility prediction metrics. The symbols indicate conditions in which the measure is well documented and recommended (green check marks), the measure can be used with caution either because it only works in some circumstances or has not been sufficiently investigated (yellow check marks), should be avoided (red crosses) or data is not available for the condition (gray question marks). The data is based on [3, 5, 7, 15, 17, 33, 35–37, 41, 53, 55–59, 62, 69, 72, 79–82, 85, 86, 90, 91, 97, 97, 100, 105].

		Stationary noise	Modulated noise	Reverberation	ITFS	Phase Jitter	Spectral subtraction
INTRUSIVE	*AI* [6,40,62]	☑	☒	☒	☒	☒	☒
	SII [7]	☑	☒	☒	☒	☒	☒
	ESII [80,81]	☑	☑	☒	☒	☒	☒
	CSII [58]	☑	☑	☒	☒	?	☑
	HASPI [59]	☑	☑	☑	☑	☑	☑
	STI [96]	☑	☒	☑	☒	☒	☒
	STMI [33]	☑	☑	☑	?	☑	☒
	STOI [97]	☑	☒	☑	☑	☑	☑
	sEPSM [55]	☑	☒	☑	☒	☒	☑
	ESTOI [53]	☑	☑	?	☑	☑	☑
	WSTOI [69]	☑	?	?	☑	☑	☑
	mr-sEPSM [56]	☑	☑	☑	☒	☒	☑
	sEPSMcorr [79]	☑	☑	☑	☑	☑	☑
NON-INTRUSIVE	*SRMR* [37]	☑	☑	☑	☒	☒	☒
	ModA [17]	☑	☑	☑	☒	☒	☒
	LCIA [90]	☑	☑	?	?	?	☑
	NISA [91]	☑	☑	?	?	?	?
	THMMB-STOI [57]	☑	☑	?	?	?	?
	NI-STOI [5]	☑	☑	?	☑	?	?
	CNN-based [3]	☑	☒	?	☑	?	?
	Spatial filtering based STOI [Paper A]	☑	☒	☒	?	?	?
	PB-STOI [Paper B]	☑	☑	☑	?	?	?
	NIC-STOI [Paper C-E]	☑	☑	?	☑	?	?
	HDSB-STOI [Paper F]	☑	☑	☑	?	?	?
	HWB-STOI [Paper F]	☑	☑	☑	?	?	?

☑ The measure is well documented and recommended for this condition
☑ The measure can be used with caution either because it only works in some circumstances or has not been sufficiently investigated
☒ The measure should be avoided to be used for this condition
? Data is not available for this condition

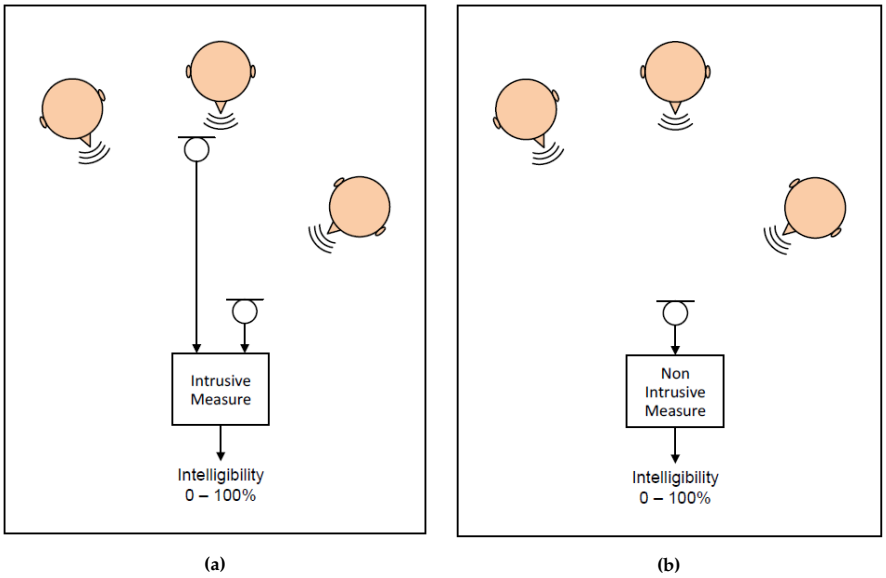


Fig. 10: (a) An intrusive measure requires access to both a clean reference signal and the degraded signal. (b) A non-intrusive measure does only require access to the degraded signal in order to estimate speech intelligibility.

speech intelligibility prediction algorithms and advantages, disadvantages and limitations of the underlying models should be considered in regards to the experimental setup. On the other hand, if these algorithms and, thus, the underlying models of human speech perception, correlate well with the subjective intelligibility results, they can possibly provide insight into how speech intelligibility is obtained [22, 55, 95]. Reliable objective measures are very useful for the development and evaluation of new algorithms, e.g. in hearing aids or telecommunication systems, in order to assure these are functioning as expected and be able to faster test and point out any problems.

Objective measures can broadly be classified as either *intrusive* or *non-intrusive* (see Figure 10). Intrusive measures predict speech intelligibility based on some type of distance metric between a clean speech reference signal and the degraded signal, whereas non-intrusive measures predict the speech intelligibility based solely on the degraded speech signal and, thus, do not require a clean reference signal. Generally, intrusive measures are more reliable, since they have access to information about the clean signal. On the other hand, non-intrusive measures are more practical for real-time applications, when the clean reference signal is not available. The remainder of this section will describe some of the most important speech intelligibility measures for the work in this thesis but will not provide an extensive review of all existing methods. An overview of the metrics described in Section 4.1

and 4.2 is provided in Table 2, which summarizes the conditions in which existing speech intelligibility prediction metrics as well as the measures proposed in this thesis are recommended to be used and to be avoided.

4.1 Intrusive Prediction of Speech Intelligibility

The research on objective speech intelligibility measures started in the 1920s at Bell Laboratories, where the first model was developed by Harvey Fletcher [2, 38]. This work led to the development of the Articulation Index (AI) by French and Steinberg [40], which was thoroughly described by Kryter [62] and later standardized by ANSI [6]. With the emergence of computers, the AI was later extended and modified into the Speech Intelligibility Index (SII) [7].

The AI and SII are based on the assumption that the different frequency bands contribute differently to speech intelligibility. The AI and SII use a weighted average of the SNR of long-term speech excerpts in several frequency bands [7, 40]. As such, the AI and SII can account for intelligibility scores in quiet and in the presence of additive noise but requires that the clean speech signal and the noise signal can be accessed separately. The need for the noise signal in separation implies that the AI and SII are unable to predict speech intelligibility for conditions, where the speech and noise mixture have been subjected to non-linearly processing. Furthermore, the long-term analysis implies that the AI and SII are insensitive to short-term fluctuations implying that the noise has to be stationary, e.g., these models are unable to account for modulated noise and the ability to listen in the dips. These limitations are important to consider, when applying the measures for practical applications, where the noise is often fluctuating, e.g. interfering speech, and speech processing might introduce non-linear distortions.

As an extension of the standard SII, the Extended SII (ESII) was developed in order to improve the measures performance in fluctuating noise conditions [80, 81]. The ESII introduces a short-term analysis, where the speech intelligibility is evaluated in short time frames and then averaged across these [80]. The ESII is able to predict the speech intelligibility well in conditions where the speech signal is additively corrupted with amplitude modulated noise and interrupted noise [81]. However, the ESII still requires that the clean speech signal and the noise signal are available separately and can, thus, not account for conditions where the noisy speech has been non-linearly processed [82]. This limitation of the SII was addressed with the coherence SII (CSII), which was introduced in order to be able to account for the impact of non-linear processing [58]. The CSII is based on the same principles as the SII but is evaluated using the coherence between the clean and degraded signal instead of the SNR [15, 58]. An updated version of the CSII, the Hearing-Aid Speech Perception Index (HASPI), includes a more detailed auditory model and bases predictions on a larger set of features making it

able to account for individual hearing loss [59].

Another approach that was designed to predict the impact of some types of non-linear processing is the Speech Transmission Index (STI) based on the concept of the modulation transfer function [46–48, 96]. The Modulation Transfer Function (MTF) measures the change in the modulation depth of a probe signal over a communication channel [46]. The STI then uses the output from the MTF to predict how well the modulations of the transmitted signal are preserved. However, the STI is not able to predict the intelligibility for pre-recorded signals, since the metric requires the signal to be transmitted across a well-defined communication. As such, several subsequent models were developed based on the STI, which instead use speech as a probe signal [41]. Furthermore, the STI cannot predict the speech intelligibility for more adverse types of non-linear processing, such as spectral subtraction, or for conditions with fluctuating interferers [41, 72]. The Spectro-Temporal Modulation Index (STMI) is an extended version of the STI with a more complex auditory model that evaluates modulations jointly across time and frequency [33].

More recently, the Short-Time Objective Intelligibility (STOI) measure [97] and the speech-based Envelope Power Spectrum Model (sEPSM) [55] have been introduced, which can account for the condition in which the SII- and STI-based approaches fail. The STOI measure assumes that speech intelligibility is related to the correlation between the clean and degraded signal. Despite its simplicity – or due to – it has become very popular as it has been shown to have a high prediction performance. The intelligibility is predicted as the average of the correlation in time-frequency (TF) regions between the temporal envelopes of short excerpts of the clean and degraded signal. The STOI measure accounts well for TF processed speech, such as Ideal Binary Masking (IBM), and different noise conditions [79, 97, 105]. However, the STOI measure is not suitable for predicting the intelligibility of highly reverberant speech [79, 97]. Furthermore, the measure is not able to account for highly fluctuating interference due to a relatively long time window (384ms). An extension of the STOI measure, the Extended STOI (ESTOI) measure [53], has been introduced to improve the performance in fluctuating noise conditions. Another extension is the Weighted STOI (WSTOI) measure [69], which takes the information content of the signal into account by weighing each TF according to this.

The other recent model, sEPSM, is based on the Envelope Power Spectrum Model (EPSM) [28, 34]. The measure estimates the SNR in the envelope-frequency domain based on the intrinsic envelope fluctuations of the degraded signal and the noise signal. The sEPSM can accurately account for the effects of reverberation, additive noise and some types of non-linear processing, such as spectral subtraction, but fails with fluctuating interferers and other types of non-linear processing, such as IBM processing and phase

jitter [79]. A short-time version of the sEPSM, the multi-resolution sEPSM (mr-sEPSM), was introduced in order to better predict the intelligibility of fluctuating interferers [56]. Another notable speech intelligibility measure, the sEPSM^{corr} [79], combines the sEPSM and STOI measure in other to overcome the limitations of each model and utilize the complementary strengths of the two models. The sEPSM^{corr} is performed in the envelope-frequency domain as the sEPSM but uses a cross-correlation back end similar to the one used in STOI. The sEPSM^{corr} is shown to account well for the effects of stationary and fluctuating additive interferers as well as the effects of non-linear processing, such as spectral subtraction, phase jitter and IBM processing but fails to account for the effect of reverberation [79].

4.2 Non-Intrusive Prediction of Speech Intelligibility

The first attempts to estimate speech intelligibility non-intrusively were based on using Automatic Speech Recognition (ASR) techniques [4]. The principle behind this approach is to transcribe the degraded test word and compare the similarity between the transcription and a syntactically fixed list of all possible responses [45, 88]. One approach was proposed by Holube and Kollmeier in 1996 based on a Dynamic-Time-Warping (DTW) ASR recognizer [84] trained using an auditory model [27]. The recognition rate of hearing impaired listeners was predicted for a test set of consonant-vowel-consonant words corrupted by SSN [45]. Another approach employing ASR is based on a Hidden Markov Model (HMM) to predict the intelligibility of a closed matrix sentence test [88]. The ASR-based non-intrusive intelligibility measures can provide a very high performance comparable to the prediction levels offered by the intrusive measures. However, the ASR-based measures are not completely non-intrusive approaches, since they are limited to predicting the intelligibility of a closed set of words. This short-coming is highly relevant for real-world conditions as they are rarely limited to a fixed set of words.

A later approach attempts to predict non-intrusive speech intelligibility based on converting the results of existing non-intrusive *quality* prediction algorithms, e.g. ITU-T P.563 [52] and ANIQUE+ [60], into intelligibility scores [4, 36]. The P.563 measure was the first standardized non-intrusive algorithm by ITU-T in 2004 [52]. The P.563 and ANIQUE+ algorithms extract a number of signal parameters, e.g. level of background noise, signal interruptions and speech robotization, from which they evaluate the unnaturalness, i.e., level of distortion affecting the perceived quality. However, an improvement in speech quality is, as already mentioned in Section 2, not necessarily equal to an improvement in speech intelligibility, and, the P.563 and ANIQUE+ algorithms have in fact been shown to be a poor predictor for speech intelligibility [36, 37].

The first proposed measures that can successfully predict speech *intelligibility* and are truly *non-intrusive* are the Speech-to-Reverberation Modulation energy Ratio (SRMR) [37] and the Modulation spectrum Area (ModA) [17]. They are both based on the principle that reverberation smears the envelopes of the speech signal (see Figure 6c), which affects the modulation spectrum of the speech signal [17, 36, 37]. The SRMR measure computes an intelligibility score based on the ratio between the average modulation energy at low modulation frequencies consistent with clean speech to the average modulation energy at high frequencies consistent with reverberated speech [37]. A number of refined versions of the SRMR measure have been introduced with the aim of improving prediction accuracy in different conditions, such as hearing impairment and cochlear implant users [35, 85, 86]. The ModA measure computes an intelligibility score by calculating the area under the modulation spectrum [17]. The SRMR and ModA can successfully account for speech intelligibility of reverberated and (stationary) noisy speech [17, 36, 37] but has been shown to fail in conditions with speech enhancement and non-linear processing [5, 36].

Another approach to estimate speech intelligibility non-intrusively is to estimate relevant features of the clean signal and use this as the clean reference signal in an intrusive intelligibility metric, the STOI measure [5, 57, 91]. Andersen et al. proposes the Non-Intrusive STOI (NI-STOI) measure [5], which estimates the clean signal envelopes by projecting the modulation magnitude spectrum of the degraded signal into a subspace trained in advance containing only the modulation magnitude spectra consistent with clean speech. The NI-STOI measure can account well for different noise conditions and non-linear processing, such as IBM processing, but fails in conditions where interfering speech is mistaken for the target speech [5]. Furthermore, being based on the STOI measure it will probably fail in conditions with reverberation and highly fluctuating interference.

Recently, machine learning has gained increasingly interest as a means to estimate the clean speech signal and use this to predict speech intelligibility [3, 57, 90, 91]. Generally, machine learning based speech intelligibility prediction measures estimate speech intelligibility from different features, e.g. spectral flatness, envelope and pitch, extracted from the degraded speech signal. One approach proposed by Sharma et al., the Low Cost Intelligibility Assessment (LCIA) measure [90], uses a Gaussian Mixture Model (GMM) to compute an intelligibility score from features such as the spectral flatness, spectral centroid, spectral dynamics and excitation variance. The GMM is trained in a supervised manner using subjective intelligibility scores as the desired output. However, since subjective scores are time-consuming and expensive to collect, the lack of subjective training data poses a limitation for how well generalizable the approach is for other conditions than the one it is trained for even though the LCIA measure shows a high correlation

with subjective intelligibility scores [90]. A refined version of the LCIA measure, the Non-Intrusive Speech Assessment (NISA) measure [91], attempts to overcome this limitation by using objective intelligibility scores from an established intrusive objective intelligibility measure, e.g. STOI, to train a tree based regression in a supervised manner. The methods shows high correlation with objective intelligibility scores from STOI but has not been further evaluated against subjective intelligibility scores [91]. A similar approach, the Twin HMM Based STOI (THMMB-STOI) measure [57], attempts to estimate the clean speech signal from the degraded speech signal using a speech synthesizer based on a twin HMM and use this as input to an intrusive intelligibility measure. The THMMB-STOI measure has been shown to provide a prediction accuracy comparable to STOI. However, it is also difficult to generalize, since a speech synthesizer requires a large amount of training data. Another machine learning-based based approach uses a Convolutional Neural Network (CNN) to predict speech intelligibility [3]. Usually, neural networks require large amounts of training data, which poses a limitation for this approach due to the lack of large databases of degraded speech signals with measured intelligibility scores. However, the CNN-based intelligibility measure is based on the assumption that speech intelligibility can be predicted from a rather small number of spectro-temporal patterns in the degraded signal, which can be estimated with a relatively small CNN structure [3]. The measure shows a high correlation with the subjective intelligibility scores for the tested conditions – even outperforming some of the intrusive intelligibility measures [3]. However, the measure is tested with a relatively limited number of noise conditions and it is uncertain how well it can be generalized to other conditions given the relatively small amount of training material.

5 Application to Hearing Aids

A reliable intelligibility measure can play a key role in the development and online processing of hearing aids. During development of hearing aids it is crucial to test the performance of new signal processing algorithms in order to assure they are behaving as expected. Traditionally, subjective listening experiments have been performed to evaluate the performance of the hearing aids. However, as already mentioned in Section 3, despite listening experiments being more reliable, they are also more expensive and time-consuming, which can pose a limitation for how thoroughly the algorithms can be evaluated. As such, it can be highly relevant to replace part of the listening experiments during development with an objective intelligibility measure. Generally, the predicted outcomes of the listening experiments can either be based on recorded speech signals from the environment in which the subjective experiment would have taken place, e.g. on a KEMAR, as in Paper A or

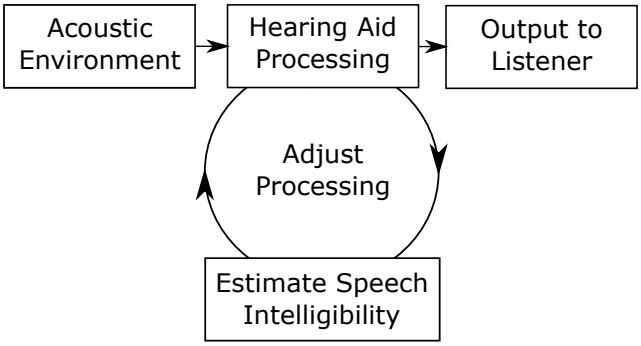


Fig. 11: The speech intelligibility of the combined acoustic environment and hearing aid processing is predicted in real-time in order to optimize the processing of the hearing aid, which would requires a non-intrusive speech intelligibility prediction algorithm.

based on simulated signals, e.g. in McRoomSim [102] as in Paper B. In either case, it is usually possible to have access to the clean speech signal during the development of new signal processing algorithms such that an intrusive measure is sufficient.

Contrarily, if an objective measure should be used to predict intelligibility on a hearing aid during online processing in order to optimize the operation of the system, it is necessary with a non-intrusive speech intelligibility measure. An application during online processing in hearing aids could be to change settings of the algorithms with regards to speech intelligibility as an alternative to environment classification (see Figure 11) [65]. While speech enhancement algorithms can be useful for improving speech intelligibility in adverse listening scenarios with high noise levels, the same algorithms can affect the quality of the speech negatively in less noisy conditions [70]. Therefore, it could be beneficial for hearing aid users if the hearing aid is designed to limit speech enhancement processing to scenarios in which it provides an actual increase in speech intelligibility and otherwise leave the signal unaltered in order to preserve the quality of the speech in quieter conditions. As such, fast and robust real-time prediction of speech intelligibility could detect whether the speech enhancement algorithms increase the speech intelligibility or not in order to weigh the trade-off between quality and intelligibility and adjust the hearing aid accordingly.

6 Contributions

The main body of this thesis is constituted by a collection of six papers. The primary contribution of the work underlying this thesis consists of the proposal of objective measures to predict speech intelligibility non-intrusively.

They are all based on the approach to extend an existing intrusive speech intelligibility measure such that it can predict speech intelligibility non-intrusively without requiring access to a clean reference signal. The proposed measures are based on the intrusive STOI measure [97], because it has been highly evaluated and proven to correlate well with subjective intelligibility scores across a wide range of conditions and processing types. Furthermore, its simplicity makes it useful for real-time applications, e.g. hearing aids. The principle of all the proposed methods is to replace the clean speech reference signal with an estimation of the relevant features representing the original reference signal. The difference between the proposed methods lies in how the reference signal of the clean signal is estimated using, respectively, the spatial content (papers [A], [B] and [F]), the fundamental frequency content (papers [B] and [F]) or the envelope content (papers [C]-[E]) to reconstruct the clean signal.

[A] Semi-Non-Intrusive Objective Intelligibility Measure using Spatial Filtering in Hearing Aids

In this paper we propose a method for predicting the speech intelligibility with an intrusive intelligibility metric, STOI, without requiring access to a clean reference signal. The reference signal is instead replaced with an estimate of the clean speech signal, which is obtained from a multi-channel signal with spatial filtering using a generalized sidelobe cancellation structure, which is chosen due to it being an already widely applied beamformer for hearing aid applications and its simplicity making it easily implementable in today's hearing aids. The results show that the principle holds for one interfering speech source, where the non-intrusively obtained scores are highly correlated with the intrusively estimated STOI scores. For multiple interfering speech sources the proposed measure correlates well for higher STOI scores but deviates for lower scores (below 0.5). However, STOI scores below this level generally correspond to a very low speech intelligibility, which is most likely not relevant in realistic situations. Thus, depending on the purpose may the functioning range of the proposed measure be adequate.

[B] Pitch-Based Non-Intrusive Objective Intelligibility Prediction

A disadvantage of the previous proposed measure is the limited ability to differentiate between multiple speech sources due to being purely based on the spatial content with a limited amount of microphones available in a haring aid setup. In order to overcome this limitation, this paper proposes to estimate the reference signal using a multi-channel spatio-temporal harmonic model dubbed Pitch-Based STOI (PB-STOI). Combining spatial and temporal cues (the location and fundamental frequency of the desired speaker) can

improve the ability to differentiate the desired speech signal from competing speakers, since also considering the voice of the individual speaker and not only the spatial content helps to resolve ambiguities. The PB-STOI measure is shown to capture the features of the original clean signal relatively well and to be well-correlated with the intrusive STOI scores for both stationary noise, a complex setup with multiple competing speakers and low reverberation.

[C] Non-Intrusive Intelligibility Prediction using a Codebook-Based Approach

This paper proposes a Non-Intrusive Codebook-Based STOI (NIC-STOI), which allows using STOI to predict the intelligibility of the noisy signal without requiring access to the clean reference signal by replacing it with an estimate of the clean speech envelope spectrum. The spectral envelope of the clean speech is estimated from its degraded version by identifying combinations of pre-trained dictionaries, i.e. codebooks of clean speech and noise spectra, parametrized by auto-regressive parameters, which best fit the data. In contrast to the previously proposed measures, the NIC-STOI measure is a single-channel solution for estimating speech intelligibility, which has the advantage that it doesn't require access to multiple microphones but can't differentiate between competing speakers based on the spatial configuration. On the other hand, NIC-STOI has the possibility to have the predefined codebook for the reference signal trained for the desired speaker such that it can use this information to differentiate the reference signal from competing speakers. In this paper, the NIC-STOI measure is evaluated objectively with the intrusive STOI as the ground truth and the results show a high correlation between the proposed NIC-STOI measure and the intrusive STOI measure.

[D] Non-Intrusive Codebook-Based Intelligibility Prediction

In this paper, the NIC-STOI measure proposed in Paper C is further investigated on a larger test set. Firstly, it is investigated how using gender specific codebooks affect the prediction accuracy compared to a generic codebook. The results show that the performance does not degrade for the case with a generic codebook compared to the gender specific codebook. Secondly, the NIC-STOI measure is evaluated against subjective listening data in which it is shown to have a high correlation with intelligibility score for additive babble noise.

[E] Validation of The Non-Intrusive Codebook-Based Short Time Objective Intelligibility Metric for Processed Speech

In this paper, the application of NIC-STOI measure is further validated against subjective listening scores for conditions with non-linearly processed speech across a wide range of noise conditions. Even though NIC-STOI is not expected to be able to account for non-linear processing, since it is based on an additive noise model, the results show a high correlation with subjective listening scores for Ideal Time-Frequency Segregation data. The NIC-STOI measure outperforms the three existing state-of-the-art non-intrusive speech intelligibility measures it is compared with and is almost on par with the performance of the intrusive STOI measure.

[F] Harmonic Beamformers for Non-Intrusive Speeh Intelligibility Prediction

This paper combines the principles in Paper A and B such that the reference signal is obtained using model-based harmonic spatial filtering, which exploits the simplicity of the beamforming approach while preserving the robustness of the spatio-temporal model, since taking both the spatial and spectral content into account resolves possible ambiguities due to competing speakers or reverberation. The model-based harmonic beamformer consists of a spatial filters, which are optimized to the spatial and spectral characteristics of the desired speech signal. However, using a harmonic model to estimate the reference signal only captures the voiced segments of the speech. This might be sufficient for capturing the relevant features of the reference signal in noisy conditions, since it is that the voiced parts of the speech that contains the most energetic spectro-temporal regions and intelligibility has been shown to be highly related to the presence of such glimpses [22]. The proposed measure is in the simulated results shown to be well correlated with the intrusively computed scores in an adverse listening scenario with multiple competing speakers and different noise and reverberation levels.

7 Conclusion

The main outcome of the work performed during this thesis is the proposal of different approaches for non-intrusive prediction of speech intelligibility, i.e. predicting how intelligible speech is without access to a clean reference signal. The direction taken in this work is to predict the speech intelligibility non-intrusively by first obtaining an estimate of the clean reference signal, which is thereafter used as input to an intrusive speech intelligibility measure. The difference between the proposed non-intrusive speech intelligibility prediction measures lies in how the reference signal is estimated.

The proposed methods can generally be divided into two different approaches to the problem; The multi-channel case (Papers A, B and F), where the proposed measures utilize the spatial and/or spectral content to extract an estimate of the desired reference signal, and the single-channel case (Papers C-E), where the desired reference signal is estimated through a combination of pre-defined dictionaries of speech and noise spectra, parametrized by auto-regressive parameters, which through a model of the speech production system models the envelope of the signal's spectrum.

The advantage of the multichannel approach is that using the spatial information can help to resolve ambiguities. Furthermore, the simplicity of these approaches is important for some applications such as hearing aids. On the other hand, the single-channel approach is a much more difficult task than the multichannel problem, which also increases the complexity of the proposed measure but is advantageous in cases, where it is not possible to have access to multiple microphones. It should be noted that both approaches achieve a high correlation with their intrusive counterpart and as such the choice of measure primarily depends on the purpose of the speech intelligibility prediction task at hand.

In the work underlying this thesis it has been investigated to use the Short-Time Objective Intelligibility (STOI) as the intrusive framework to which the estimated reference signals are given as input. The rationale behind choosing this measure is due to its simplicity while it still has been shown to have a high performance for a large number of noise conditions and processing types. However, it should be noted that it is the front-end of the proposed measures that forms the basis of the present work and as such the estimated reference signals could also be used for other intrusive speech intelligibility prediction measures given that they are based on the type of features that the different approaches extract.

In future work, it could be interesting to look into other intrusive metrics as back-end in the non-intrusive intelligibility metric for conditions, where STOI is expected to fail. For example, the performance of STOI is poor in highly fluctuating noise conditions, i.e., modulated noise, due to the long analysis window length. In this case, it could be relevant to replace STOI with an intrusive metric, which is known to perform well in such a condition, e.g., ESTOI has been developed as an extension of STOI to work well for modulated noise. Similarly, STOI is not expected to work well in highly reverberant conditions in which in could be interesting to investigate the possibility for using intrusive metrics that are more suitable for such conditions such as HASPI and mr-sEPSM.

Although the proposal of the non-intrusive metrics in the work underlying this thesis has narrowed the gap for the usability of such metrics for predicting speech intelligibility in real-time during usage on, e.g., a hearing aid, it is also important with a more extensive and exhaustive evaluation of

these metrics in order to gain the necessary trust in the application of these. Therefore, future work should include more comprehensive experiments in different conditions and realistic scenarios as well as more extensive subjective listening tests.

References

[1] J. B. Allen, "How do humans process and recognize speech?" *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct 1994.

[2] ——, "Harvey Fletcher's role in the creation of communication acoustics," *J. Acoust. Soc. Am.*, vol. 99, no. 4, pp. 1825–1839, 1996.

[3] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 26, no. 10, pp. 1925–1939, 2018.

[4] A. H. Andersen, "Speech intelligibility prediction for hearing aid systems," Ph.D. dissertation, Aalborg University, 2017, industrial PhD Supervisors: Prof. Jesper Jensen, Oticon A/S and Aalborg University Tekn. Dr. Jan Mark de Haan, Oticon A/S University PhD Supervisor: Prof. Zheng-Hua Tan, Aalborg University.

[5] A. H. Andersen, J. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *ICASSP*, March 2017, pp. 5085–5089.

[6] ANSI S3.5, 1969, *Methods for the Calculation of the Articulation Index*, American National Standards Institute, New York, USA Std., 1969.

[7] ANSI S3.5, 1997, *Methods for the Calculation of the Speech Intelligibility Index*, American National Standards Institute, New York, USA Std., 1997.

[8] C. Binns and J. F. Culling, "The role of fundamental frequency contours in the perception of speech against interfering speech," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1765–1776, 2007.

[9] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *17th European Signal Processing Conference*, 2009, pp. 1849–1853.

[10] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2801–2810, 2002.

[11] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, pp. 117–128, 01 2000.

[12] ——, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, perception & psychophysics*, vol. 77, no. 5, p. 1465–1487, Jul. 2015.

[13] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1508–1516, 1988.

[14] C. A. Brown and S. P. Bacon, "Fundamental frequency and speech intelligibility in background noise," *Hearing Research*, vol. 266, no. 1, pp. 52 – 59, 2010, special Issue: Annual Reviews 2010.

[15] G. Carter, C. Knapp, and A. Nuttall, "Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 4, pp. 337–344, 1973.

[16] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *EUROSPEECH*, vol. 1, 18-21 September 1995, pp. 867–870.

[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] E. C. Cherry, "Some experiments on the recognition of speech with one and two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953.

[19] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[20] J. Clark, C. Yallop, and J. Fletcher, *An Introduction to Phonetics and Phonology*, 3rd ed. Blackwell, 2007.

[21] M. Cooke, "Modelling auditory processing and organisation," Ph.D. dissertation, Cambridge University Press, 1993.

[22] ——, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[23] M. Cooke and J. Barker, "An Audio-visual corpus for speech perception and automatic speech recognition (L)," *J. Acoust. Soc. Am.*, vol. 120(5), pp. 2421–2424, Nov. 2006.

[24] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20(5), pp. 1644–1657, 2012.

[25] C. J. Darwin, "Listening to speech in the presence of other sounds," *Philosophical Transactions of the Royal Society B*, vol. 363, p. 1011–1021, 2008.

[26] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2913–2922, 2003.

[27] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. i. model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[28] T. Dau, J. Verhey, and A. Kohlrausch, "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2752–2760, 1999.

[29] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, 1st ed. Wiley-Interscience, 2000.

[30] M. Dorman, P. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2403–2411, 1997.

[31] R. Drullman, "Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level," *The Journal of the Acoustical Society of America*, vol. 98, no. 3, pp. 1796–1798, 1995.

[32] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[33] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech Communication*, vol. 41, pp. 331–348, 2003.

[34] S. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *J. Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1181–1196, Sep. 2000.

[35] T. H. Falk, S. Cosentino, J. Santos, D. Suelzle, and V. Parsa, "Non-intrusive objective speech quality and intelligibility prediction for hearing instruments in complex listening environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7820–7824.

[36] 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 Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[37] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[38] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell System Technical Journal*, vol. 8, no. 4, pp. 806–854, Oct. 1929.

[39] D. Fogerty, "Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 977–988, 2011.

[40] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[41] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[42] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-Time Processing of Speech Signals*. Prentice-Hall, 1987.

[43] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3108–3118, 2004.

[44] J. M. Hillenbrand and T. M. Nearey, "Identification of resynthesized /hvd/ utterances: Effects of formant contour," *Journal of the Acoustical Society of America*, vol. 406, pp. 3509—-3523, 1999.

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

[46] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

[47] ——, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[48] T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics," *Acta Acustica United with Acustica*, vol. 46, no. 1, pp. 60–72, 1980.

[49] 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.

[50] ——, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[51] Y. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 124, no. 2, pp. 1306–1319, 2004.

[52] ITU P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*, International Telecommunication Union, Standard Std., 2004.

[53] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov 2016.

[54] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acustica United with Acta Acustica*, vol. 101, pp. 1016–1025, 2015.

[55] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[56] 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, pp. 436–446, 07 2013.

[57] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-hmm-based non-intrusive speech intelligibility prediction," in *ICASSP*, March 2016, pp. 624–628.

[58] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[59] ——, "The hearing-aid speech perception index (haspi)," *Speech Commun.*, vol. 65, pp. 75 – 93, 2014.

[60] D.-S. Kim and A. Tarraf, "Anique+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell System Technical Journal*, vol. 12, no. 1, pp. 221–236, 2007.

[61] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *The Journal of the Acoustical Society of America*, vol. 15, no. 1, pp. 362–378, 2004.

[62] K. D. Kryter, "Validation of the articulation index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1698–1702, 1962.

[63] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[64] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Brill, 1996.

[65] L. Lamarche, C. Gigure, W. Gueaieb, T. Aboulnasr, and H. Othman, "Adaptive environment classification system for hearing aids," *J. Acoust. Soc. Am.*, vol. 127, no. 5, pp. 3124–3135, 2010.

[66] W. J. Levelt, "Models of word production," *Trends in cognitive sciences*, vol. 3, no. 6, p. 223–232, 1999.

[67] H. Levitt, "Transformed up-down methods in psychoacoustics." *The Journal of the Acoustical Society of America*, vol. 49, no. 2, pp. 467–477, 1971.

[68] 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.

[69] L. Lightburn and M. Brookes, "A weighted stoi intelligibility metric based on mutual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5365–5369.

[70] P. C. Loizou, *Speech Enhancement: Theory and Practice*, ser. Signal processing and communications. Taylor & Francis, 2007.

[71] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.

[72] C. Ludvigsen, C. Elberling, and G. Keidser, "Evaluation of a noise reduction method comparison between observed scores and scores predicted from STI," *Scand. Audiol. Suppl.*, vol. 38, pp. 50–55, 1993.

[73] D. Maurer, *Acoustics of the Vowel*. Bern, Switzerland: Peter Lang, 2016. [Online]. Available: https://www.peterlang.com/view/title/36750

[74] B. T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition depending on speech-intrinsic variability." *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3126–41, 2010.

[75] B. C. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Brill, 2012.

[76] R. W. Peters, B. C. J. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 577–587, 1998.

[77] J. O. Pickles, *An introduction to the physiology of hearing*, 4th ed. Brill, 2013.

[78] C. J. Plack, *The Sense of Hearing*, 2nd ed. Psychology Press, 2014.

[79] H. Relaño-Iborra, T. May, J. Zaar, C. Scheidiger, and T. Dau, "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," *J. Acoust. Soc. Am.*, vol. 140, no. 4, p. 2670–2679, 2016.

[80] 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. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[81] 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," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, 2006.

[82] ——, "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise," *J. Acoust. Soc. Am.*, vol. 126, no. 6, pp. 3236–3245, 2009.

[83] P. Rubin and E. Vatikiotis-Bateson, *Measuring and Modeling Speech Production*. Germany: Springer, Berlin, Heidelberg, 1998.

[84] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, February 1978.

[85] J. F. Santos and T. H. Falk, "Updating the srmr-ci metric for improved intelligibility prediction for cochlear implant users," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2197–2206, Dec 2014.

[86] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *IWAENC*, Sept. 2014, p. 55–59.

[87] A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter, "Objective measures of listening effort: Effects of background noise and noise reduction," *Journal of Speech, Language, and Hearing Research*, vol. 52, p. 1230–1240, Oct. 2009.

[88] 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, pp. 1–8, 2015.

[89] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[90] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, and M. Brookes, "Data driven method for non-intrusive speech intelligibility estimation," in *Proceedings of the 2010th European Signal Processing Conf. (EUSIPCO). Denmark*, 2010.

[91] 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, 2016.

[92] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception." *Nature*, vol. 416, no. 6876, p. 87, 2002.

[93] M. S. Sommers and J. Barcroft, "Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2406–2416, 2006.

[94] C. Sørensen, A. Xenaki, J. Boldt, and M. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *ICASSP*, March 2017, pp. 386–390.

[95] S. Srinivasan and D. Wang, "A model for multitalker speech perception," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3213–3224, 2008.

[96] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[97] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[98] M. Tomasello, *Origins of human communication*. MIT press, 2010.

[99] T. Van den Bogaert, T. Klasen, M. Moonen, L. Van Deun, and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 515–526, 2006.

[100] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 26, no. 11, pp. 2153,2166, 2018-11.

[101] H. von Helmholtz, *Tonempfindungen als physiologische grundlage für die theorie der musik*, 3rd ed. Verlag von Friedrich Vieweg u. Sohn, 1870.

[102] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, 2010, pp. 1–6.

[103] 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, 2003.

[104] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2336–2347, 2007.

[105] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Boston, MA: Springer US, 2005, pp. 181–197.

Part II

Papers

Paper A

Semi-Non-Intrusive Objective Intelligibility Measure using Spatial Filtering in Hearing Aids

Charlotte Sørensen
Jesper Bünsow Boldt
Fredrik Gran
Mads Græsbøll Christensen

The paper has been presented at the
24th European Signal Processing Conference (EUSIPCO), pp. 1358–1362,
Budapest, Hungary, 2016.

The layout has been revised.

Abstract

Reliable non-intrusive online assessment of speech intelligibility can play a key role for the functioning of hearing aids, e.g. as guidance for adjusting the hearing aid settings to the environment. While existing intrusive metrics can provide a precise and reliable measure, the current non-intrusive metrics have not been able to achieve acceptable intelligibility predictions. This paper presents a new semi-non-intrusive intelligibility measure based on an existing intrusive measure, STOI, where an estimate of the clean speech is extracted using spatial filtering in the hearing aid. The results indicate that the STOI score obtained with the proposed method using an estimate of the clean speech correlates well with the STOI score having the original clean speech signal available.

1 Introduction

For users of hearing aids speech intelligibility depends highly on the specific listening environment. One of the main issues is significantly decreased speech intelligibility in noisy multi-talker environments termed the "cocktail party problem" [1, 2]. Therefore, a lot of research has gone into the development of various speech enhancement algorithms (e.g., noise and echo suppression) to overcome this challenge. However, noise suppression techniques, such as adaptive directional filtering, can have a negative impact on localization performance of hearing aid users [3]. The fact that hearing aid users receive distorted localization cues can lead to decreased intelligibility due to losing a binaural advantage of 3-12 dB [3, 4]. As such, it is important to quantify, whether the gain from the noise suppression techniques are advantageous if localization cues are lost in return by assessing the intelligibility of the current environment. For the users of assistive listening devices it would be a great benefit, if the devices were able to automatically detect when advanced speech enhancement actually provides an improvement and adjust the hearing aid settings accordingly. Generally, the remaining hearing of the hearing aid user should be relied on as much as possible such that speech enhancement processing is limited to when it provides a benefit and the proposed method could facilitate exactly this. Fast and robust online evaluation of the listening environment could assure that speech enhancement processing is only applied when necessary and selected without requiring an action of the hearing aid user [5, 6]. As such, the proposed method can be seen as an alternative to environment classification based on intelligibility rather than classifying the different environments [7].

Thus, it would be preferable if objective intelligibility measures could become a crucial part of the online processing of assistive listening devices. Intrusive objective measures (e.g., the short-time objective intelligibility (STOI)

metric [8], the normalized covariance metric (NCM) [9]) with access to both the clean and noisy speech can generally provide a precise and reliable measure for the speech intelligibility [6]. However, online processing in a hearing aid requires a non-intrusive objective measure, since access to the clean speech is rarely available. Over the years a number of non-intrusive metrics have been developed (e.g., the modulation spectrum area (ModA) [10], the speech-to-reverberation modulation energy ratio (SRMR) [11]). However, according to a recent comprehensive review none of the tested the existing non-intrusive measures have achieved acceptable results [6].

This paper is concerned with a method in between the intrusive and non-intrusive technique that can be processed online in a hearing aid while taking advantage of the reliability of existing intrusive metrics. The approach is to extract an estimate of the clean speech with directional spatial filtering in the hearing aid and use this in existing intrusive objective intelligibility metrics. In other words, an estimate of the intelligibility is obtained by comparing the output of a beamformer at the direction of the desired talker with the output of an omnidirectional microphone using an existing objective measure such as STOI. The online processed intelligibility prediction of the specific environment can then be used to determine, whether the intelligibility is below a certain threshold and apply speech enhancement processing when it is beneficial.

2 Method

In this section the approach and method behind the proposed semi-non-intrusive objective intelligibility measure is presented. A block diagram incorporating the whole semi-non-intrusive objective intelligibility measure with both the beamformer and the existing intrusive intelligibility measure STOI is shown in Figure A.1. The principles behind the beamforming structure and notation are explained in Section 2.1. The STOI metric gives a prediction, $\mathrm{d}(t)$, of the speech intelligibility on a 0-1 scale by comparing the correlation of a clean and degraded version of the same speech signal [8]. As illustrated in the diagram the noisy signal from an omnidirectional microphone is both used as the degraded speech input to STOI as well as reference of the source to an adaptive noise cancellation (ANC) stage in the beamformer. The remaining microphone signals are used in a fixed spatial filtering stage in the beamformer to extract a reference of the interference.

2.1 Generalized sidelobe cancellation

An estimate of the clean speech is obtained using a widely applied beamformer for hearing aid applications based on the generalized sidelobe can-

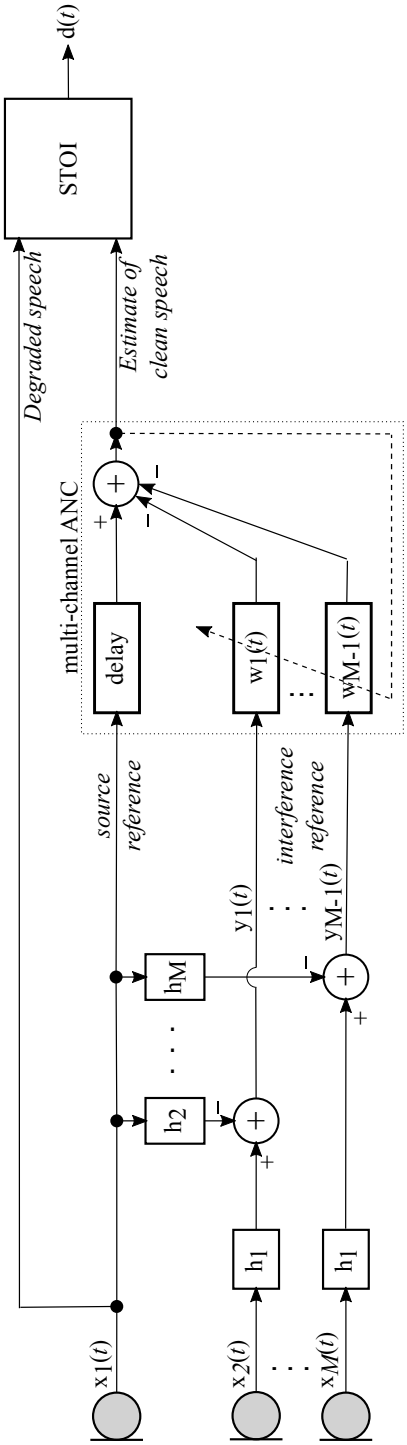


Fig. A.1: Block diagram of the proposed semi-non-intrusive objective intelligibility measure in which an estimate of the clean speech is extracted with the GSC structure and compared with the output of an omnidirectional microphone using STOI.

cellation (GSC) structure [12–14]. The beamformer has four microphones by exploiting the front and rear microphone of two BTE hearing aids assuming a bilateral wireless link between them. The implemented GSC structure consists of a fixed spatial preprocessor and an ANC unit similar to the approach of [14] extended to four microphones as illustrated in Figure A.1 with $M = 4$.

It is assumed that each microphone signal x_k, $k = 1, \ldots, M$ is the desired source additively interfered with a number of interferers, N i.e.

$$x_k(t) = h_k * s(t) + \sum_{n=1}^{N} h_{k,n}^{\text{interf}} * s_n^{\text{interf}}(t) \tag{A.1}$$

where h_k and $h_{k,n}^{\text{interf}}$ are the acoustic impulse responses between the kth microphone and the desired source, $s(t)$, and interferers, $s_n^{\text{interf}}(t)$, respectively and $*$ denotes convolution. Ambient noise can be created by adding up multiple interferers with reverberation included in the acoustic impulse responses.

During periods of interference-only, $s(t) = 0$, each microphone signal is the sum of the interferers convolved with the acoustic impulse response between each interferer and the kth microphone, i.e.

$$x_k(t) = \sum_{n=1}^{N} h_{k,n}^{\text{interf}} * s_n^{\text{interf}}(t) \tag{A.2}$$

A reference of the interference is created by steering a zero towards the direction of the desired speaker. The location of the desired speaker is assumed to be in the front of the listener at zero degrees but can easily be relaxed to other positions. The desired source is canceled using spatial filters, which give an estimate of interference-only at the kth microphone for $k = 2, \ldots, M$, where h_1 is the acoustic impulse response between the desired source at 0° and the first microphone:

$$\begin{aligned} y_{k-1}(t) &= x_k(t) * h_1 - x_1(t) * h_k \\ &= h_k * s(t) * h_1 + \sum_{n=1}^{N} h_{k,n}^{\text{interf}} * s_n^{\text{interf}}(t) * h_1 \\ &- (h_1 * s(t) * h_k + \sum_{n=1}^{N} h_{1,n}^{\text{interf}} * s_n^{\text{interf}}(t) * h_k) \\ &= \sum_{n=1}^{N} h_{k,n}^{\text{interf}} * s_n^{\text{interf}}(t) * h_1 - \sum_{n=1}^{N} h_{1,n}^{\text{interf}} * s_n^{\text{interf}}(t) * h_k \end{aligned} \tag{A.3}$$

where y_{k-1}, $k = 2, \ldots, M$ is the interference reference at the kth microphone. It can be seen that the filters block out $s(t)$ in the derivation of y_{k-1}. The coefficients of the blocking filters have been determined based on the impulse responses between the source at 0° and the kth microphone measured on a KEMAR artificial head and torso as described in Section 3.

The ANC unit attenuates the interference in the desired source reference that is correlated with the interference reference using the filters $\mathbf{w}_k(t) = [w_{k,1}(t), w_{k,2}(t), \ldots, w_{k,L}(t)]$, where L is the length of the filter. The ANC unit is updated with a least squares (LS) approach but can in online processing easily be implemented as a least mean square (LMS) algorithm.

The incorporation of the fixed spatial filter in the preprocessor reduces the amount of speech leakage into the interference reference but cannot completely prevent it [13, 15]. Therefore, the ANC is adapted during periods of interference-only in order to avoid possible cancellation of the desired speech source. For this purpose a robust speech detector is assumed available in this paper.

3 Experimental methodology

The acoustic impulse responses have been measured using the front and rear microphones on a GN ReSound Alera 312 BTE hearing aid on a KEMAR artificial head and torso in an anechoic room with a maximum length sequence (MLS) with a code length of 11 and averaged over 30 repetitions. The KEMAR artificial head and torso was rotated in the horizontal plane with a resolution of 2 degrees using a Brüel & Kjær Turntable system type 9640.

The speech samples of both the desired source and the interferers were taken from the EUROM_1 database as 5 second recordings of the English sentence corpus [16]. The level of the interferers were varied according to the level of the desired speech source as the source-to-interference ratio (SIR) [17]. The clean speech of the desired source was convolved with the acoustic impulse responses from 0° to each each microphone and the interfering speech sources were convolved with the impulse responses from 140°, 270°, 50° and 300° for one, two, three or four speakers, respectively. Compared to current state-of-the-art studies four interferers can be considered a relatively complex scenario with speech-on-speech masking being a difficult task [1, 2, 6].

4 Results

The performance of the proposed semi-non-intrusive objective intelligibility measure is evaluated by comparing the STOI score of the noisy speech obtained using the estimate of the clean speech as reference with the STOI score obtained using the original clean speech as reference. Figure A.2-A.5 show the STOI scores as function of SIR for one, two, three and four interferers, respectively. For one interferer located at 140° (Figure A.2) it can be seen that the STOI score obtained using the output from the implemented GSC beamformer as reference (dashed line) correlates well with the STOI scores

obtained with access to the original clean speech signal (solid line) for all SIRs.

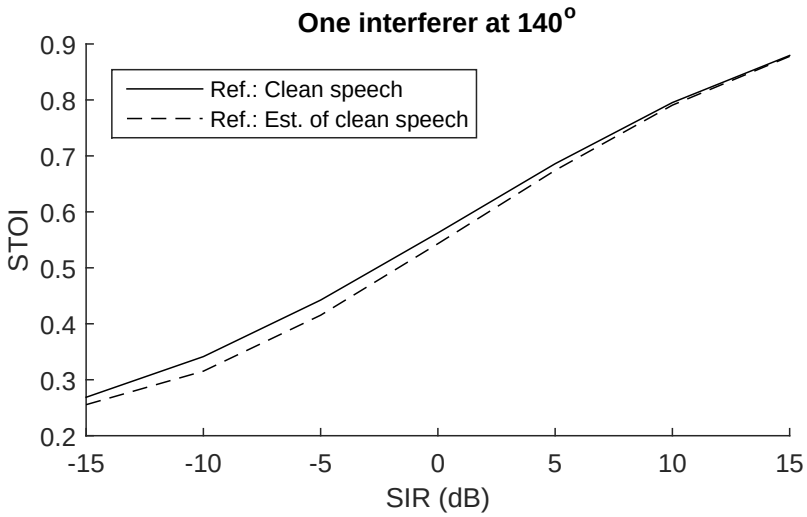


Fig. A.2: STOI score as function of SIR with one interferer at 140° using the clean speech signal (solid line) and the estimate of the clean speech extracted with the implemented 4 microphone GSC beamformer (dashed line) as reference.

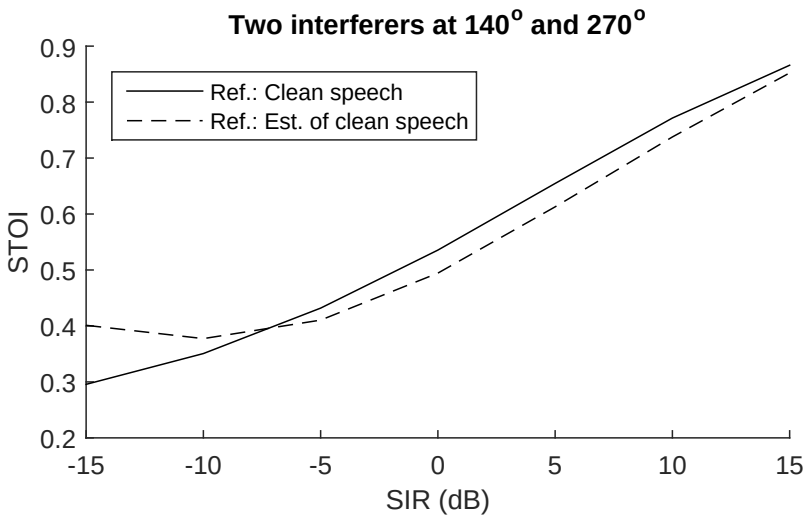


Fig. A.3: STOI score as function of SIR with two interferers at 140° and 270° using the clean speech signal (solid line) and the estimate of the clean speech extracted with the implemented 4 microphone GSC beamformer (dashed line) as reference.

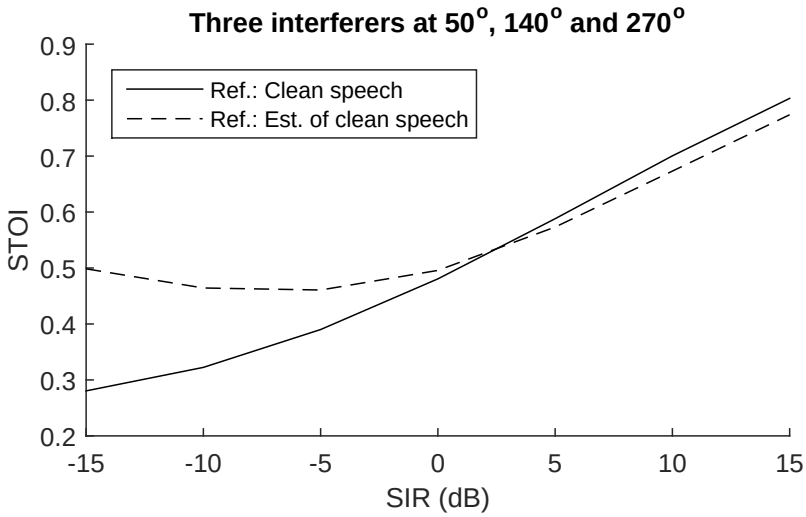


Fig. A.4: STOI score as function of SIR with three interferers at 50°, 140° and 270° using the clean speech signal (solid line) and the estimate of the clean speech extracted with the implemented 4 microphone GSC beamformer (dashed line) as reference.

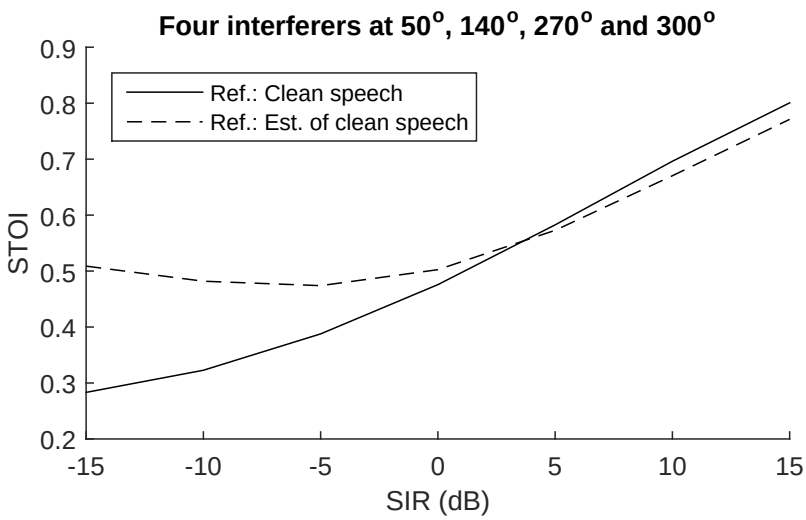


Fig. A.5: STOI score as function of SIR with four interferers at 50°, 140°, 270° and 300° using the clean speech signal (solid line) and the estimate of the clean speech extracted with the GSC beamformer (dashed line) as reference.

In the case of two interferers located at 140° and 270° (Figure A.3) the STOI score using the beamformed estimate of the clean speech as reference correlates well with the intrusive STOI score having access to the clean speech signal for STOI scores and SIRs higher than 0.4 and -5 dB, respectively. However, below this level the output from STOI using the estimate of the clean speech as reference starts to deviate from the STOI score obtained using the original clean speech as reference. In the cases of three interferers at 50°, 140° and 270° (Figure A.4) and four interferers at 50°, 140°, 270° and 300° (Figure A.5) the STOI scores with the estimate of the clean speech and the original clean speech as references, respectively, correlates well for STOI scores above 0.5 and SIRs above 5 dB but deviates below these levels. Noteworthy, the performance of the proposed method does not decrease substantially going from the case with three interferers to four interferers.

5 Discussion

A reliable objective intelligibility measure in the online processing of hearing aids could be of great advantage to predict whether speech enhancement would provide a benefit for the user and adjust the hearing aid settings accordingly. Online processing would require a non-intrusive metric and even though a number of promising non-intrusive measures have been developed over the recent years none of them have achieved sufficient results for the purpose [6, 10, 11]. Previous studies have shown that STOI scores correlate well with subjective intelligibility scores and thus gives a reliable estimate for the speech intelligibility [6, 8]. As such, a non-intrusive measure performing similarly to STOI could yield a promising method for online processing of speech intelligibility in hearing aids. The intelligibility scores obtained with the proposed semi-non-intrusive technique correlates well with the intrusive STOI scores obtained with access to the clean speech for STOI scores above 0.5 but deviates for lower scores. This may or may not be a problem for the intended purpose provided it reflects so little speech intelligibility that it conforms to the threshold for applying speech enhancement anyway. A STOI score below 0.6 may correspond to very low speech intelligibility depending on the speech material and the psychometric function relating STOI scores to subjective scores [6]. Furthermore, the proposed method could easily be implemented in today's hearing aids. The acoustic impulse responses used for the spatial filter design in the blocking matrix could either be the standard acoustic impulse responses measured on KEMAR or personalized acoustic impulse responses measured during adjustment of the hearing aid.

In future work it could be interesting to test the proposed method with added reverberation as this is known to affect the performance of the GSC beamformer [13, 14]. In order to properly simulate reverberation 3 dimen-

sional acoustic room impulses would be required. Additionally, the objective intelligibility scores obtained with the proposed semi-non-intrusive technique could be tested against subjective listening tests in future work. In a similar manner to using the proposed method for prediction of the speech intelligibility the same approach could be used to evaluate speech quality with e.g. the perceptual evaluation of speech quality (PESQ [18]) by using the estimate of the clean speech to evaluate the speech quality before and after speech processing in the hearing aid. Furthermore, the proposed method could also be extended to include personalized hearing losses in the speech intelligibility prediction similarly to the technique in e.g. the hearing-aid speech perception index (HASPI) [19].

Recently, binaural speech intelligibility methods have with limited success attempted to predict the speech intelligibility by including the effects of spatial masking [20]. The proposed technique in this paper does not take advantage of the multiple channels used in the beamformer to predict the effects of spatial masking on the speech intelligibility. In future work this could be an interesting extension of the proposed technique.

6 Conclusion

This paper has presented a new feasible technique for online processing of speech intelligibility in hearing aids. The technique is based on an existing intrusive objective metric, where an estimate of the clean speech to be used as reference is obtained using a GSC structure with spatial filters as blocking matrix. The GSC structure is implemented using the front and rear microphones on two wirelessly linked BTE hearing aids. The results indicate that the obtained STOI scores using the estimate of the clean speech as reference correlate well with the intrusive STOI having access to the original clean speech for STOI scores above 0.5. Thus, the proposed method yields a promising and feasible technique for online processing of speech intelligibility in hearing aids.

References

[1] R. W. Peters, B. C. J. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 577–587, 1998.

[2] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1725–1736, 1990.

[3] T. Van den Bogaert, T. Klasen, M. Moonen, L. Van Deun, and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 515–526, 2006.

[4] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferera," *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 833–843, 2006.

[5] V. Hamacher, J. Chalupper, E. Eggers, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Applied Signal Process.*, vol. 18, pp. 2915–2929, 2005.

[6] 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 Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[7] L. Lamarche, C. Gigure, W. Gueaieb, T. Aboulnasr, and H. Othman, "Adaptive environment classification system for hearing aids," *J. Acoust. Soc. Am.*, vol. 127, no. 5, pp. 3124–3135, 2010.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[10] 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.

[11] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[12] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[13] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 487–503, 2005.

[14] A. Spriet, L. Van Deun, K. Eftaxiadis, J. Laneau, M. Moonen, B. van Dijk, A. van Wieringen, and J. Wouters, "Speech understanding in background noise with the two-microphone adaptive beamformer beam in the nucleus freedom cochlear implant system," *Ear & Hearing*, vol. 28, pp. 62–71, 2007.

[15] S. Nordebo, I. Claesson, and S. Nordholm, "Adaptive beamforming: Spatial filter designed blocking matrix," *IEEE J. Ocean. Eng.*, vol. 19, no. 14, pp. 583–590, 1994.

[16] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *EUROSPEECH*, vol. 1, 18-21 September 1995, pp. 867–870.

[17] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] 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 *Proc IEEE Int Conf Acoust Speech Signal Process*, vol. 2, 2001, pp. 749–752.

[19] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi)," *Speech Commun.*, vol. 65, pp. 75 – 93, 2014.

[20] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension, and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, 2010.

Paper B

Pitch-Based Non-Intrusive Objective Intelligibility Prediction

Charlotte Sørensen
Angeliki Xenaki
Jesper Bünsow Boldt
Mads Græsbøll Christensen

The paper has been presented at the
42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 386–390, New Orleans, United States, 2017.

The layout has been revised.

Abstract

Automatic adjustment of the hearing aid according to the intelligibility for the user in the environment could be beneficial. While most intelligibility metrics require a clean speech reference, i.e. intrusive methods, this is rarely available in real-life. This paper proposes a non-intrusive intelligibility metric in which a reconstruction of the clean speech is used in the established intrusive short-time objective intelligibility (STOI) metric. The reconstruction of the clean speech is based on pitch-features of the desired source using a spatio-temporal harmonic model. This model takes advantage of both the spatial and spectral separation of the desired source and interferers to reconstruct the clean signal. The simulations show a high correlation between the proposed pitch-based STOI (PB-STOI) and the original intrusive STOI and hence is promising for online processing of intelligibility.

1 Introduction

One of the main issues encountered by hearing aid (HA) users is severely degraded speech intelligibility in noisy multi-talker environments such as the "cocktail party problem" [1, 2]. Generally, the speech intelligibility for users of assistive listening devices depends highly on the specific listening environment. As such, additional speech enhancement processing may be beneficial in some listening environments whereas the exact same algorithms can have a negative impact on the quality and intelligibility in other listening environments [3, 4]. In HA technology, automatic intelligibility assessment of the listening environment would be beneficial for the user such that speech enhancement is only applied when necessary [5, 6]. This could be facilitated by an online intelligibility evaluation of the listening environment. Thus, it could be beneficial if objective intelligibility metrics could be used in the online processing of HAs.

There are various intrusive methods to predict the speech intelligibility with acceptable reliability such as the short-time objective intelligibility (STOI) metric [7] and the normalized covariance metric (NCM) [8]. However, these methods are intrusive, i.e., they all require access to the clean-speech reference which is rarely available in real-life. A number of non-intrusive methods have been introduced that do not require access to the clean speech signal, e.g. the modulation spectrum area (ModA) [9] or the speech-to-reverberation modulation energy ratio (SRMR) [10]. However, both of these non-intrusive measures are limited to the assessment of reverberated speech signals and are still inferior to the intrusive measures according to a recent review [6].

This paper proposes a method that non-intrusively estimates the speech intelligibility in the listening environment for HAs. Similar to the approaches

in [11, 12] a prediction of the speech intelligibility is obtained by comparing a reconstruction of the clean speech with the noisy speech using an established and reliable intrusive framework, e.g. STOI [6, 13]. The clean speech is obtained by estimating relevant signal features assuming the desired source consists of a number of narrowband signals with harmonically related carrier frequencies using a spatio-temporal model. Combining spatial (i.e. direction of arrival) and temporal (i.e. pitch) cues improves the accuracy of the reconstruction as it resolves ambiguities, e.g. due to reverberation or competing speakers. The proposed method can then potentially be used as an alternative to environment classification by determining, whether the intelligibility is below a certain threshold [14].

2 Method

In this section the approach behind the PB-STOI metric is presented. A block diagram incorporating the framework is shown in Fig. B.1. In the first step, the sound field is recorded with a microphone array. Then, the pitch of the desired speech signal is estimated and the speech is reconstructed using the pitch and direction of arrival of the desired speech signal. Finally, a non-intrusive prediction, $\mathrm{d}(n)$, is given on a 0-1 scale by comparing the correlation of the reconstructed clean speech with the noisy version using the intrusive STOI framework.

2.1 Signal model

A multi-channel spatio-temporal harmonic model is applied based on the model from [15] in order to reconstruct the clean speech signal as input to the intrusive intelligibility metric. In the proposed method it is assumed that K microphones are used to obtain the desired signal added to a mixture of interfering sources and background noise for a frame length of N such for the k'th microphone, the data vector $\mathbf{x}_k = [x_k(0) \;\; x_k(1) \;\; \ldots \;\; x_k(N-1)]^T$ for $k = 0, \ldots, K-1$. The desired source is assumed to be periodic, which is an appropriate assumption for short segments of voiced speech [16]. As such, the data vector $\mathbf{x}_k$ can be modeled as

$$\mathbf{x}_k = \beta_k \mathbf{Z}\mathbf{D}(k)\boldsymbol{\alpha} + \mathbf{e}_k, \tag{B.1}$$

with $\mathbf{Z} = [\mathbf{z}(\omega_0) \ldots \mathbf{z}(L\omega_0)], \mathbf{z}(l\omega_0) = [1 \ldots e^{jl\omega_0(N-1)}]$ for $n = 0, \ldots, N-1$, $\mathbf{D}(k) = \mathrm{diag}([e^{-j\omega_0 f_s \tau_k} \ldots e^{-jL\omega_0 f_s \tau_k}])$ for $l = 1, \ldots, L$ with all other entries equal to zero and $\mathbf{e}_k$ is the sum of the recorded noise and interference. Furthermore, ω_0 is the fundamental frequency, f_s is the sampling frequency and τ_k is the delay of the desired target source between microphone 0 and the k'th microphone giving the direction of arrival (DOA). Moreover, β_k is the

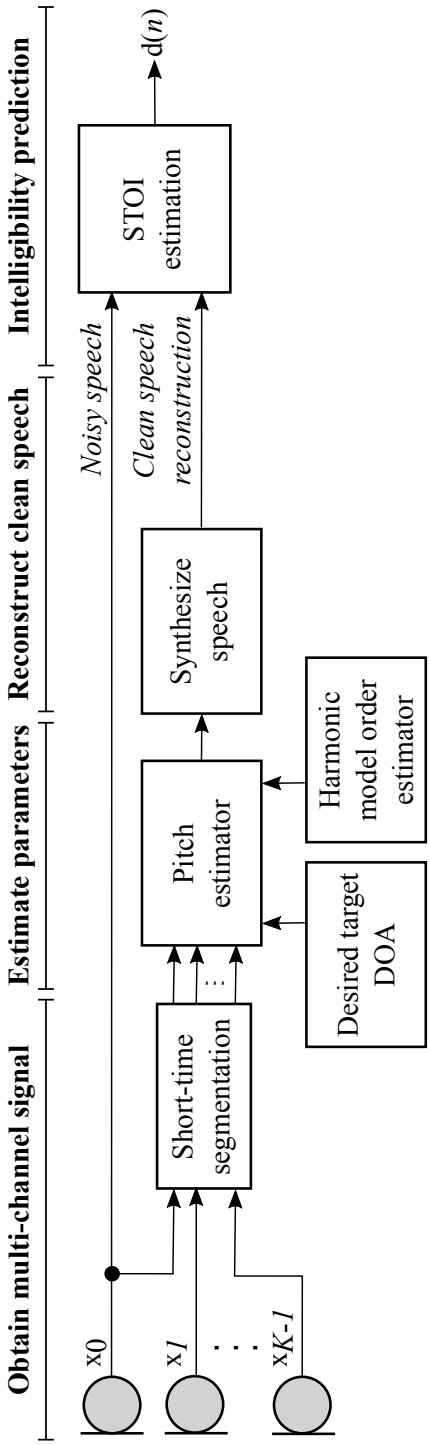


Fig. B.1: Block diagram of the proposed pitch-based non-intrusive objective intelligibility measure in which reconstruction of the clean speech is obtained using the estimated pitch and compared with the output of an omnidirectional microphone using the original intrusive STOI.

attenuation of the desired source at the k'th microphone, $\boldsymbol{\alpha} = [\alpha_1 \dots \alpha_L]^T$ is the complex amplitudes given by $\alpha_l = A_l e^{j\phi_l}$, L is the number of harmonics, $A_l > 0$ and ϕ_l are the real amplitude and phase of the l'th harmonic, respectively.

2.2 Pitch-based intelligibility prediction

The pitch of the desired target source is found by exploiting the spatio-temporal harmonic model structure of the multi-channel signal using the joint pitch and DOA estimation method presented in [15]. In the following, the basic principles and deviations from the original method are explained.

Assuming the noise is uncorrelated white Gaussian with variance σ_k^2 in each channel, the log-likelihood function of the complex data vector $\mathbf{x}_k$ can be written as [15]

$$\ln p(\mathbf{x}_k; \psi) = \\ - NK \ln \pi - N \sum_{k=0}^{K-1} \ln \sigma_k^2 - \sum_{k=0}^{K-1} \frac{\|\mathbf{e}_k\|^2}{\sigma_k^2} \tag{B.2}$$

with the vector ψ containing the signal parameters for $\mathbf{x}_k$. Even though this assumption may seem unreasonable the white Gaussian noise distribution maximizes the entropy of the noise and is a good choice for the noise probability density function [15]. Then, the pitch can be estimated by maximizing the log-likelihood function by differentiating with respect to the amplitudes, $\hat{\boldsymbol{\alpha}}$, the attenuation factor, β_k, and the noise variance, σ_k^2, respectively. As mentioned in [15] these parameters are dependent on each other and are therefore estimated by initially setting the β_k's and σ_k^2's to 1 and iterating over the expressions in Equation (B.3), (B.4) and (B.5). The estimated complex amplitudes are given by

$$\hat{\boldsymbol{\alpha}} = \left[\sum_{k=0}^{K-1} \frac{\beta_k^2}{\sigma_k^2} \mathbf{D}^H(k) \mathbf{Z}^H \mathbf{Z} \mathbf{D}(k) \right]^{-1} \sum_{k=0}^{K-1} \frac{\beta_k}{\sigma_k^2} \mathbf{D}^H(k) \mathbf{Z}^H \mathbf{x}_k \tag{B.3}$$

The estimated attenuation of the desired source at the k'th microphone can be obtained as

$$\hat{\beta}_k = \frac{\mathrm{Re}\{\boldsymbol{\alpha}^H \mathbf{D}^H(k) \mathbf{Z}^H \mathbf{x}_k\}}{\boldsymbol{\alpha}^H \mathbf{D}^H(k) \mathbf{Z}^H \mathbf{Z} \mathbf{D}(k) \boldsymbol{\alpha}} \tag{B.4}$$

Moreover, the noise variance can be found as

$$\hat{\sigma}_k^2 = N^{-1} \|\hat{\mathbf{e}}_k\|^2, \tag{B.5}$$

where $\hat{\mathbf{e}}_k = \mathbf{x}_k - \beta_k \mathbf{Z} \mathbf{D}(k) \boldsymbol{\alpha}$. The maximum likelihood estimator of the pitch can then be written as

$$\hat{\omega}_0 = \arg \min_{\omega_0 \in \Omega_0} \sum_{k=0}^{K-1} \ln \|\mathbf{x}_k - \hat{\beta}_k \mathbf{Z} \mathbf{D}(k) \hat{\boldsymbol{\alpha}}\|^2 \tag{B.6}$$

where Ω_0 is a set of possible pitch candidates. Contrary to the original method in [15], the DOA of the desired target source is assumed known such that the problem reduces to spatial filtering rather than DOA estimation and the estimation is only performed over a one-dimensional search. This assumption limits computational complexity as well as makes the model more robust against stronger interfering harmonic sources from other directions. Finally, a reconstruction of the clean speech for the k'th microphone can be obtained given the estimated pitch, ω_0 and the delay, τ,

$$\hat{\mathbf{s}}_k = \Pi_{\mathbf{ZD}(k)}\mathbf{x}_k \tag{B.7}$$

with the projection matrix $\Pi_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$. The reconstructed clean speech signal to be used as input to the non-intrusive objective intelligibility metric is then obtained by summing the estimated signal over all microphone channels

$$\hat{\mathbf{s}} = \frac{1}{K}\sum_{k=0}^{K-1}\hat{\mathbf{s}}_k \tag{B.8}$$

Alternatively, the variance estimates in (B.5) can be used to form a weighted estimate.

2.3 Experimental methodology

The proposed metric PB-STOI is evaluated using two different microphone array setups: A broadside uniform linear array (ULA) consisting of $K = 10$ microphones and a behind the ear (BTE) HA setup consisting of two bilateral wireless linked HAs with $K = 4$ microphones. The ULA has a microphone spacing of $d = c/f_s$ and the delay of the desired source between microphone 0 and the k'th microphone is given by $\tau_k = kdc^{-1}\sin\theta$, where the wave propagation speed was $c = 343$ m/s. The DOA of the desired source was $\theta = 0°$ and the sampling frequency was $f_s = 8$ kHz. For the BTE HA setup the spacing between the microphone on each HA was 1 cm and the spacing between the two HAs was 25 cm.

In the experimental evaluation the set of fundamental frequencies was set to the range $\Omega_0 = 100 - 400$ Hz, the model order was estimated using the maximum a posteriori (MAP) criterion [18], the short-time segmentation window block size was 30 ms and reconstructed by overlap-and-add using a Hanning window with 50% overlap. The simulations were performed using a complex multi-talker scenario with 8 interfering speakers (Fig. B.2), reverberation (RT60 = 0.3 s) and ambient white noise in a room with dimensions of 10x6x4 m simulated for 2.5 s using the toolbox McRoomSim [17]. The simulations were carried out in three scenarios at SNRs ranging from -20 to 20 dB; a white noise only scenario, one with interferers and white noise and one with interferers, white noise and reverberation. The desired speech was the

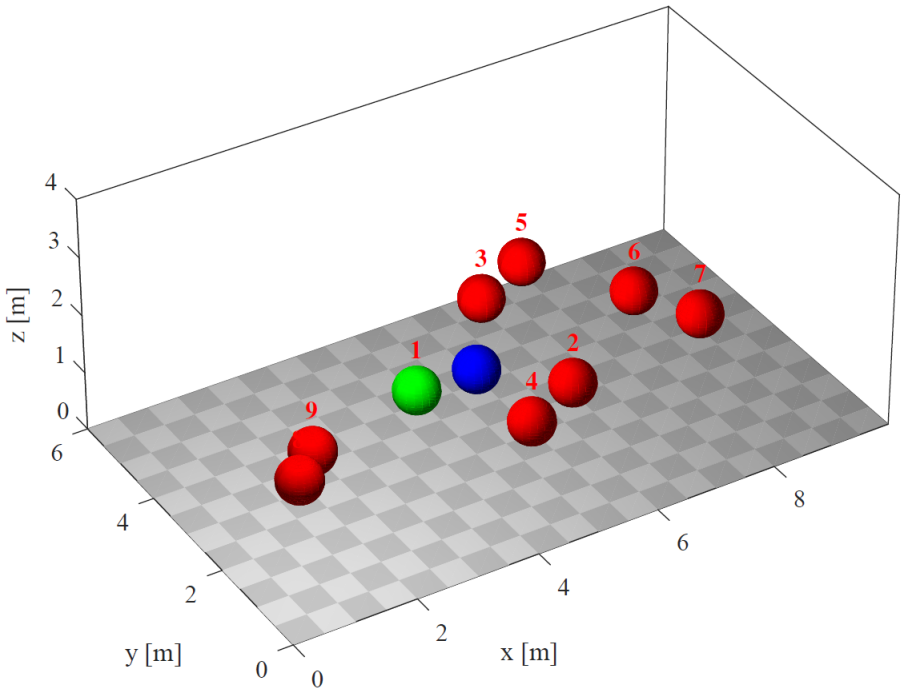


Fig. B.2: The experimental setup simulated with the software toolbox McRoomSim [17]. The blue, green and red balls illustrate the location of the listener, the desired target source and the interferers, respectively.

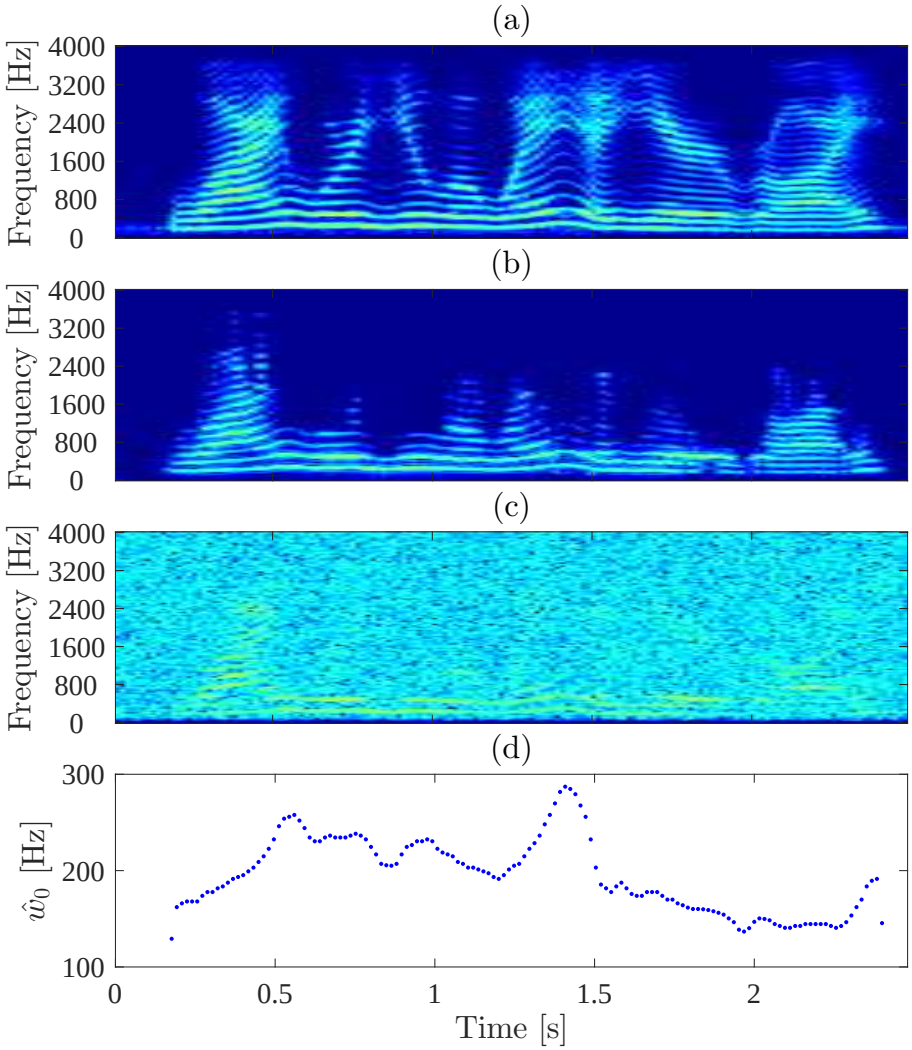
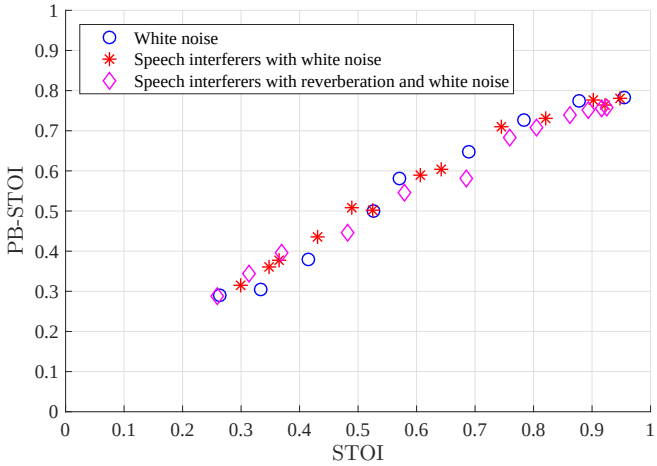
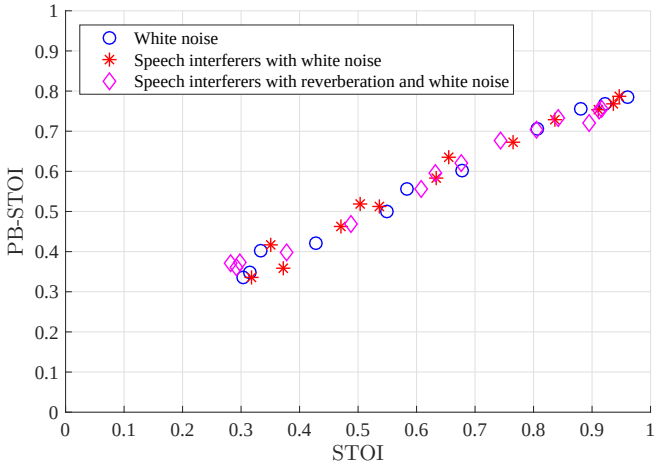


Fig. B.3: Spectrograms of (a) the clean voiced utterance "Why were you away a year, Roy", (b) the reconstructed speech signal using the estimated pitch from the harmonic model, and (c) the noisy signal at 0 dB SNR, and plot of (d) the estimated fundamental frequency from the noisy signal.



(a) Results from PB-STOI using a ULA setup.



(b) Results from PB-STOI using a BTE HA setup.

Fig. B.4: Scatter plots of the non-intrusive PB-STOI metric versus the intrusive STOI metric. The pitch of the PB-STOI metric is estimated using a multi-channel signal from (a) a ULA with $K = 10$ microphones and (b) two bilateral BTE HAs setup. The circles, asterisks and diamonds show the simulated results for white noise only, multiple interferers with white noise without and with reverberation, respectively.

utterance "Why were you away a year, Roy" from the voiced corpus in [19] and the interferers were speech samples from the EUROM_1 database of the English sentence corpus [20].

3 Results and discussion

The spectrograms of (a) the original clean speech, (b) the equivalent reconstructed signal and (c) the degraded noisy signal at 0 dB as well as (d) the estimated pitch from the noisy signal are depicted in Fig. B.3. Comparison of Figs. B.3(a) and (b) indicates that the reconstructed speech signal has captured relatively well the features of the original clean signal.

The performance of the proposed intelligibility measure is evaluated by comparing the correlation between the non-intrusive PB-STOI scores against the original intrusive STOI scores in Fig. B.4 for (a) the ULA setup and (b) the bilateral BTE HA setup. It can be observed that the PB-STOI scores correlate well with the original intrusive scores with a strong linear trend between the two metrics for both microphone array setups. Thus, it is promising that a small microphone array such as the HA setup can give acceptable results.

The performance of the proposed PB-STOI metric is evaluated in Table B.1 using three performance criteria often used for assessing objective intelligibility metrics [6, 11]. Pearson's correlation (ρ) quantifies the linear relationship, while Spearman's rank (ρ_{spear}) and Kendall's tau (τ) characterize the ranking capability. The values are close to one for all performance criteria indicating high correlation between the intrusive and non-intrusive metric. Hence, the proposed non-intrusive PB-STOI metric can offer a comparable performance to the original intrusive intelligibility metric.

Compared with the study in [11] which uses a similar approach for non-intrusive intelligibility prediction, the proposed PB-STOI metric only requires a calibration of the conversion between PB-STOI and STOI scores depending on the array configuration without any training to the data. However, the experimental evaluation only contained voiced speech and should also be tested on utterances containing unvoiced parts. This could be done by only assessing the intelligibility in the voiced parts of the speech using a voiced speech detector. It is expected to obtain similar results for sentences also containing unvoiced parts, since the most energetic regions occur during the voiced parts. According to the glimpsing model of speech in noise the most energetic regions of the desired speech are most important for intelligibility and thus a good predictor for intelligibility [21]. As such, it is a reasonable assumption that using only the voiced regions of the speech can yield a promising predictor for speech intelligibility.

Table B.1: Performance of the proposed metric in terms of Pearson's correlation (ρ), the Spearman rank ($\rho_{\mathbf{spear}}$) and Kendall's tau (τ) between PB-STOI and STOI as well as their linear regression lines for a ULA and bilateral BTE HA setup.

Setup	ρ	$\rho_{\mathbf{spear}}$	τ	Regression line
ULA	0.9886	0.9887	0.9287	$0.74x + 0.11$
BTE HA	0.9812	0.9004	0.9922	$0.67x + 0.16$

4 Conclusion

This paper proposes a non-intrusive intelligibility metric for online processing in HAs. A clean speech signal is reconstructed by its spatio-temporal characteristics (i.e. direction of arrival and pitch) using only the noisy speech signal and utilized inside an established and reliable intrusive intelligibility metric, which requires a clean reference. The proposed non-intrusive metric has a high correlation with the original intrusive counterpart and thus is a promising method for online assessment of speech intelligibility in HAs.

References

[1] R. W. Peters, B. C. J. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 577–587, 1998.

[2] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1725–1736, 1990.

[3] P. C. Loizou, *Speech Enhancement: Theory and Practice*, ser. Signal processing and communications. Taylor & Francis, 2007.

[4] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[5] V. Hamacher, J. Chalupper, E. Eggers, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Applied Signal Process.*, vol. 18, pp. 2915–2929, 2005.

[6] 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 Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[8] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[9] 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.

[10] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[11] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-hmm-based non-intrusive speech intelligibility prediction," in *ICASSP*, March 2016, pp. 624–628.

[12] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *EUSIPCO*, August 2016, pp. 1358–1362.

[13] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," *Comput. Speech Lang.*, vol. 35, no. C, pp. 73–92, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2015.06.002

[14] L. Lamarche, C. Gigure, W. Gueaieb, T. Aboulnasr, and H. Othman, "Adaptive environment classification system for hearing aids," *J. Acoust. Soc. Am.*, vol. 127, no. 5, pp. 3124–3135, 2010.

[15] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Statistically efficient methods for pitch and doa estimation," in *ICASSP*, May 2013, pp. 3900–3904.

[16] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Process.*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[17] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, 2010, pp. 1–6.

[18] P. M. Djuric, "Asymptotic map criteria for model selection," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2726–2735, Oct 1998.

[19] M. Cooke, "Modelling auditory processing and organisation," Ph.D. dissertation, Cambridge University Press, 1993.

[20] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *EUROSPEECH*, vol. 1, 18-21 September 1995, pp. 867–870.

[21] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

Paper C

Non-Intrusive Intelligibility Prediction using a Codebook-Based Approach

Charlotte Sørensen
Mathew Shaji Kavalekalam
Angeliki Xenaki
Jesper Bünsow Boldt
Mads Græsbøll Christensen

The paper has been presented at the
25th European Signal Processing Conference (EUSIPCO), pp. 216–220, Kos, Greece, 2017.

The layout has been revised.

Abstract

It could be beneficial for users of hearing aids if these were able to automatically adjust the processing according to the speech intelligibility in the specific acoustic environment. Most speech intelligibility metrics are intrusive, i.e., they require a clean reference signal, which is rarely available in real-life applications. This paper proposes a method, which allows using an intrusive short-time objective intelligibility (STOI) metric without requiring access to a clean signal. The clean speech reference signal is replaced by the clean speech envelope spectrum estimated from the noisy signal. The spectral envelope has been shown to be an important cue for speech intelligibility and is used as the reference signal inside STOI. The spectral envelopes are estimated as a combination of predefined dictionaries, i.e., codebooks, that best fits the noisy speech signal. The simulations show a high correlation between the proposed non-intrusive codebook-based STOI (NIC-STOI) and the intrusive STOI indicating that NIC-STOI is a suitable metric for automatic classification of speech signals.

1 Introduction

Speech is a fundamental tool for human communication. Understanding speech becomes a challenging task in adverse listening conditions such as "the cocktail party scenario" especially for hearing impaired individuals [1, 2]. Speech enhancement algorithms aim to improve speech intelligibility for hearing aid users [3–5]. However, speech enhancement algorithms may be beneficial in some acoustic scenarios whereas the same algorithms can have a negative impact on quality and intelligibility in other conditions [5, 6]. Thus, it would be beneficial for HA users if speech enhancement algorithms are automatically limited to scenarios in which they provide an improvement in speech intelligibility [3, 4]. This could be facilitated by an objective speech intelligibility metric processed online in the HA.

Several methods can with an acceptable accuracy predict the speech intelligibility intrusively, i.e., they require access to a clean speech reference [4]. Some of the earliest intrusive metrics that predict the intelligibility well for a limited type of degradations, like linear filtering and additive noise, include the articulation index (AI) [7] and the speech transmission index (STI) [8]. Later, the short-time objective (STOI) metric [9] and the speech-based envelope power spectrum model (sEPSM) [10] were introduced for more complex distortion types and are reported to have an useful reliability [4]. However, the need for a clean speech signal would be a limitation for real-time prediction of speech intelligibility, since this is rarely available. More recently, a number of non-intrusive metrics not requiring access to a clean speech reference signal have been introduced, e.g., the speech-to-reverberation modu-

lation energy ratio (SRMR) [11], the modulation spectrum area (ModA) [12]. These methods are, however, either limited to assessment of reverberated speech or still inferior to the intrusive metrics [4].

This paper proposes a non-intrusive intelligibility prediction method referred to as the non-intrusive codebook-based STOI (NIC-STOI). The method estimates the intelligibility of noisy speech non-intrusively by comparing relevant features of the clean speech with the features of the noisy speech inside a well-established intrusive intelligibility framework, STOI, similar to [13, 14]. The relevant features of the clean speech are based on the spectral envelope of the speech, which has been shown to be an important cue for speech intelligibility [15]. The spectral envelopes of the clean speech and the noise signal are estimated as the most suitable combination from a predefined speech and noise spectra dictionary, a codebook, which best fits the noisy speech signal using a codebook-based approach [16, 17]. These codebooks consist of filter coefficients that capture the overall structure of the spectral envelope.

2 The NIC-STOI measure

NIC-STOI allows predicting the intelligibility from the noisy signal only using an intrusive metric (STOI) without requiring access to the clean speech signal. The approach behind the method is to replace the clean reference signal with an estimate of the clean speech features obtained from the noisy signal. An estimate of the clean speech spectral envelope is used as the relevant features of speech intelligibility in the method. Then, NIC-STOI gives a non-intrusive intelligibility prediction by comparing the correlation of the estimated clean speech spectrum with the noisy spectrum with the intrusive STOI measure. The framework of the measure is illustrated by a block diagram in Fig. C.1. The framework can be divided into three main steps: (1) The parameters needed to obtain the clean speech reference are estimated, (2) time-frequency-spectra of the clean and noisy speech signals are composed from the estimated parameters, and, (3) an intelligibility score is predicted with the intrusive STOI framework.

2.1 Signal model

The proposed method is based on an additive noise model assuming the speech and noise are statistically uncorrelated from [16, 17], i.e.,

$$y(n) = s(n) + w(n), \tag{C.1}$$

where $y(n)$, $s(n)$ and $w(n)$ represent the sampled noisy speech, clean speech and noise, respectively. The clean speech signal can be modeled as a stochas-

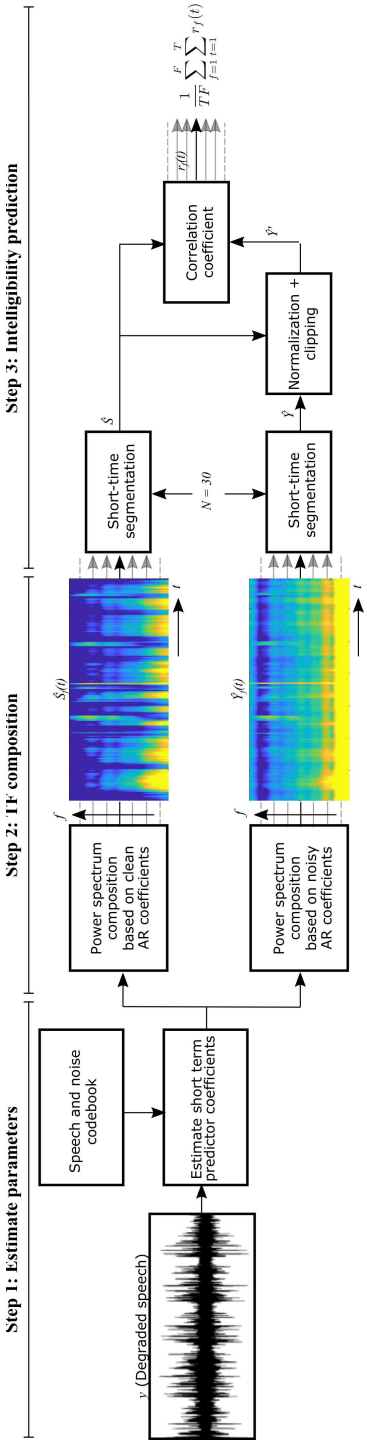


Fig. C.1: Block diagram illustrating the proposed non-intrusive codebook-based STOI metric in which the relevant features of the clean and noisy speech signals are composed as time-frequency power spectra using a codebook-based approach and utilized within the intrusive framework, STOI.

tic autoregressive (AR) process

$$s(n) = \sum_{i=1}^{P} a_{s_i}(n)s(n-i) + u(n) = \mathbf{a}_s(n)^T\mathbf{s}(n-1) + u(n), \tag{C.2}$$

where $\mathbf{s}(n-1) = [s(n-1), \ldots, s(n-P)]^T$ with the P past speech samples, $\mathbf{a}_s(n) = [a_{s_1}(n), a_{s_2}(n), \ldots, a_{s_P}(n)]^T$ is a vector containing the speech linear prediction coefficients (LPC), and $u(n)$ is zero mean white Gaussian noise with excitation variance $\sigma_u^2(n)$. Similarly, the noise signal can be modeled as

$$w(n) = \sum_{i=1}^{Q} a_{w_i}(n)w(n-i) + v(n) = \mathbf{a}_w(n)^T\mathbf{w}(n-1) + v(n), \tag{C.3}$$

where $\mathbf{w}(n-1) = [w(n-1), \ldots, w(n-Q)]^T$ with the Q past noise samples, $\mathbf{a}_w(n) = [a_{w_1}(n), a_{w_2}(n) \ldots, a_{w_Q}(n)]^T$, and $v(n)$ is zero mean white Gaussian noise with excitation variance $\sigma_v^2(n)$.

The AR model is used to model the speech and noise signals as well as training the codebook dictionaries.

2.2 Step 1: Estimate parameters

The spectra of the clean and noisy speech signals are estimated from the LPC and the excitation variances concatenated in the vector $\theta = [\mathbf{a}_s \ \mathbf{a}_w \ \sigma_u^2(n) \ \sigma_v^2(n)]$. These parameters are estimated using a priori information from a trained codebook about the speech and noise spectral shapes in the form of LPC based on the approach in [16–18], where more details on the derivation of this method can be found. Given the observed vector of noisy samples $\mathbf{y} = [\ y(0) \ y(1) \ \ldots \ y(N-1)\ ]$ for the current frame of length N, the MMSE (minimum mean square error) estimate of θ can be given as $\hat{\theta} = \mathrm{E}(\theta|\mathbf{y})$ for the support space of the parameters to be estimated, Θ, and using Bayes' theorem can be reformulated as

$$\hat{\theta} = \int_{\Theta} \theta p(\theta|\mathbf{y}) d\theta = \int_{\Theta} \theta \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} d\theta. \tag{C.4}$$

The vector, $\theta_{ij} = [\mathbf{a}_{s_i} \ \mathbf{a}_{w_j} \ \sigma_{u,ij}^{2,\mathrm{ML}}(n) \ \sigma_{v,ij}^{2,\mathrm{ML}}(n)]$, is then defined for each ith entry of the speech codebook and j^{th} entry of the noise codebook, respectively. The maximum likelihood (ML) estimates of the speech and noise excitation variances, $\sigma_{u,ij}^{2,\mathrm{ML}}$ and $\sigma_{v,ij}^{2,\mathrm{ML}}$, respectively, are then given by [16, 18]

$$\mathbf{C} \begin{bmatrix} \sigma_{u,ij}^{2,\mathrm{ML}} \\ \sigma_{v,ij}^{2,\mathrm{ML}} \end{bmatrix} = \mathbf{D}, \tag{C.5}$$

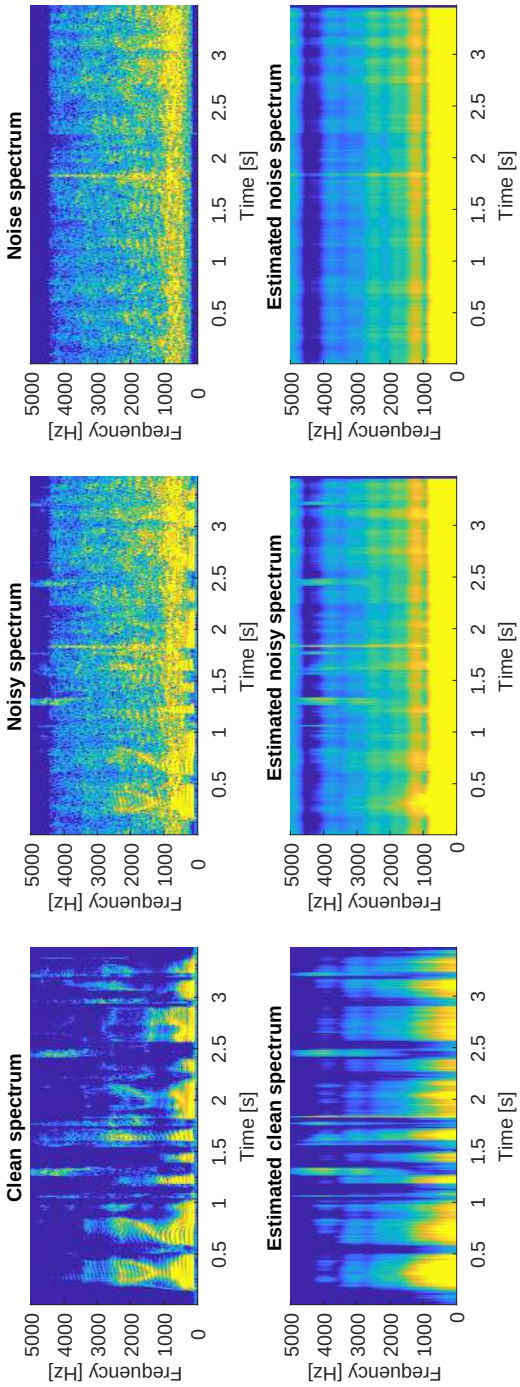


Fig. C.2: Spectrograms of the original clean speech signal, noisy speech signal at 0 dB SNR and noise signal are depicted in the top panel from left to right, respectively, as well as their corresponding estimated power spectra from the codebook-based approach in the bottom panel.

where

$$\mathbf{C} = \begin{bmatrix} \| \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^4} \| & \| \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2} \| \\ \| \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2} \| & \| \frac{1}{P_y^2(\omega)|A_w^j(\omega)|^4} \| \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} \| \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2} \| \\ \| \frac{1}{P_y^2(\omega)|A_w^j(\omega)|^2} \| \end{bmatrix} \tag{C.6}$$

where A_s^i and A_w^j are the spectra of the ith and j^{th} vector from the speech codebook and noise codebook, respectively, and with $\|f(\omega)\| = \int |f(\omega)| d\omega$. The spectral envelope of the speech codebook, the noise codebook and the noisy signal are given by $\frac{1}{|A_s^i(\omega)|^2}$, $\frac{1}{|A_w^j(\omega)|^2}$ and $P_y(\omega)$, respectively. In practice, the MMSE estimate of θ in Eq. C.4 is evaluated as a weighted linear combination of θ_{ij} by

$$\hat{\theta} = \frac{1}{N_s N_w} \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} \theta_{ij} \frac{p(\mathbf{y}|\theta_{ij}) p(\sigma_{u,ij}^{2,\mathrm{ML}}) p(\sigma_{v,ij}^{2,\mathrm{ML}})}{p(\mathbf{y})}, \tag{C.7}$$

where N_s and N_w are the the number of entries in the speech and noise codebooks, respectively. The weight of the MMSE estimate, $p(\mathbf{y}|\theta_{ij})$, can be computed as

$$p(\mathbf{y}|\theta_{ij}) = e^{-d_{\mathrm{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega))} \tag{C.8}$$

$$\hat{P}_y^{ij}(\omega) = \frac{\sigma_{u,ij}^{2,\mathrm{ML}}}{|A_s^i(\omega)|^2} + \frac{\sigma_{v,ij}^{2,\mathrm{ML}}}{|A_w^j(\omega)|^2} \tag{C.9}$$

$$p(\mathbf{y}) = \frac{1}{N_s N_w} \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} p(\mathbf{y}|\theta_{ij}) p(\sigma_{u,ij}^2) p(\sigma_{v,ij}^2), \tag{C.10}$$

where the Itakura-Saito distortion between the noisy spectrum and the modeled noisy spectrum is given by $d_{\mathrm{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega))$ [17, 19]. The weighted summation of the LPC should be performed in the line spectral frequency domain in order to insure stable inverse filters [16, 17].

2.3 Step 2: TF composition

Time-frequency (TF) power spectrum of the estimated reference signal, $\hat{S}$, are composed from the estimated AR filter coefficients of the clean speech signal $\hat{\mathbf{a}}_s$ for each time frame:

$$\hat{S}(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2}, \tag{C.11}$$

where $\hat{A}_s(\omega) = \sum_{k=0}^{P} \hat{a}_{s_k} e^{-j\omega k}$. In the same manner, the estimated noise AR filter coefficients, $\hat{\mathbf{a}}_w$, are used to compose a TF spectrum of the noise:

$$\hat{W}(\omega) = \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}, \tag{C.12}$$

where $\hat{A}_w(\omega) = \sum_{k=0}^{Q} \hat{a}_{w_k} e^{-j\omega k}$. The LPC, i.e. $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_w$, determine the shape of the envelope of the corresponding signals $\hat{S}(\omega)$ and $\hat{W}(\omega)$, respectively. The excitation variances, $\hat{\sigma}_u$ and $\hat{\sigma}_v$, determine the overall signal magnitude. Finally, the noisy spectrum is composed as the combined sum of the clean and the noise power spectra:

$$\hat{Y}(\omega) = \hat{S}(\omega) + \hat{W}(\omega). \tag{C.13}$$

These time-frequency spectra replace the discrete Fourier transform of the clean reference signal and the noisy signal in the original STOI measure [9].

2.4 Step 3: Intelligibility Prediction

In the final step, the intelligibility prediction is carried out in exactly the same manner as for the STOI measure [9]. The power spectra of the noisy speech, $\hat{Y}$, are further clipped by a normalisation procedure expressed in Eq. C.14 in order to de-emphasize the impact of region in which noise dominates the spectrum:

$$\hat{Y}' = \max(\min(\lambda \cdot \hat{Y}, (1 + 10^{-\beta/20}) \cdot \hat{S}), (1 - 10^{-\beta/20}) \cdot \hat{S}), \tag{C.14}$$

where $\hat{S}$ is the power spectrum of the estimated reference signal, $\lambda = \sqrt{\sum \hat{S}^2 / \sum \hat{Y}^2}$ is a scale factor for normalizing the noisy TF bins and $\beta = -15$ dB is the lower signal-to-distortion ratio. Given the local correlation coefficient, $r_f(t)$, between $\hat{Y}$ and $\hat{S}$ at frequency f and time t, the NIC-STOI prediction is given by averaging across all bands and frames:

$$\text{NIC-STOI} = \frac{1}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} r_f(t). \tag{C.15}$$

3 Simulation methodology

The proposed metric NIC-STOI is evaluated on speech samples from of 5 male and 5 female speakers from the EUROM_1 database of the English sentence corpus [20]. The interfering additive noise signal is simulated in the range of -30 to 30 dB SNR as multi-talker babble from the NOIZEUS database [6]. The LPC and variances of both the speech and noise signal are

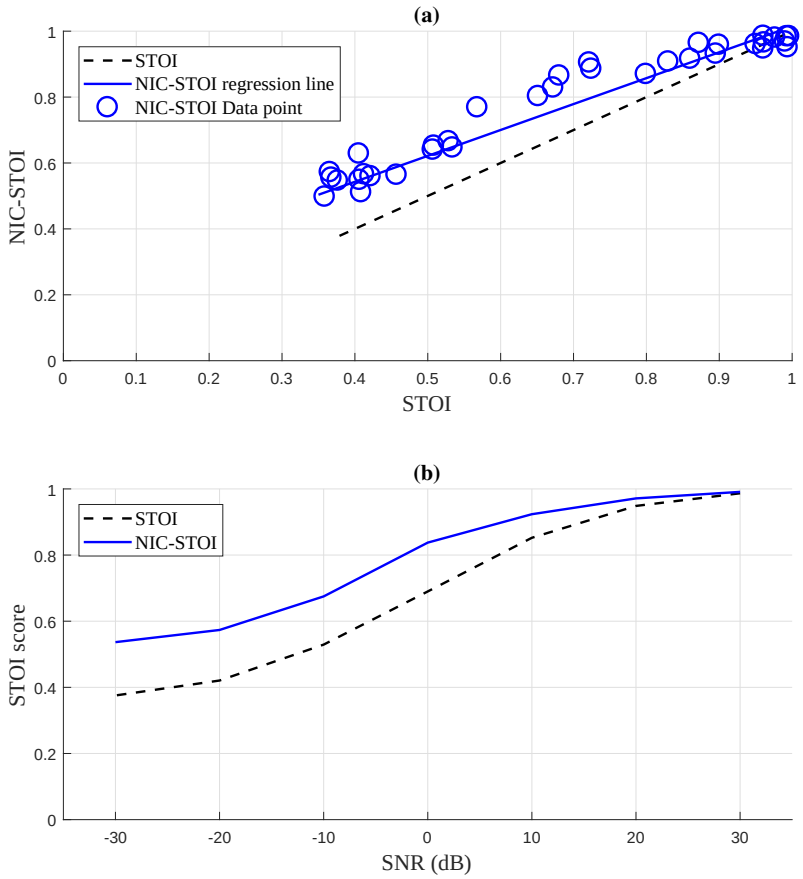


Fig. C.3: (a) Scatter plot of the non-intrusive codebook-based STOI (NIC-STOI) metric versus the intrusive STOI metric and (b) STOI and NIC-STOI as a function of SNR.

estimated from 25.6 ms frames with sampling frequency 10 kHz. The speech and, thus, the STP parameters are assumed to be stationary over these very short frames. The AR model order P and Q of both the speech and noise, respectively, is set to 14 according to literature [16–18]. The speech codebook is generated on a training sample of 15 minutes of speech from multiple speakers in the EUROM_1 database in order to assure a generic speech model using the generalized Lloyd algorithm (GLA) [16, 21]. The speech codebook training sample does not include speech samples from the speakers used in the test set. The noise codebook is trained on 2 minutes of babble talk. The sizes of the speech and noise codebooks are $N_s = 64$ and $N_w = 8$, respectively. The performance of the metric is evaluated using three performance criteria

common for assessment of objective intelligibility metrics [4, 14]; Pearson's correlation (ρ) which characterizes the linear relationship, Spearman's rank (ρ_{spear}) and Kendall's tau (τ) which both quantify the ranking capability.

4 Results and Discussion

The spectra of an example speech signal in the test set is shown in Fig. C.2 for the original clean speech signal, the noisy speech signal at 0 dB SNR and the noise signal in the top panel from left to right, respectively. In the bottom panel the corresponding estimated power spectra of relevant signal features are composed using trained codebooks of speech and noise spectral shapes parametrized as LPC to model the a priori information in a Bayesian MMSE scheme.

It can be observed that the method only captures the overall envelope structure and not the fine structure of speech, since it is based on an AR model [17, 19]. Only modeling the overall envelope structure is assumed to be sufficient for depicting the essential features of clean speech, since the envelope structure has long been identified as an important cue for speech intelligibility used within other intrusive intelligibility prediction frameworks, i.e., STI and EPSM [8, 10, 15]. This viewpoint can also be supported by extensive vocoder simulations, where it has been shown that envelope cues from only four spectral bands are sufficient to yield a high intelligibility of speech perception in quiet [15]. As such, it seems to be a reasonable assumption that only depicting the overall envelope structure can be a good predictor for speech intelligibility.

The performance of the NIC-STOI metric is evaluated in relation to the corresponding original STOI scores. In Fig. C.3a there is a clear monotonic correspondence between the NIC-STOI score (blue solid line) and the intrusive STOI measure (black dashed line), such that a higher NIC-STOI score also corresponds to a higher STOI score. Furthermore, a strong linear trend can be observed between the NIC-STOI and STOI measures. This observation is also supported by the performance criteria given in Table C.1, where Pearson's correlation and the Spearman Rank is close to one implying a high correlation. This indicates that the proposed non-intrusive version of STOI can offer a comparable performance to the original intrusive STOI. In Fig. C.3b the STOI measure (black dashed line) and the NIC-STOI measure (blue solid line) are depicted as function of SNR. There is a clear monotonic correspondence between NIC-STOI and STOI, such that a higher STOI measure results in a higher NIC-STOI score. Furthermore, the NIC-STOI scores also increase with increasing SNRs. The offset between the two graphs can be accounted for by the linear trend described in Table C.1, which gives the translation between NIC-STOI and STOI scores.

Table C.1: Performance of the proposed metric in terms of Pearson's correlation (ρ), the Spearman rank ($\rho_{\mathbf{spear}}$) and Kendall's tau (τ) between NIC-STOI and STOI as well as the linear regression line.

Metric	ρ	$\rho_{\mathbf{spear}}$	τ	Regression line
NIC-STOI	0.972	0.961	0.8521	$0.730 \cdot \text{STOI} + 0.285$

In future work, it would be interesting to investigate how the method performs with different noise types and environments as well as unseen noise conditions. Additionally, the objective results could be tested against subjective listening experiments for further validation in future work .

5 Conclusion

This paper proposes a method for objective prediction of speech intelligibility. The proposed method, NIC-STOI, allows using an intrusive intelligibility metric (STOI) without requiring access to the clean speech signal. Hence, NIC-STOI is essentially a non-intrusive metric. In principle, the method predicts the speech intelligibility by replacing the clean reference signal with an estimate of its spectrum. The features of the clean speech signal are estimated using a codebook-based approach, where the spectral shape of the speech is trained and parametrized using LPC. The proposed NIC-STOI metric shows a high correlation with the intrusive original STOI score and, hence, seems promising for predicting speech intelligibility non-intrusively using an intrusive intelligibility metric.

Acknowledgment

This work was supported by the Innovation Fund Denmark, Grant No. 99-2014-1.

References

[1] R. W. Peters, B. C. J. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 577–587, 1998.

[2] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1725–1736, 1990.

[3] V. Hamacher, J. Chalupper, E. Eggers, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Applied Signal Process.*, vol. 18, pp. 2915–2929, 2005.

[4] 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 Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[5] P. C. Loizou, *Speech Enhancement: Theory and Practice*, ser. Signal processing and communications. Taylor & Francis, 2007.

[6] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[7] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[8] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[10] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[11] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[12] 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.

[13] C. Sørensen, A. Xenaki, J. Boldt, and M. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *ICASSP*, March 2017, pp. 386–390.

[14] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-hmm-based non-intrusive speech intelligibility prediction," in *ICASSP*, March 2016, pp. 624–628.

[15] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[16] M. Kavalekalam, M. Christensen, F. Gran, and J. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *ICASSP*, March 2016, pp. 191–195.

[17] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[18] ——, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[19] K. Paliwal and W. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*. Elsevier Science, 1995, pp. 433–468.

[20] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *EUROSPEECH*, vol. 1, 18-21 September 1995, pp. 867–870.

[21] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

Paper D

Non-Intrusive Codebook-Based Intelligibility Prediction

Charlotte Sørensen
Mathew Shaji Kavalekalam
Angeliki Xenaki
Jesper Bünsow Boldt
Mads Græsbøll Christensen

The paper has been published in
Speech Communication Vol. 101, pp. 85–93, 2018.

The layout has been revised.

Abstract

In recent years, there has been an increasing interest in objective measures of speech intelligibility in the speech processing community. Important progress has been made in intrusive measures of intelligibility, where the Short-Time Objective Intelligibility (STOI) method has become the de facto standard. Online adaptation of signal processing in, for example, hearing aids, in accordance with the listening conditions, requires a non-intrusive measure of intelligibility. Presently, however, no good non-intrusive measures exist for noisy, nonstationary conditions. In this paper, we propose a novel, non-intrusive method for intelligiblity prediction in noisy conditions. The proposed method is based on STOI, which measures long-term correlations in the clean and degraded speech. Here, we propose to estimate the clean speech using a codebook-based approach that jointly models the speech and noisy spectra, parametrized by auto-regressive parameters, using pre-trained codebooks of both speech and noise. In experiments, the proposed method is demonstrated to be capable of accurately predicting the intelligibility scores obtained with STOI from oracle information. Moreover, the results are validated in listening tests that confirm that the proposed method can estimate intelligibility from noisy speech over a range of signal-to-noise ratios.

1 Introduction

Human interaction depends on communication where speech has a central role. Inability to understand speech, e.g., due to hearing impairment, noisy background, or distortion in communication systems, can lead to ineffective communication and social isolation, and the development of speech enhancement methods [1, 2] is, therefore, a key concern in many applications. These include challenging applications such as hearing aids [3], telecommunication systems [4, 5], and architectural acoustics [6]. To assess the listening conditions in which speech processing would be beneficial, but also to evaluate the speech processing algorithms as such, a speech intelligibility measure is required [3, 5, 7].

A natural way of assessing the intelligibility of a degraded, i.e., processed, distorted or noisy speech signal is by performing subjective listening tests. Subjective speech intelligibility scores gives the percentage of correctly identified information from a degraded speech signal. However, subjective speech intelligibility experiments are time-consuming, expensive and cannot be used for real-time applications. Hence, there is a great interest in developing objective measures for speech intelligibility prediction. As opposed to subjective listening tests, objective intelligibility prediction algorithms are faster, cheaper and can be used for real-time processing.

The Articulation Index (AI) [8, 9] and the Speech Intelligibility Index (SII) [10] are some of the earliest metrics for prediction of speech intelligi-

bility scores. The AI and SII use the signal-to-noise ratio (SNR) of speech excerpts in several frequency bands to estimate the intelligibility, hence they require that both the clean speech signal and the noise are available and uncorrelated as well as the noise to be stationary. The Extended SII (ESII) [11] and the Coherence SII (CSII) [12], are variants of SII which account for fluctuating noise and nonlinear distortions from clipping, respectively. The Speech Transmission Index (STI) [4] was introduced to predict the intelligibility of an amplitude modulated signal at the output of a transmission channel based on changes in the modulation depth across frequency of a probe signal. The STI, which requires a probe signal as reference, offers good prediction of speech intelligibility in reverberant and noisy conditions [4], but not for more adverse nonlinear distortions, such as those caused by spectral subtraction [13]. The Short-Time Objective Intelligibility (STOI) metric [14] predicts the intelligibility of a signal by its short-time correlation with its clean counterpart which is required as input. STOI estimates are accurate for time-frequency processed speech [15, 16]. The speech-based Envelope Power Spectrum Model (sEPSM) [17] estimates the SNR in the envelope-frequency domain and uses the noise signal alone as reference. The sEPSM accounts for the effects of additive noise and reverberation and some types of nonlinear processing such as spectral subtraction [17], but fails with other types of nonlinear processing such as ideal binary masks and phase jitter [16]. More recent work includes that of [18], which takes an information theoretical approach to the problem.

All the aforementioned methods are intrusive, i.e., they require either the clean speech signal or the noise interference as reference to estimate the intelligibility of the degraded signal. Access to the clean speech signal is impractical for many real-life applications or real-time processing systems. To overcome this limitation, a number of non-intrusive objective intelligibility measures have been proposed. The Speech to Reverberation Modulation energy Ratio (SRMR) [19] and the average Modulation-spectrum Area (ModA) [20] both provide intelligibility predictions based on the modulation spectrum of the degraded speech signal, i.e., in a non-intrusive manner. Other notable work includes the reduced dynamic range (rDR) based intelligibility measure [21], wherein the intelligibility is predicted directly from the dynamic range of the noisy speech, and the across-band envelope correlation (ABEC) metric [22], which is based on temporal envelope waveforms. Another approach to predict speech intelligibility non-intrusively is to first obtain an estimate of the clean speech signal which is thereafter used as reference to an intrusive method. Machine learning [23, 24], principal component analysis [7] or noise reduction [25, 26] methods have been proposed to reconstruct the clean signal from its degraded version and use it as input to the intrusive STOI metric for objective intelligibility prediction.

The present paper, which is an extension of our prior work [27], pro-

poses a non-intrusive intelligibility metric, which uses the STOI measure non-intrusively by estimating the features of the clean reference signal from its degraded version. The proposed method, however, estimates the reference signal by identifying the entries of pre-trained codebooks of speech and noise spectra which best fit the data, i.e., the noisy speech signal. The resulting new metric is dubbed Non-Intrusive Codebook-based STOI (NIC-STOI). The method is inspired by the work [28, 29] which demonstrates that codebook-based approaches offer effective speech enhancement, even under nonstationary noise such as babble noise. Moreover, the approaches of [28, 29] are based on low-dimensional parametrizations of both the noise and speech spectra, more specifically, via auto-regressive (AR) models, something that engenders both effective training leading to small codebooks and computationally fast implementations. Furthermore, an AR process models the envelope of the signal's spectrum rather than its fine structure. Such models are suitable in this context since it is shown that the spectral envelope of speech is an important cue for intelligibility [30]. Compared to our previous work [26], which can be interpreted as sampling the speech spectrum at high-SNR frequencies based on the pitch, something that is consistent with the glimpsing model of speech perecetion [31], the new method is based on the complete speech spectrum. It should also be noted that we here address the problem of single-channel non-intrusive intelligibility prediction, which is a much more difficult task than the multichannel problem [25, 26], as the latter can use spatial information.

The rest of the paper is organized as follows. First, the principles of intelligibility prediction in the STOI method are described in Section 2. Then, the signal model that the proposed method is based on is detailed in Section 3, and the proposed non-intrusive method is described in Section 4. The experimental details and results, which include both experiments with objective measures and a listening test, are first described in Section 5 and then discussed in Section 6. Finally, Section 7 concludes on the work.

2 Background

The STOI [14] metric predicts the speech intelligibility based on the correlation between the temporal envelopes of the clean and the degraded speech signal (see Fig. D.1). First, the clean and degraded speech signals are decomposed in time-frequency representations using a discrete Fourier transform. Then, these time-frequency representations are grouped in one-third octave frequency bins and short-time segments (384 ms). The short-time segments are normalized in order to account for global level differences of the input signals. Furthermore, the short-time segments are clipped to prevent time-frequency units that are already completely degraded from excessively

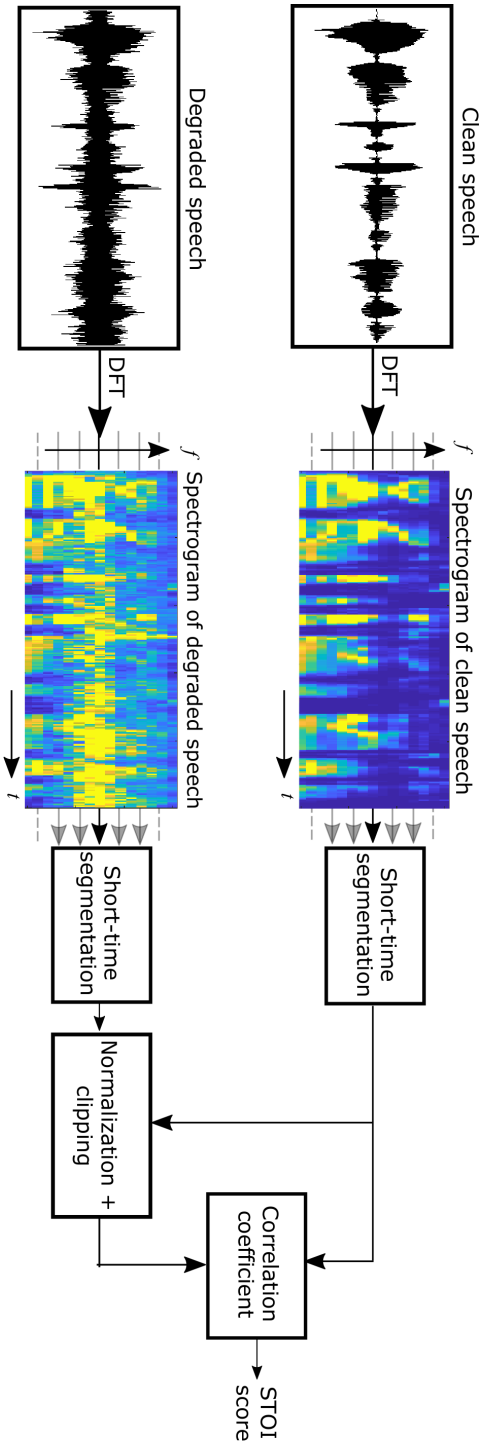


Fig. D.1: Block diagram of the STOI measure [14] that forms the basis for the proposed non-intrusive method. The STOI metric is based on the correlation between temporal envelopes of the clean and degraded speech in short time segments.

influencing the intelligibility score. Finally, the correlation of the signals is calculated over the short-time segments per frequency band. The STOI output is the average of the correlation coefficients across frequency bands and time-segments, i.e., a scalar value in the range 0-1 which relates monotonically to the average speech intelligibility scores.

3 Signal model

Assuming that a speech signal and a noise signal are generated by uncorrelated random processes, the corresponding noisy speech signal, $y(n)$, at time instance n is $y(n) = s(n) + w(n)$. In the proposed method, both the speech and the noise are modeled as stochastic processes, namely AR processes [28, 29]. Using such a stochastic AR model, a segment of the speech signal is expressed as

$$s(n) = -\sum_{i=1}^{P} a_s(i)s(n-i) + u(n), \tag{D.1}$$

which can also be expressed in vector notation as

$$u(n) = \mathbf{a}_s^T \mathbf{s}(n) \tag{D.2}$$

where P is the order of the AR process, $\mathbf{s}(n) = [s(n), s(n-1), \ldots, s(n-P)]^T$ is a vector collecting the P past speech samples, $\mathbf{a}_s = [1, a_s(1), a_s(2), \ldots, a_s(P)]^T$ is a vector containing the speech auto-regressive parameters with $a_s(0) = 1$, and $u(n)$, which here models the excitation, is zero mean white Gaussian noise with excitation variance σ_u^2. Transforming the AR model into the frequency domain, $A_s(\omega)S(\omega) = U(\omega) \Leftrightarrow S(\omega) = U(\omega)/A_s(\omega)$, results in the following power spectrum:

$$P_s(\omega) = |S(\omega)|^2 = \frac{\sigma_u^2}{|A_s(\omega)|^2}, \tag{D.3}$$

where $A_s(\omega) = \sum_{k=0}^{P} a_s(k)e^{-j\omega k}$. Similarly to the speech signal, the noise signal can be modeled as

$$w(n) = -\sum_{i=1}^{Q} a_w(i)w(n-i) + v(n), \tag{D.4}$$

which can also be expressed as

$$v(n) = \mathbf{a}_w^T \mathbf{w}(n), \tag{D.5}$$

where Q is the order of the AR process, $\mathbf{w}(n) = [w(n), w(n-1), \ldots, w(N-Q)]^T$ is a vector collecting the Q past noise samples, $\mathbf{a}_w =$

$[1, a_w(1), a_w(2) \ldots, a_w(Q)]^T$ where $a_w(0) = 1$, and $v(n)$ is zero mean white Gaussian noise with excitation variance σ_v^2. The noisy power spectrum is likewise given by

$$P_w(\omega) = |W(\omega)|^2 = \frac{\sigma_v^2}{|A_w(\omega)|^2}. \tag{D.6}$$

where $A_w(\omega) = \sum_{k=0}^{Q} a_w(k) e^{-j\omega k}$.

The models of the the speech and noise in (D.2) and (D.5), respectively, can be motivated as follows. The AR model has a long history in speech processing, where one of its uses is in modeling the speech production system (see, e.g., [32]), where it corresponds to a cylinder model of the vocal tract which is excited by a noise signal generated by the lungs. The model is, though, well-known not to be perfect. For example, it does not account for the nasal cavity and the Gaussian model is only a good model for unvoiced speech and less so for voiced speech [33]. Nevertheless, it remains useful for many purposes and here it is used as a low-dimensional representation of the speech spectrum. Regarding the noise, the model is good for many natural noise sources, but, in any case, it can be used for modeling arbitrary, smooth spectra of Gaussian signals [34].

4 The NIC-STOI measure

The proposed method provides an objective measure for speech intelligibility prediction given solely the degraded speech signal, i.e., non-intrusively.

The method is based on the speech and noise being additive and the AR models of the speech (D.2) and noise (D.5) signals. The speech and noise spectra are simultaneously estimated from the degraded speech signal using a Bayesian approach which uses the AR parameters as prior information for inference. The prior information is obtained from trained codebooks (dictionaries) of speech and noise AR parameters. The estimation is performed on short-time frames in order to account for non-stationary noise.

Figure D.2 depicts a block diagram of the NIC-STOI algorithm. The methodology comprises three main steps: 1) estimation of the parameters for the speech and noise AR models, 2) computation of the time-frequency representations for the clean, s, and noisy speech, y, signals from the estimated parameters, 3) prediction of speech intelligibility of the noisy speech signal with the STOI framework from the estimated spectra.

4.1 Step 1: Parameter Estimation

Let the column vector $\boldsymbol{\theta} = [\mathbf{a}_s;\ \mathbf{a}_w;\ \sigma_u^2;\ \sigma_v^2]$ comprise all parameters to be estimated, i.e., the AR coefficients and the excitation variances of the models of both speech and noise.

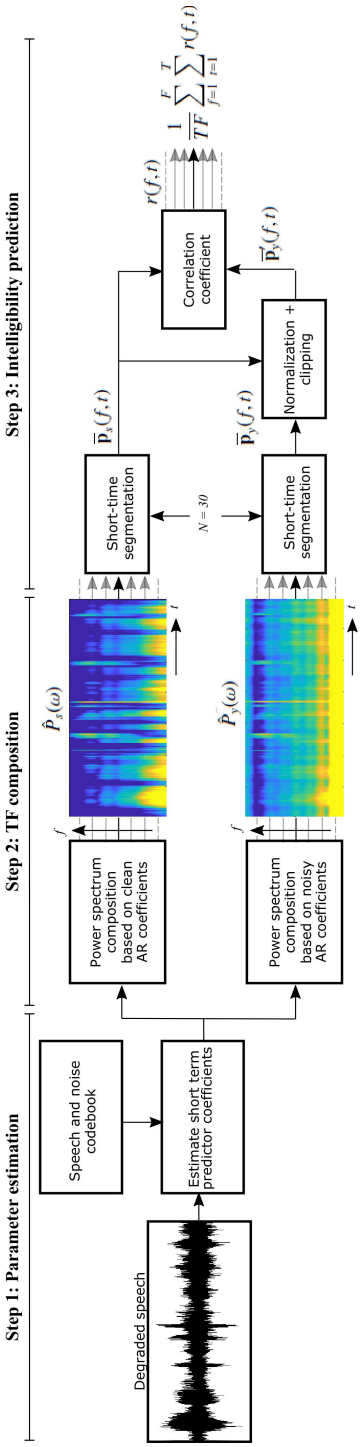


Fig. D.2: Block diagram depicting the processing scheme of the proposed non-intrusive codebook-based short-time objective intelligibility (NIC-STOI) metric. The relevant features of the clean and degraded speech signals are estimated using a codebook-based approach as time-frequency power spectra, which replace the estimates in the front-end of the STOI method.

Bayes' theorem facilitates the computation of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ of the model parameters $\boldsymbol{\theta}$ conditioned on the observation of N noise samples, i.e., $\mathbf{y} = [\ y(0)\ y(1)\ \ldots\ y(N-1)\ ]$, from the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, the prior distribution of the model parameters $p(\boldsymbol{\theta})$, and the marginal distribution of the data $p(\mathbf{y})$ [28, 29, 35]:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{D.7}$$

Based on the signal model introduced previously, the data likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, is a multi-variate zero-mean Gaussian distribution with covariance matrix, $\mathbf{R}_Y = \mathbf{R}_s + \mathbf{R}_w$, where $\mathbf{R}_s = \sigma_u^2(\mathbf{G}_s^T\mathbf{G}_s)^{-1}$ and $\mathbf{G}_s$ is a $N \times N$ lower triangular Toeplitz matrix defined by the AR parameters $\mathbf{a}_s$. More specifically, it is given by

$$\mathbf{G}_s = \begin{bmatrix} 1 & 0 & \ldots & & 0 \\ a_s(1) & 1 & & & \\ \vdots & a_s(1) & & & \\ a_s(P) & \vdots & \ddots & & \vdots \\ 0 & a_s(P) & \ddots & & \\ \vdots & \vdots & & 1 & \\ 0 & 0 & \ldots & a_s(1) & 1 \end{bmatrix} \tag{D.8}$$

while the noise covariance matrix can be expressed as $\mathbf{R}_w = \sigma_v^2(\mathbf{G}_w^T\mathbf{G}_w)^{-1}$ with $\mathbf{G}_w$ being defined in a similar manner as $\mathbf{G}_s$ but from $\mathbf{a}_w$. Then, the minimum mean square error (MMSE) estimate is given by [36]

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{MMSE}} &= \arg\min_{\hat{\boldsymbol{\theta}}} \mathrm{E}\left[(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta})^2\right] = \mathrm{E}(\boldsymbol{\theta}|\mathbf{y}) \\ &= \int_\Theta \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} = \int_\Theta \boldsymbol{\theta} \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} d\boldsymbol{\theta}, \end{aligned} \tag{D.9}$$

where Θ is the support space of the parameters to be estimated. Based on the independence of speech and noise signals, and further assuming that the AR process and excitation variances are independent, the prior distribution of the model parameters can be simplified as

$$p(\boldsymbol{\theta}) = p(\mathbf{a}_s, \sigma_u^2)p(\mathbf{a}_w, \sigma_v^2) \approx p(\mathbf{a}_s)p(\sigma_u^2)p(\mathbf{a}_w)p(\sigma_v^2).$$

Limiting the support of the AR parameter vectors $\mathbf{a}_s$ and $\mathbf{a}_w$ to predefined codebooks of size N_s and N_w, respectively, the corresponding excitation variances are estimated through a maximum likelihood (ML) approach

$$\{\sigma_{u,ij}^{2,\text{ML}}, \sigma_{v,ij}^{2,\text{ML}}\} = \arg\max_{\sigma_u^2, \sigma_v^2} \log p(\mathbf{y}|\mathbf{a}_{s_i}^{\text{CB}};\ \mathbf{a}_{w_j}^{\text{CB}};\ \sigma_u^2;\ \sigma_v^2),$$

where $\mathbf{a}_{s_i}^{\mathrm{CB}}$ is the ith entry of the speech codebook and $\mathbf{a}_{w_j}^{\mathrm{CB}}$ is the jth entry of the noise codebook. The Gaussian likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ can be expressed in the frequency domain in terms of the Itakura-Saito distortion measure between the observed, $P_y(\omega)$, and modeled, $\hat{P}_y^{ij}(\omega)$, noisy data power spectrum, i.e.,

$$p(\mathbf{y}|\mathbf{a}_{s_i}^{\mathrm{CB}};\ \mathbf{a}_{w_j}^{\mathrm{CB}};\ \sigma_{u,ij}^2;\ \sigma_{v,ij}^2) \propto e^{-d_{\mathrm{IS}}(P_y(\omega),\hat{P}_y^{ij}(\omega))}, \tag{D.10}$$

where $d_{\mathrm{IS}}(\cdot,\cdot)$ is the Itakura-Saito divergence, which is given by [29, 37]

$$d_{\mathrm{IS}}(P_y(\omega),\hat{P}_y^{ij}(\omega)) = \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{P_y(\omega)}{\hat{P}_y^{ij}(\omega)} - \ln\left(\frac{P_y(\omega)}{\hat{P}_y^{ij}(\omega)}\right) - 1\right)d\omega. \tag{D.11}$$

Equation (D.11) makes use of the modeled noisy power spectrum, which is here given by

$$\hat{P}_y^{ij}(\omega) = \frac{\sigma_u^2}{|A_s^i(\omega)|^2} + \frac{\sigma_v^2}{|A_w^j(\omega)|^2}, \tag{D.12}$$

where $A_s^i(\omega) = \sum_{k=0}^{P} a_s^{i,\mathrm{CB}}(k)e^{-j\omega k}$ and $A_w^j(\omega) = \sum_{k=0}^{Q} a_w^{j,\mathrm{CB}}(k)e^{-j\omega k}$ being the spectra of the ith and j^{th} vector from the speech codebook and noise codebook, respectively.

Assuming that the modeling error between $P_y(\omega)$ and $\hat{P}_y^{ij}(\omega)$ is small and by using a second-order Taylor series approximation of $\ln(\cdot)$, the Itakura-Saito divergence can be approximated as [29]

$$d_{\mathrm{IS}}(P_y(\omega),\hat{P}_y^{ij}(\omega)) \approx \frac{1}{2}d_{\mathrm{LS}}\left(P_y(\omega),\hat{P}_y^{ij}(\omega)\right), \tag{D.13}$$

where the log-spectral distortion between the observed and modeled noisy spectrum, $d_{\mathrm{LS}}\left(P_y(\omega),\hat{P}_y^{ij}(\omega)\right)$, which is given by

$$d_{\mathrm{LS}}\left(P_y(\omega),\hat{P}_y^{ij}(\omega)\right) = \frac{1}{2\pi}\int_0^{2\pi}\left|\ln\left(\frac{\sigma_u^2}{|A_s^i(\omega)|^2} + \frac{\sigma_v^2}{|A_w^j(\omega)|^2}\right) - \ln\left(P_y(\omega)\right)\right|^2 d\omega \tag{D.14}$$

Finally, the ML estimates of the speech and noise excitation variances, $\sigma_{u,ij}^{2,\mathrm{ML}}$ and $\sigma_{v,ij}^{2,\mathrm{ML}}$ can be obtained by

$$\{\sigma_{u,ij}^{2,\mathrm{ML}},\sigma_{v,ij}^{2,\mathrm{ML}}\} = \arg\min_{\sigma_u^2,\sigma_v^2} d_{\mathrm{LS}}\left(P_y(\omega),\hat{P}_y^{ij}(\omega)\right), \tag{D.15}$$

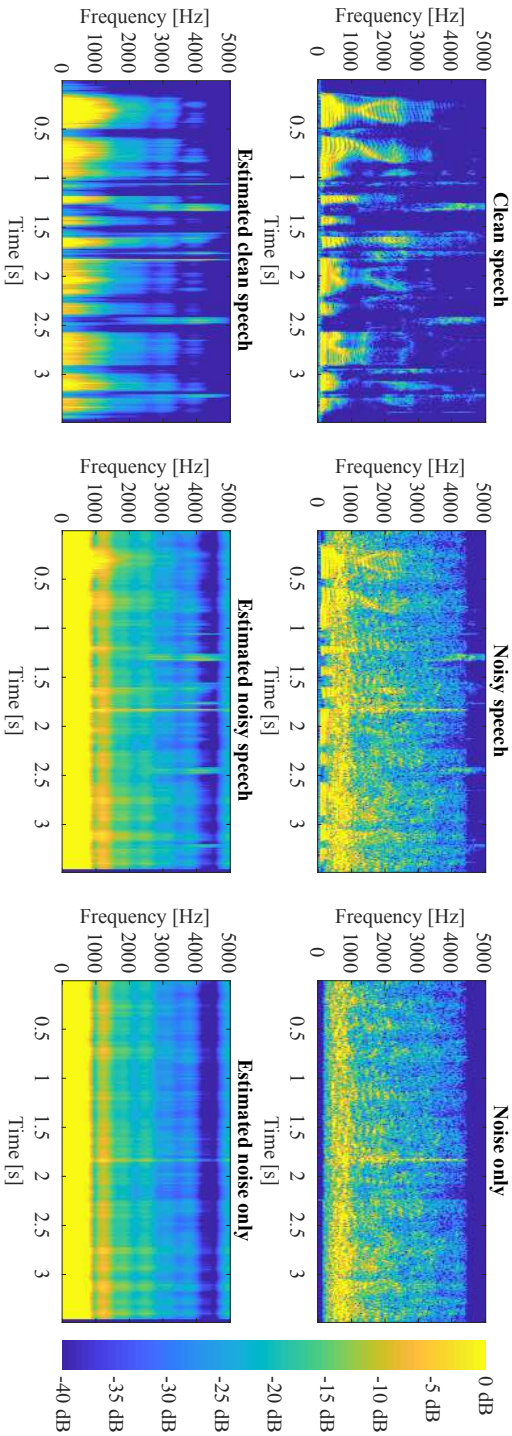


Fig. D.3: The top panel depicts from left to right, respectively, the spectra of the original clean speech signal, the degraded noisy speech signal at 0 dB SNR and noise only. In the bottom panel their corresponding estimated spectra using the codebook-based approach are depicted.

which is solved by differentiating (D.14) with respect to σ_u^2 and σ_v^2 and setting the result equal to zero [28, 35]. This results in the following estimate of the excitation variance for the speech:

$$\sigma_{u,ij}^{2,\mathrm{ML}} = \frac{1}{\Psi_{ij}} \Bigg(\sum_{\omega} \frac{1}{P_y^2(\omega)|A_w^j(\omega)|^4} \sum_{\omega} \frac{1}{P_y(\omega)|A_s^i(\omega)|^2} - \sum_{\omega} \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2} \sum_{\omega} \frac{1}{P_y(\omega)|A_w^j(\omega)|^2} \Bigg).$$

Similarly, the estimate of for excitation variance of the noise is given by

$$\sigma_{v,ij}^{2,\mathrm{ML}} = \frac{1}{\Psi_{ij}} \Bigg(\sum_{\omega} \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^4} \sum_{\omega} \frac{1}{P_y(\omega)|A_w^j(\omega)|^2} - \sum_{\omega} \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2} \sum_{\omega} \frac{1}{P_y(\omega)|A_s^i(\omega)|^2} \Bigg).$$

The quantity Ψ_{ij} is given by

$$\Psi_{ij} = \sum_{\omega} \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^4} \sum_{\omega} \frac{1}{P_y^2(\omega)|A_w^j(\omega)|^4} - \left(\sum_{\omega} \frac{1}{P_y^2(\omega)|A_s^i(\omega)|^2|A_w^j(\omega)|^2} \right)^2. \tag{D.16}$$

Finally, based on these estimates, the quantities in (D.9) are estimated from their discrete counterparts, which are given by

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_s N_w} \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} \boldsymbol{\theta}_{ij} \frac{p(\mathbf{y}|\boldsymbol{\theta}_{ij})}{p(\mathbf{y})} \tag{D.17}$$

and

$$p(\mathbf{y}) = \frac{1}{N_s N_w} \sum_{i=1}^{N_s} \sum_{j=1}^{N_w} p(\mathbf{y}|\boldsymbol{\theta}_{ij}), \tag{D.18}$$

where $\boldsymbol{\theta}_{ij} = [\mathbf{a}_{s_i}^{\mathrm{CB}};\ \mathbf{a}_{w_j}^{\mathrm{CB}};\ \sigma_{u,ij}^{2,\mathrm{ML}};\ \sigma_{v,ij}^{2,\mathrm{ML}}]$ is the resulting parameter vector for the ith entry of the speech codebook and the jth entry of the noise codebook and the final estimates are denoted as $\hat{\boldsymbol{\theta}} = [\hat{\mathbf{a}}_s;\ \hat{\mathbf{a}}_w;\ \hat{\sigma}_u^2;\ \hat{\sigma}_v^2]$. These estimates can be thought of as being obtained from an average over all possible models with each model being weighted by its posterior. We remark that codebook combinations that result in infeasible, negative values for either the speech

or noise excitation variances should be neglected. Since all ML estimates of the excitation variances and the predefined codebook entries contribute with equal probability, the prior is non-informative and is omitted in (D.9). It should also be noted that the weighted summation of the AR parameters can be performed in the line spectral frequency (LSF) domain whereby a stable inverse filters is ensured, something that is not always the case when operating directly on the AR parameters [28, 29].

4.2 Step 2: TF composition

The estimated parameters in $\hat{\boldsymbol{\theta}}$, obtained using (D.17), are then used to compute the time-frequency (TF) power spectra of the estimated speech and noise spectra as

$$\hat{P}_s(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2}, \tag{D.19}$$

where $\hat{A}_s(\omega) = \sum_{k=0}^{P} \hat{a}_s(k) e^{-j\omega k}$, and

$$\hat{P}_w(\omega) = \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}, \tag{D.20}$$

where $\hat{A}_w(\omega) = \sum_{k=0}^{Q} \hat{a}_w(k) e^{-j\omega k}$. The AR parameters, i.e., $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_w$, determine the shape of the envelope of the corresponding signals $\hat{S}(\omega)$ and $\hat{W}(\omega)$, respectively. The excitation variances, $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$, determine the overall signal power. Finally, the noisy spectrum is composed as the combined sum of the clean and the noise power spectra:

$$\hat{P}_y(\omega) = \hat{P}_s(\omega) + \hat{P}_w(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2} + \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}. \tag{D.21}$$

These time-frequency spectra replace the discrete Fourier transform of the clean reference signal and the noisy signal in the original STOI measure, respectively.

4.3 Step 3: Intelligibility Prediction

The STOI measure is used for intelligibility prediction with the estimated spectra $\hat{P}_s(\omega)$ (D.19) and $\hat{P}_y(\omega)$ (D.21) as inputs. First, the frequency bins of $\hat{P}_s(\omega)$ and $\hat{P}_y(\omega)$ are grouped into 15 one-third octave bands denoted by $\overline{P}_s(f,t)$ and $\overline{P}_y(f,t)$, respectively, with the lowest center frequency set to 150 Hz and the highest set to 4.3 kHz. The short-time region of the temporal envelopes of the clean speech is defined as $\overline{\mathbf{p}}_s(f,t) = [\overline{P}_s(f,t-N+1), \overline{P}_s(f,t-N+2), \ldots, \overline{P}_s(f,t)]^T$, where N is the length of the short-time regions and is set to 30, resulting in a short-time region of 384 ms as in the original STOI

implementation [14]. In the same manner, the short-time region of the degraded speech is given by $\overline{\mathbf{p}}_y(f,t)$. The short-time regions of the degraded speech, $\overline{\mathbf{p}}_y(f,t)$, are further clipped by a normalization procedure in order to de-emphasize the impact of region in which noise dominates the spectrum:

$$\overline{\mathbf{p}}'_y(f,t) = \min\left(\frac{\|\overline{\mathbf{p}}_s(f,t)\|_2}{\|\overline{\mathbf{p}}_y(f,t)\|_2}\overline{\mathbf{p}}_y(f,t), (1+10^{-\beta/20})\overline{\mathbf{p}}_s(f,t)\right)$$

where $\|\cdot\|_2$ denotes the l_2 norm and $\beta = -15$ dB is the lower signal-to-distortion ratio. The local correlation coefficient, $r(f,t)$, between $\overline{\mathbf{p}}'_y(f,t)$ and $\overline{\mathbf{p}}_s(f,t)$ at frequency f and time t, is defined as

$$r(f,t) = \frac{(\overline{\mathbf{p}}_s(f,t) - \mu_{\overline{\mathbf{p}}_s(f,t)})^T(\overline{\mathbf{p}}'_y(f,t) - \mu_{\overline{\mathbf{p}}'_y(f,t)})}{\sqrt{(\overline{\mathbf{p}}_s(f,t) - \mu_{\overline{\mathbf{p}}_s(f,t)})^2}\sqrt{(\overline{\mathbf{p}}'_y(f,t) - \mu_{\overline{\mathbf{p}}'_y(f,t)})^2}},$$

where $\mu(\cdot)$ denotes the sample average of the corresponding vector. Given the local correlation coefficient, the NIC-STOI prediction is given by averaging across all bands and frames as

$$d_{NS} = \frac{1}{TF}\sum_{f=1}^{F}\sum_{t=1}^{T} r(f,t). \tag{D.22}$$

5 Experimental Details and Results

5.1 Performance Measures

The non-intrusive intelligibility prediction is given by d_{NS}, for the different conditions to be evaluated. Whereas the ground truth, denoted by d_S, for these conditions are given by the intrusive STOI scores. Similarly to the approach in [24], the original true STOI score is expected to be well-correlated with the subjective intelligibility. Thus, the purpose is to predict the intrusive STOI score of a given condition using a non-intrusive method. The performance of the objective intelligibility predictions are evaluated using three performance metrics often used for assessing objective intelligibility predictions [3, 14, 38]:

- The Pearson correlation coefficient (ρ) quantifies the linear relationship between the predicted non-intrusive intelligibility scores and true STOI scores or subjective intelligibility scores, where a higher ρ indicates higher correlation.
- Kendall's Tau (τ) characterizes the ranking capability by describing the monotonic relationship between the predicted intelligibility

Table D.1: Sentence syntax of the GRID database [39] which is used in the subjective listening test. Each sentence is constructed from (in order) a combination of a command, color, preposition, letter digit, and adverb.

Command	Color	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	0-9	again
lay	green	by	(no W)		now
place	red	in			please
set	white	with			soon

scores and true STOI scores or subjective intelligibility scores, where a higher τ represents better performance [40]. It is defined as $\tau = 2(n_c - n_d)/N(N-1)$, where n_c is the number of concordant pairs, i.e. ordered in the same way, and n_d is the number of discordant pairs, i.e. ordered differently.

- The standard deviation of the prediction error (σ) is given as a measure of the estimation accuracy of the predicted non-intrusive intelligibility scores, where a lower σ implies better results.

5.2 Experimental Details

The results reported in this paper are based on both objective measurements and subjective listening tests. For the results based on the objective measures, the proposed metric, NIC-STOI, is evaluated on a test set of 100 speech utterances (full sentences), 50 male and 50 female, randomly selected from the EUROM_1 database of the English corpus [41]. The interfering additive noise signal is babble noise from the AURORA database. The babble noise contains many speakers in a reverberant acoustical environment. The sentences and interfering additive noise signal are both resampled to 10 kHz. Segments randomly selected from the additive noise signal are added to the EUROM_1 sentences at different SNR levels in the range of -30 to 30 dB SNR in steps of 10 dB SNR.

For further evaluation of the proposed metric, a subjective listening test has also been carried out to provide a data set for comparing NIC-STOI and SRMR. Stimuli were the fixed-syntax sentences from the GRID corpus database [39] mixed with the babble signal from the AURORA database with an SNR range -8 to 0 dB. The grid corpus consists of sentences spoken by 34 talkers (16 female and 18 male). The sentences are simple, syntactically identical phrases, e.g. "place blue in A 4 again", and the listeners are asked to identify the color, letter, and digit after listening to the stimuli using a user-controlled MATLAB interface. The syntax and words of the GRID corpus are shown in Table D.1. A total of nine subjects were used for the experiment

which took around 30 minutes per subject. Intelligibility was defined as the number of keywords correctly identified per stimulus resulting in a fraction of either 0, 1/3, 2/3, or 1 being correct. A total of 220 stimuli were used, approximately 2 s in duration each, with the same stimuli being used for both NIC-STOI and SRMR: 5 SNR levels times 44 different sentences. We remark that to reduce intra- and intersubject variability the condition-averaged results are used for comparison and mapping of the objective results to subjective performance [3, 42]. Measuring intelligibility on a short time-scale (i.e., from short stimuli less than 2 s in duration each) with non-stationary noise types implies a high variance for both subjective and objective evaluations, i.e., precise estimation of intelligibility requires multiple sentences and not only a few words. However, it is difficult to execute subjective listening tests using long sentences or phrases as stimulus for which reason the average of many shorter sentences is here used instead.

The AR parameters and excitation variances of both the speech and noise signal are estimated on frames with a length of 256 samples. The speech and, thus, the estimated parameters are assumed to be stationary over these very short 25.6 ms frames. The frames are windowed using a Hann window with 50 % overlap between adjacent frames. The AR model orders P and Q of the speech and noise, respectively, are set to 14 in accordance with the literature [28, 29, 35]. The speech codebook is trained using the generalized Lloyd algorithm (GLA) on 10 minutes of speech from multiple speakers in the EUROM_1 database in order to ensure a sufficiently general speech model [28, 43]. We stress that the speakers included in the test set are not used for the training of the speech codebook. The noise codebook is trained on 2 minutes of babble talk. The sizes of the speech and noise codebooks are $N_s = 64$ and $N_w = 8$, respectively.

5.3 Experimental Results

An example of the spectrum of a speech signal from the test set is shown in Fig. D.3 . The spectra of the original clean speech signal, the degraded noisy signal at 0 dB SNR and the noisy only are depicted in the top panel from left to right, respectively. The corresponding estimated spectra of the relevant signal features are shown in the bottom panel. The spectra are generated using trained codebooks of speech and noise spectral shapes. The estimated clean spectrum (bottom left panel) and estimated noisy spectrum (bottom middle panel) are used as input to the intrusive STOI framework.

The performance of the NIC-STOI metric is evaluated against the intrusively computed scores of the original STOI metric as ground truth. In Fig. D.4, the estimated NIC-STOI scores have been plotted against the intrusive STOI scores. The plot shows good performance by means of a strong monotonic relationship between NIC-STOI and STOI, such that a higher NIC-

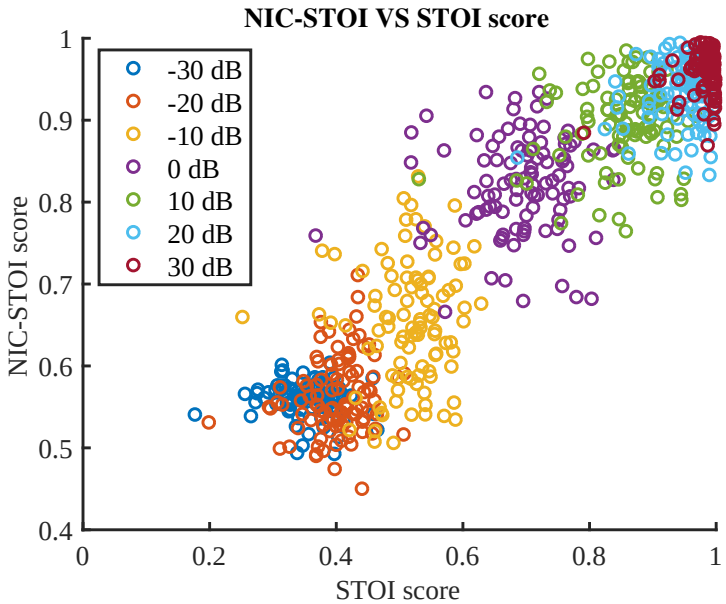


Fig. D.4: Scatter plot of the predicted STOI scores using the non-intrusive codebook-based STOI, NIC-STOI, metric.

STOI score also corresponds to a higher STOI score. Furthermore, a strong linear correlation can be observed between the two measures. This observation is also supported by the performance criteria, where NIC-STOI achieves a Pearson's correlation of $\rho = 0.94$, Kendall's Tau of $\tau = 0.70$ and a standard deviation of the prediction error $\sigma = 0.14$ for STOI, implying a high correlation. This indicates that the proposed non-intrusive version of STOI can offer a comparable performance to the original intrusive STOI.

Fig. D.5 depicts the averaged predictions ($\pm$ standard deviation) of the NIC-STOI scores in the scatter plot in Fig. D.4 for male (blue line), female (red line) and both genders (yellow line), where the performance measures are given in Tab. D.2. As it can be observed, the measure performs equally well whether the method is tested using either a gender specific clean speech codebook or a generic clean speech codebook. This suggests that the method generalizes well and does not capture gender specific effects due to the very generic and smooth structure of the spectra of the auto-regressive processes.

In Fig. D.6 the STOI measure (purple line) and the NIC-STOI measure (male: blue line; female: red line; both genders: yellow line) are depicted as function of SNR. There is a clear monotonic correspondence between NIC-STOI and STOI, such that a higher STOI measure results in a higher NIC-STOI score. Furthermore, the NIC-STOI scores also increase with increasing SNR.

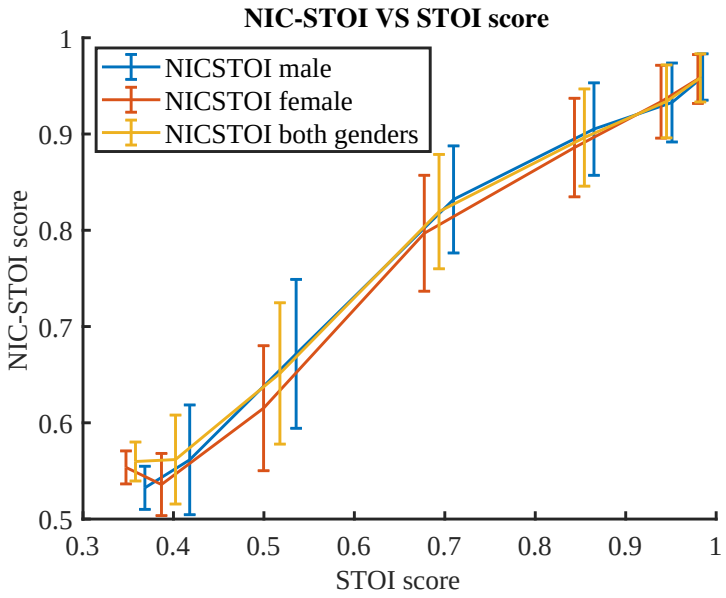


Fig. D.5: Averaged NIC-STOI scores (± standard deviation) against the intrusively computed STOI score.

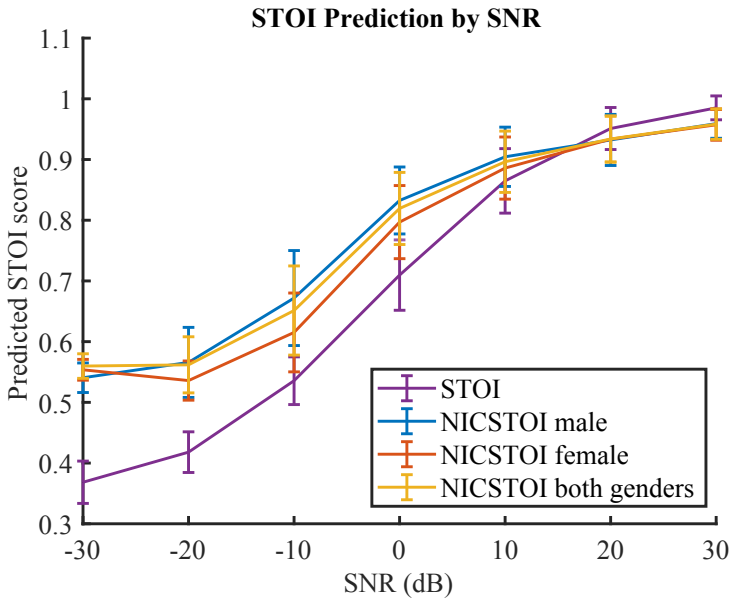


Fig. D.6: Averaged NIC-STOI and STOI scores (± standard deviation) per SNR condition.

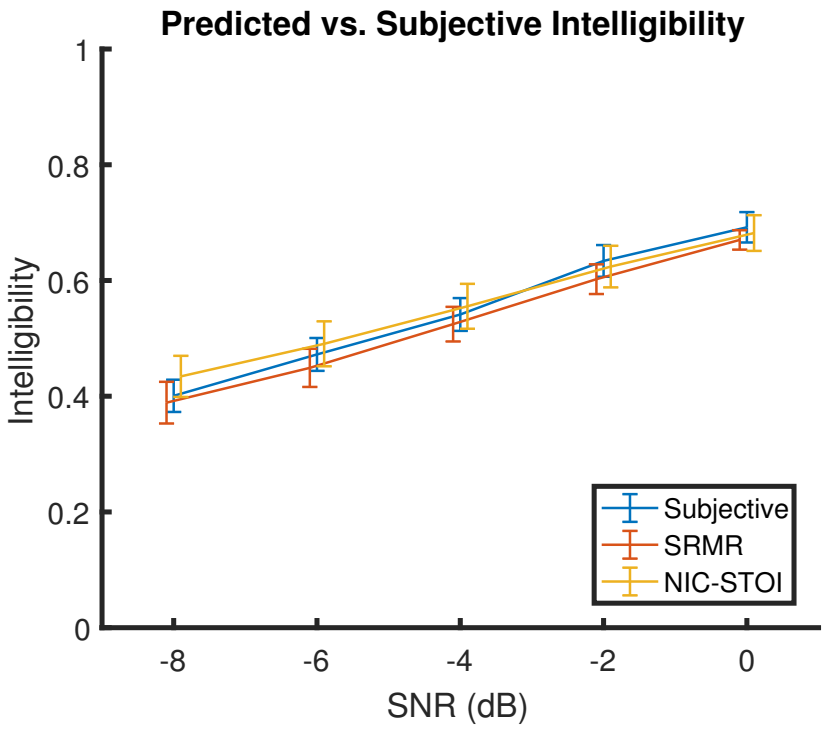


Fig. D.7: Intelligibility as a function of SNR for subjective listening experiments and as predicted by the proposed NIC-STOI and SRMR. Shown are the means and their 95 % confidence intervals.

Subjective results, in terms of intelligibility as a function of SNR, are shown in Fig. D.7 together with objective results obtained using the proposed NIC-STOI and SRMR. The error bars in the Figure are 95 % confidence intervals computed using a normal distribution for the SRMR and NIC-STOI methods and the normal approximation for the binomial confidence interval of the subjective results from the listening test. Note that to map the objective results to subjective intelligibility, a sigmoid function has been fitted to the average data as described in Section 5.2. As can be seen, the proposed method performs well and is capable of predicting the speech intelligibility with similar variance over a range of SNRs. The results do not, however, enable the conclusion that NIC-STOI is superior to SRMR although NIC-STOI has a better alignment with the subjective data, as both metrics have a good performance, even at low SNRs, and the confidence intervals overlap. Concerning the probability intervals, the intervals for both NIC-STOI and SRMR are large, as is to be expected, due to the short sentences in the GRID corpus and the limited number of stimuli for each SNR level. One thing to note is that the variance for SRMR increases as the SNR decreases, whereas

Table D.2: Performance of the proposed metric in terms of Pearson's correlation (ρ), and Kendall's tau (τ) and the standard deviation of the prediction error (σ) between NIC-STOI and STOI.

Condition	ρ	τ	σ
Male	0.93	0.70	0.14
Female	0.94	0.71	0.13
Both genders	0.94	0.70	0.14

NIC-STOI exhibits a similar variance across SNRs.

6 Discussion

Since the framework of NIC-STOI is based on an AR model, it only captures the overall envelope structure and not the fine structure of the speech signal as illustrated in Fig. D.3 [29, 37]. The envelope of the speech has been shown to be a good predictor for speech intelligibility in previous intrusive intelligibility frameworks, i.e. STI and EPSM [4, 17, 30]. Extensive vocoder simulations also support these findings, where a high speech intelligibility can be obtained in quiet solely from the envelope content in only four spectral bands [30]. As such, only modeling the envelope structure of the clean speech as the essential features in NIC-STOI is assumed to be an appropriate predictor for speech intelligibility. Moreover, the promising results in [28], which show improvements of STOI scores for single channel enhancement over the noise signal, also support that the proposed model captures the essential features of the speech, as the estimated AR parameters and excitation variances are used in a speech production model in [28] to enhance the noisy speech with a Kalman filter.

Both the reported objective and subjective results show that the proposed method works well. The subjective results show that the proposed method can predict the intelligibility of a listening experiment over a range of 10 dB. Although the predicted values exhibit a high variance, as is to be expected of this type of experiment, this variance is similar to the one obtained with SRMR. The objective results indicate that NIC-STOI performs very well for a broad range of SNRs, even down to -30 dB SNR where the noisy speech is expected to be unintelligible. It should be noted that while NIC-STOI appears to deviate from STOI for very low SNRs, this is less important as, according to [3], a STOI score of 0.6 approximately corresponds to zero intelligibility. Even though the absolute value of STOI depends highly on the specific speech material and listening environment, the broad working range of NIC-STOI should cover the range of intelligibility. Hence, any score below this threshold can be simply assumed unintelligible. Here, it is also important to stress

that the overall aim of NIC-STOI is to have a monotonic relation with the intrusively computed STOI scores, and not necessarily to predict the absolute STOI scores. However, the offset observed between the predicted NIC-STOI scores and STOI scores in Fig. D.6 can easily be accounted for by the observed linear trend between the two measures depicted in Fig. D.4, such that the absolute STOI score can be predicted by means of the estimated NIC-STOI score.

It should be noted that STOI was among the first intrusive intelligibility metrics with very good performance, but since it was first introduced other intrusive metrics have also been proposed that show good performance. The front-end of NIC-STOI, that forms the basis of the present work, could also quite possibly be used for other intrusive frameworks, provided that they are also based on spectral features of the noisy and clean speech. Regarding this, it is interesting to note that the estimation of the parameters in short-time segments based on the current observation makes the front-end suitable for non-stationary noise conditions. However, STOI does not work well for highly non-stationary interferers due to the analysis window length. Therefore, it could be interesting to investigate using the Extended STOI (ESTOI) as a back-end to NIC-STOI instead, since this method has been developed to work well for highly modulated noise sources [44].

Correlation-based metrics including STOI are generally not suitable for predicting the intelligibility of reverberant speech and, thus, it is likely that NIC-STOI will fail in such conditions [14, 45]. Furthermore, the short time frames used in STOI might also have a negative impact on the application of NIC-STOI to reverberant speech, as short time frames cannot capture all the effects of reverberation, such as temporal smearing [14]. Currently, SRMR and ModA are the most well-studied non-intrusive intelligibility metrics. They have both been proposed for predicting the intelligibility of reverberant speech, where they both show good performance [3, 19, 20]. Even though these metrics are aimed for reverberant speech, they have also been tested for noisy and processed speech [3], where they perform reasonably well. However, it would seem that SRMR and ModA are a more suitable choice for reverberant speech, while our proposed method, NIC-STOI, which takes into account the presence of noise, is a more suitable choice for additive degradations, such as background noise and interferences. In this connection, it should also be mentioned that the proposed method is computationally much more demanding than SRMR and ModA, mainly due to the codebook search, although approximate methods for implementation of this exist [46].

In closing, we remark that the proposed method is not expected to account well for non-linear signal processing, since it is based on an additive noise model as well as the codebooks being trained on clean speech signals and noise signals. However, testing the method on the Ideal Time-Frequency Segregation (IFTS) data set from [47], which was used for evaluating the original

STOI measure [14], results in a Pearson correlation of 0.70, which is surprisingly good. For comparison, NIC-STOI outperforms the non-intrusive intelligibility metric, SRMR [3, 19], which achieves a Pearson correlation of 0.24 [7], although it should be noted that SRMR, as already mentioned, was designed for reverberant speech. However, the newly proposed Non-Intrusive STOI (NI-STOI) measure [7] achieves a Pearson correlation of 0.71 for the data set [47], which is on par with the results obtained for NIC-STOI. We remark that NI-STOI is not completely non-intrusive, as it is based on the ideal voice activity detector used in the intrusive STOI metric [7].

7 Conclusion

In this paper, a non-intrusive codebook-based short-time objective intelligibility metric, called NIC-STOI, has been proposed. It is based on an intrusive intelligibility metric, STOI, but, unlike STOI, it does not require access to the clean speech signal. Instead, the proposed method estimates the spectrum of the reference signal by identifying the entries of pre-trained spectral codebooks of speech and noise spectra, parametrized by auto-regressive parameters, which best fit the observed signal, i.e., the noisy speech signal. This is done in a statistical framework wherein parameters are estimated by minimizing the Itakura-Saito divergence for combinations of speech and noise models. This is equivalent to maximum likelihood estimation for Gaussian distributed signals. The proposed NIC-STOI metric is shown, in experiments, to be highly correlated with STOI (with a Pearson correlation of 0.94 and a standard deviation of the prediction error of 0.14) and is also validated in a listening experiment assessing speech intelligibility. Hence, it can be used for the assessment of speech intelligibility when a clean reference signal is not available. This could be used, for example, for online optimization of hearing aids.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, ser. Signal processing and communications. Taylor & Francis, 2007.

[2] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[3] 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 Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[4] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[5] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acustica United with Acta Acustica*, vol. 101, pp. 1016–1025, 2015.

[6] T. Houtgast and H. J. M. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[7] A. Heidemann Andersen, J. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *ICASSP*, March 2017, pp. 5085–5089.

[8] J. B. Allen, "Harvey Fletcher's role in the creation of communication acoustics," *J. Acoust. Soc. Am.*, vol. 99, no. 4, pp. 1825–1839, 1996.

[9] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[10] ANSI S3.5, 1997, *Methods for the Calculation of the Speech Intelligibility Index*, American National Standards Institute, New York, USA Std., 1997.

[11] 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," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, 2006.

[12] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[13] C. Ludvigsen, C. Elberling, and G. Keidser, "Evaluation of a noise reduction method comparison between observed scores and scores predicted from STI," *Scand. Audiol. Suppl.*, vol. 38, pp. 50–55, 1993.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[15] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Boston, MA: Springer US, 2005, pp. 181–197.

[16] H. Relaño-Iborra, T. May, J. Zaar, C. Scheidiger, and T. Dau, "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," *J. Acoust. Soc. Am.*, vol. 140, no. 4, p. 2670–2679, 2016.

[17] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[18] S. van Kuyk, W. B. Kleijn, and R. Hendriks, "An instrumental intelligibility metric based on information theory," in *ICASSP*, 2018.

[19] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[20] 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.

[21] F. Chen, "Modeling noise influence in speech intelligibility non-intrusively by reduced speech dynamic range," in *Interspeech*, 2016.

[22] ——, "Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation," *Biomedical Signal Processing and Control*, vol. 24, pp. 109 – 113, 2016.

[23] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-hmm-based non-intrusive speech intelligibility prediction," in *ICASSP*, March 2016, pp. 624–628.

[24] 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, 2016.

[25] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *EUSIPCO*, August 2016, pp. 1358–1362.

[26] C. Sørensen, A. Xenaki, J. Boldt, and M. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *ICASSP*, March 2017, pp. 386–390.

[27] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Non-intrusive intelligibility prediction using a codebook-based approach," in *EUSIPCO*, August 2017, pp. 226–230.

[28] M. Kavalekalam, M. Christensen, F. Gran, and J. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *ICASSP*, March 2016, pp. 191–195.

[29] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[30] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[31] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[32] J. H. L. Hansen, J. G. Proakis, and J. R. Deller, Jr., *Discrete-Time Processing of Speech Signals*. Prentice-Hall, 1987.

[33] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20(5), pp. 1644–1657, 2012.

[34] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, vol. 63(4), pp. 561–580, Apr. 1975.

[35] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[36] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[37] K. Paliwal and W. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*. Elsevier Science, 1995, pp. 433–468.

[38] A. H. Andersen, J. M. de Hann, Z.-H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE Tran. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.

[39] M. Cooke and J. Barker, "An Audio-visual corpus for speech perception and automatic speech recognition (L)," *J. Acoust. Soc. Am.*, vol. 120(5), pp. 2421–2424, Nov. 2006.

[40] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun. 1938.

[41] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language

resource for the EU," in *EUROSPEECH*, vol. 1, 18-21 September 1995, pp. 867–870.

[42] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

[43] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[44] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov 2016.

[45] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[46] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1993.

[47] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, 2009.

Paper E

Validation of The Non-Intrusive Codebook-Based Short Time Objective Intelligibility Metric for Processed Speech

Charlotte Sørensen
Jesper Bünsow Boldt
Mads Græsbøll Christensen

The paper has been presented at the
20th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 4270– 4274, Graz, Austria, 2019.

The layout has been revised.

Abstract

In recent years, objective measures of speech intelligibility have gained increasing interest. However, most speech intelligibility metrics require a clean reference signal, which is often not available in real-life applications. In a recent publication, we proposed a method, the Non-Intrusive Codebook-based Short-Time Objective Intelligibility (NIC-STOI) metric, which allows using an intrusive method without requiring access to the clean signal. The statistics of the reference signal is estimated as a combination of predefined codebooks that best fit the degraded signal by modeling the speech and noisy spectra. In this paper, we perform additional validation of the NIC-STOI in more diverse noise condition as well as for speech processed non-linearly with binary masks, where it is shown to outperform existing non-intrusive metrics.

1 Introduction

In recent years, objective measures of speech intelligibility have gained increasing interest as a tool for objectively optimizing the speech intelligibility of speech enhancement algorithms in e.g. hearing aids [1]. The articulation index (AI) [2] and the speech transmission index (STI) [3] are some of the earliest metrics that predict the intelligibility for a limited type of degradations, like linear filtering and additive noise. Recently, the speech-based envelope power spectrum model (sEPSM) [4] and the short-time objective (STOI) metric [5] were developed for more complex distortion types and are reported to have high prediction accuracy [1].

However, these metrics are all intrusive, i.e., they require a clean reference in order to predict the speech intelligibility of a degraded signal. In some scenarios, e.g., real-time processing, it is impractical to use intrusive metrics for predicting speech intelligibility. To overcome this limitation, a number of non-intrusive intelligibility prediction methods have been introduced. The Speech to Reverberation Modulation energy Ratio (SRMR) [6] and the average Modulation-spectrum Area (ModA) [7] both provide a non-intrusive estimate of the speech intelligibility based on the modulation spectrum of the degraded speech signal. Another way to predict speech intelligibility non-intrusively is to first obtain an estimate of the clean signal from its degraded version and then use this as reference to an intrusive metric. For instance, machine learning [8, 9], noise reduction [10, 11], principal component analysis [12] and neural network [13] methods have been proposed as approaches to obtain a reference signal to use inside the STOI framework from the degraded speech signal. Another non-intrusive version of the STOI metric, the non-intrusive codebook-based STOI (NIC-STOI), is proposed in [14, 15]. This is based on estimating the spectrum of the reference signal from its

degraded version by identifying combinations of pre-trained codebook entries of speech and noise spectra, parametrized by Auto-Regressive (AR) parameters, which best fit the degraded speech signal. The evaluation of the NIC-STOI metric in [15] is shown to be highly correlated with STOI and subjective listening scores for additive babble noise interference. However, since methods for predicting speech intelligibility are often used to evaluate the effects of non-linear processing, a method that is also suitable for such types of processing is desirable [1]. Therefore, in this paper, the NIC-STOI metric is further validated on speech in different noise conditions, which has been non-linearly processed with Ideal Binary Masks (IBMs) [16].

2 The NIC-STOI metric

The NIC-STOI metric, proposed in [14, 15], is based on STOI but does not require access to a clean reference signal. Figure C.1 depicts an overview of the NIC-STOI algorithm. The algorithm consists of three main steps: 1) estimation of the AR speech and noise model parameters 2) computation of the clean and noisy time-frequency spectra 3) prediction of intelligibility within STOI. In the following, a condensed description of the NIC-STOI metric is presented. A more thorough description is available in [15].

2.1 Step 1: Estimate parameters

It is assumed that a speech and noise signal are random uncorrelated processes such that the noisy speech signal is given by $y(n) = s(n) + w(n)$ [17, 18]. The speech and noise are modeled as stochastic AR processes expressed as $u(n) = \mathbf{a}_s^T \mathbf{s}(n)$ and $v(n) = \mathbf{a}_w^T \mathbf{w}(n)$, respectively, where $\mathbf{s}(n) = [s(n), s(n-1), \ldots, s(n-P)]^T$ and $\mathbf{w}(n) = [w(n), w(n-1), \ldots, w(N-Q)]^T$ are vectors collecting the P and Q past samples, $\mathbf{a}_s = [1, a_s(1), a_s(2), \ldots, a_s(P)]^T$ and $\mathbf{a}_w = [1, a_w(1), a_w(2) \ldots, a_w(Q)]^T$ are vectors containing the AR parameters with $a_s(0) = 1$ and $a_w(0) = 1$. Finally, $u(n)$ and $v(n)$ models the speech and noise excitations as zero mean white Gaussian noise with excitation variance σ_u^2 and σ_v^2, respectively.

The parameters to be estimated, i.e., the speech and noise AR coefficients and excitation variances are given by the vector $\theta = [\mathbf{a}_s;\ \mathbf{a}_w;\ \sigma_u^2(n);\ \sigma_v^2(n)]$. Using Bayes' theorem, the minimum mean square error (MMSE) estimate given N noisy samples, i.e., $\mathbf{y} = [\ y(0)\ y(1)\ \ldots\ y(N-1)\ ]$ can be given by [17–19]:

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \text{E}(\boldsymbol{\theta}|\mathbf{y}) = \int_{\Theta} \theta \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} d\theta, \tag{E.1}$$

where Θ denotes the support space to be estimated.

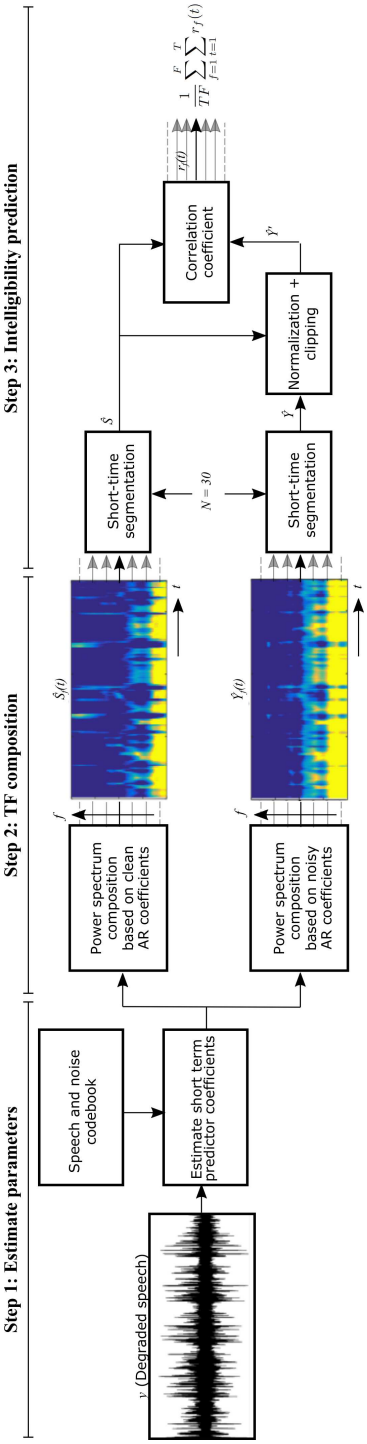


Fig. E.1: A block diagram of the NIC-STOI metric adapted from [14, 15]. Using a codebook-based approach the speech and noisy spectra are jointly modeled from pre-trained codebooks of both speech and noise and are then used within STOI.

The excitation variances are estimated through a maximum likelihood (ML) approach by limiting the AR parameters $\mathbf{a}_s$ and $\mathbf{a}_w$ to predefined codebooks of size N_s and N_w:

$$\{\sigma_{u,ij}^{2,\mathrm{ML}}, \sigma_{v,ij}^{2,\mathrm{ML}}\} = \underset{\sigma_u^2,\sigma_v^2}{\arg\max} \log p(\mathbf{y}|\mathbf{a}_{s_i}^{\mathrm{CB}};\ \mathbf{a}_{w_j}^{\mathrm{CB}};\ \sigma_u^2;\ \sigma_v^2),$$

where $\mathbf{a}_{s_i}^{\mathrm{CB}}$ and $\mathbf{a}_{w_j}^{\mathrm{CB}}$ are the ith and j^{th} entry of the speech and noise codebook, respectively. The Gaussian likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is given by:

$$p(\mathbf{y}|\mathbf{a}_{s_i}^{\mathrm{CB}};\ \mathbf{a}_{w_j}^{\mathrm{CB}};\ \sigma_{u,ij}^2;\ \sigma_{v,ij}^2) \propto e^{-d_{\mathrm{IS}}(P_y(\omega),\hat{P}_y^{ij}(\omega))}, \tag{E.2}$$

where $d_{\mathrm{IS}}(\cdot,\cdot)$ is the Itakura-Saito divergence between the observed, $P_y(\omega)$, and modeled, $\hat{P}_y^{ij}(\omega)$, noisy spectrum expressed as [18, 20]:

$$d_{\mathrm{IS}}(P_y(\omega), \hat{P}_y^{ij}(\omega)) = \frac{1}{2\pi}\int_{\Psi}\left(\frac{P_y(\omega)}{\hat{P}_y^{ij}(\omega)} - \ln\left(\frac{P_y(\omega)}{\hat{P}_y^{ij}(\omega)}\right) - 1\right)d\omega, \tag{E.3}$$

where $\hat{P}_y^{ij}(\omega) = \frac{\sigma_u^2}{|A_s^i(\omega)|^2} + \frac{\sigma_v^2}{|A_w^j(\omega)|^2}$, and A_s^i and A_w^j are the ith and j^{th} entry from the speech codebook and noise codebook, respectively. The support space, Ψ, excludes values below a threshold in order to disregard time-frequency units with low energy or where the binary masks renders the presented signal inaudible. This threshold is here set to 40 dB below peak energy.

Finally, (E.1) is computed from its discrete counterpart:

$$\hat{\boldsymbol{\theta}} = \frac{1}{N_s N_w}\sum_{i=1}^{N_s}\sum_{j=1}^{N_w}\boldsymbol{\theta}_{ij}\frac{p(\mathbf{y}|\boldsymbol{\theta}_{ij})}{p(\mathbf{y})} \tag{E.4}$$

and

$$p(\mathbf{y}) = \frac{1}{N_s N_w}\sum_{i=1}^{N_s}\sum_{j=1}^{N_w} p(\mathbf{y}|\boldsymbol{\theta}_{ij}), \tag{E.5}$$

where $\boldsymbol{\theta}_{ij} = [\mathbf{a}_{s_i}^{\mathrm{CB}};\ \mathbf{a}_{w_j}^{\mathrm{CB}};\ \sigma_{u,ij}^{2,\mathrm{ML}};\ \sigma_{v,ij}^{2,\mathrm{ML}}]$. The priors in (E.1) are non-informative, since the codebook entries and the ML excitation variance estimates contribute with equal probability and are, thus, omitted.

2.2 Step 2: TF composition

Using the estimated parameters, $\hat{\boldsymbol{\theta}}$, from (E.4) the Time-Frequency (TF) spectrum of the estimated speech and noise signal are given by:

$$\hat{P}_s(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2}, \tag{E.6}$$

where $\hat{A}_s(\omega) = \sum_{k=0}^{P} \hat{a}_s(k) e^{-j\omega k}$, and

$$\hat{P}_w(\omega) = \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}, \tag{E.7}$$

where $\hat{A}_w(\omega) = \sum_{k=0}^{Q} \hat{a}_w(k) e^{-j\omega k}$. The shape of the envelope of the estimated signals are given by the AR parameters, i.e., $\hat{\mathbf{a}}_s$ and $\hat{\mathbf{a}}_w$, while the overall signal power is given by the excitation variances, i.e., $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$. Then, the noisy spectrum is given by the sum of the speech and noise power spectra:

$$\hat{P}_y(\omega) = \hat{P}_s(\omega) + \hat{P}_w(\omega) = \frac{\hat{\sigma}_u^2}{|\hat{A}_s(\omega)|^2} + \frac{\hat{\sigma}_v^2}{|\hat{A}_w(\omega)|^2}. \tag{E.8}$$

2.3 Step 3: Intelligibility Prediction

The estimated speech and noise TF spectra, i.e., $\hat{P}_s(\omega)$ (E.6) and $\hat{P}_y(\omega)$ (E.8), are then used as inputs in the original STOI metric as replacement for the discrete Fourier transform of the clean and noisy signal, respectively.

The TF spectra $\hat{P}_s(\omega)$ and $\hat{P}_y(\omega)$ are grouped into 15 one-third octave bands and short-time regions of 384 ms denoted by $\overline{\mathbf{p}}_s(f,t)$ and $\overline{\mathbf{p}}_y(f,t)$ as in the original STOI implementation [5]. In order to de-emphasize the impact of noise dominated regions $\overline{\mathbf{p}}_y(f,t)$ are clipped by a normalization procedure:

$$\overline{\mathbf{p}}'_y(f,t) = \min\left(\frac{\|\overline{\mathbf{p}}_s(f,t)\|_2}{\|\overline{\mathbf{p}}_y(f,t)\|_2} \overline{\mathbf{p}}_y(f,t), (1 + 10^{-\beta/20}) \overline{\mathbf{p}}_s(f,t) \right)$$

where $\|\cdot\|_2$ is the l_2 norm and $\beta = -15$ dB is the lower signal-to-distortion ratio. The local correlation coefficient between $\overline{\mathbf{p}}'_y(f,t)$ and $\overline{\mathbf{p}}_s(f,t)$ is computed as

$$r(f,t) = \frac{(\overline{\mathbf{p}}_s(f,t) - \mu_{\overline{\mathbf{p}}_s(f,t)})^T (\overline{\mathbf{p}}'_y(f,t) - \mu_{\overline{\mathbf{p}}'_y(f,t)})}{\sqrt{(\overline{\mathbf{p}}_s(f,t) - \mu_{\overline{\mathbf{p}}_s(f,t)})^2} \sqrt{(\overline{\mathbf{p}}'_y(f,t) - \mu_{\overline{\mathbf{p}}'_y(f,t)})^2}},$$

where $\mu(\cdot)$ is the mean of the vector. Finally, the NIC-STOI intelligibility prediction is given by averaging the correlation coefficient, i.e. $r(f,t)$, across all bands and frames as

$$d_{NS} = \frac{1}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} r(f,t). \tag{E.9}$$

3 Experimental Details

In order to further validate the performance of the NIC-STOI metric presented in [14, 15] we here evaluate it on the same data as in the original

Table E.1: Performance of the intelligibility metrics in terms of Pearson's correlation (ρ), Kendall's tau (τ) and the standard deviation of the prediction error (RMSE).

Condition	ρ	τ	RMSE
STOI [5]	0.955	0.821	8.9 %
NIC-STOI [15]	0.940	0.791	11.4 %
NI-STOI[1] [12]	0.711	0.529	25.2 %
SRMR-norm [21]	0.392	0.155	38.4 %
SRMR [6]	0.235	0.034	45.0 %

paper on STOI [5, 16]. Subjective intelligibility scores have been obtained from 15 normal hearing subjects. Stimuli were the Dantale II sentence material [22] mixed with four different noise types: bottling factory hall noise, cafe noise, Speech Shaped Noise (SSN) and car noise at three different Signal to Noise Ratios (SNRs). The noisy signals were processed with IBMs at eight different Relative Criterion (RC) values, which determines the density of the computed binary mask [16]. Materials from 5 subjects is used to train the codebooks, whereas the results from the remaining 10 subjects are used for testing. The data and experimental details are described in detail in [16].

The speech and noise AR parameters and variances are estimated from 25.6 ms frames windowed using a Hann window with 50% overlap. Over these short time frames the estimated parameters are assumed to be stationary. The signals were resampled to 10 kHz as in the original STOI metric. The speech and noise AR model order P and Q, respectively, are set to 14 according to literature [17–19]. The speech codebook is trained on 50 clean speech sentences from the Dantale II data set not included in the training set using the generalized Lloyd algorithm (GLA) [17, 23]. The noise codebook is trained on 50 sentences of each noise type condition without IBM processing concatenated into a single vector. The sizes of the speech and noise codebooks are $N_s = 64$ and $N_w = 8$, respectively. The support space of the Itakura-Saito divergence, Ψ, is computed by taking the Fourier transform of the input signal and limiting the dynamic range to 40 dB from the highest value such that TF units below this threshold are not included in the calculation. In order to reduce intra- and intersubject variability the results are condition-averaged per noise and SNR combination and are then mapped to subjective performance across all conditions [1]. The performance of the metric is evaluated using Pearson's correlation (ρ) which gives the linear relationship, Kendall's tau (τ) which gives the ranking capability and the root mean square error (RMSE).

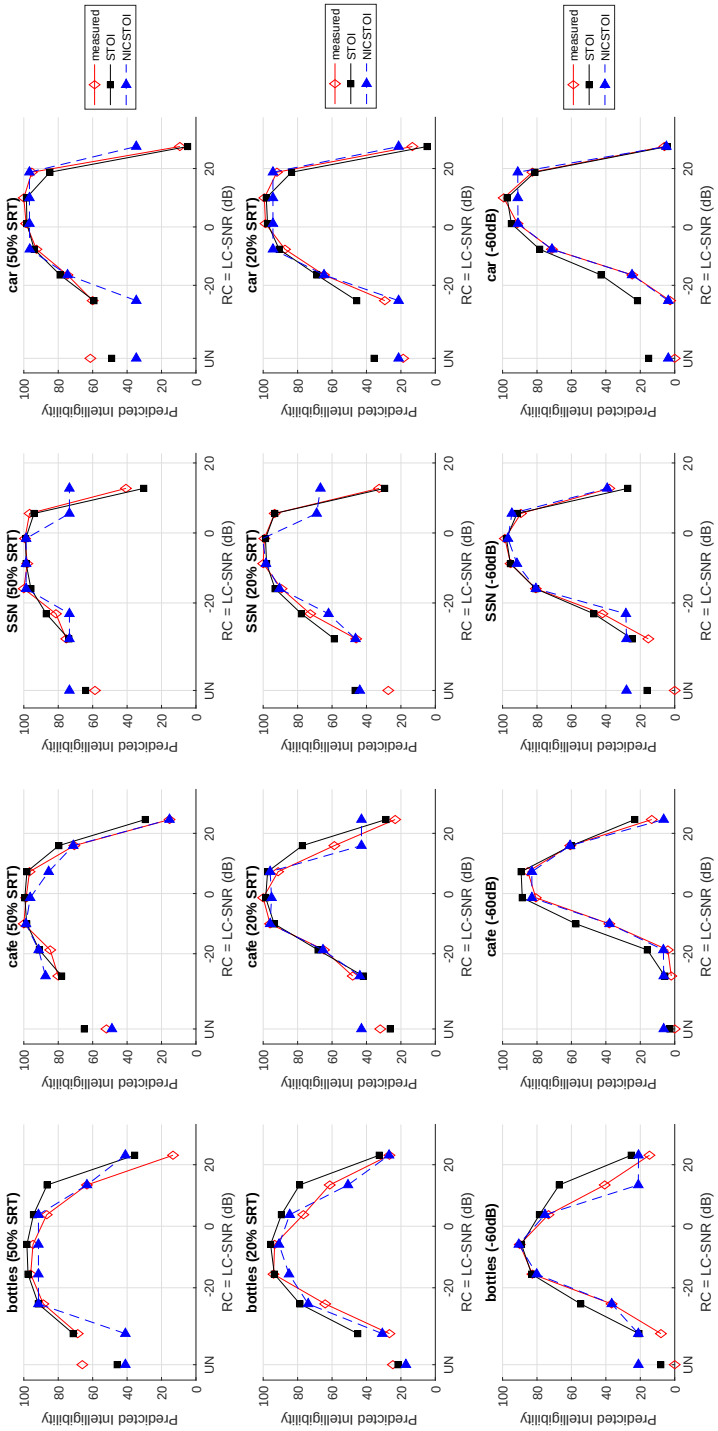


Fig. E.2: NIC-STOI (blue) predictions evaluated against STOI (black) predictions and subjective (red) intelligibility scores. Each row refers to the three different SNRs used in the data set (50 % SRT, 20 % SRT and -60 dB) with the the first row corresponding to the highest SNR and the last row the lowest SNR. The columns refer to the four different noise types (bottling factory hall, cafe, Speech Shaped Noise (SSN) and car noise). The x-axis refers to the Relative Criterion (RC) values, which determines the density of the computed Ideal Binary Masks (IBM). "UN" refers to unprocessed conditions.

4 Results and Discussion

The performance of NIC-STOI is depicted in Fig E.2 (blue) against measured subjective scores (red) and the original intrusive STOI metric (black). It can be observed that NIC-STOI is highly correlated with the subjective scores across all noise conditions. Furthermore, NIC-STOI is also highly correlated with STOI, which supports the earlier findings in [14, 15]. The highest deviation can be observed for the SSN noise condition, which can perhaps be explained by the noise codebook weighting this condition less when being trained on all the noise conditions concatenated.

In Table E.1, NIC-STOI is evaluated against existing state of the art intelligibility metrics. The best performance is obtained by the intrusive metric STOI. However, even though NIC-STOI is non-intrusive, it comes close to being on par with the performance of STOI. NIC-STOI is compared to three other non-intrusive metrics: NI-STOI [12], SRMR [6] and SRMR-norm [21]. The results from NI-STOI are obtained from [12], since it was not possible to obtain an implementation, while implementations of the latter two are publicly available. The NI-STOI metric is aimed to predict the intelligibility of non-linearly processed speech, while the SRMR metric and the improved SRMR-norm are aimed to predict the intelligibility of reverberated speech, but have successfully been applied for noisy and processed speech [1]. As shown in Table E.1, NIC-STOI outperforms all three existing non-intrusive intelligibility metrics. It should, however, be noted that NI-STOI is only trained using clean speech material [12] and SRMR and SRMR-norm is not trained at all. The cafe noise condition is primarily composed by a single interfering speaker such that additional information is needed in order to determine, which speaker is the target. NIC-STOI is trained with both clean speech and noise material, which makes it able to account for the cafe condition. Excluding the cafe noise condition in NI-STOI, NIC-STOI still has the best performance ($\rho = 0.940$, $\tau = 0.791$, RMSE $= 11.4\%$) even though the performance of NI-STOI comes close to that of NIC-STOI ($\rho = 0.907$, $\tau = 0.777$, RMSE $= 13.9\%$) [12].

5 Conclusion

In this paper, the Non-Intrusive Codebook-based Short-Time Objective Intelligibility metric, NIC-STOI, has been investigated more thoroughly on a large data set with subjective scores in diverse noise conditions. NIC-STOI non-intrusively estimates the spectrum of a reference signal from its degraded version and uses this as input to an intrusive intelligibility metric, STOI. The reference signal is estimated as combinations of entries from pre-trained speech and noise spectral codebooks, parametrized by auto-regressive pa-

rameters, which best fit the degraded signal by minimizing the Itakura-Saito divergence. In order to account for binary mask processing a small adjustment of NIC-STOI is implemented in which only time-frequency units above a certain threshold is included in the Itakura-Saito divergence. The NIC-STOI metric is highly correlated with subjective intelligibility scores on the non-linearly processed speech data set and outperforms existing non-intrusive metrics.

References

[1] 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 Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[2] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[3] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[4] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[6] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[7] 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.

[8] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-hmm-based non-intrusive speech intelligibility prediction," in *ICASSP*, March 2016, pp. 624–628.

[9] 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, 2016.

[10] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *EUSIPCO*, August 2016, pp. 1358–1362.

[11] C. Sørensen, A. Xenaki, J. Boldt, and M. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *ICASSP*, March 2017, pp. 386–390.

[12] A. Heidemann Andersen, J. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *ICASSP*, March 2017, pp. 5085–5089.

[13] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 26, no. 10, pp. 1925–1939, 2018.

[14] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Non-intrusive intelligibility prediction using a codebook-based approach," in *EUSIPCO*, August 2017, pp. 226–230.

[15] ——, "Non-intrusive codebook-based intelligibility prediction," *Speech Communication*, vol. 101, pp. 85 – 93, 2018.

[16] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, p. 1415–1426, 2009.

[17] M. Kavalekalam, M. Christensen, F. Gran, and J. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *ICASSP*, March 2016, pp. 191–195.

[18] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[19] ——, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, 2006.

[20] K. Paliwal and W. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis*. Elsevier Science, 1995, pp. 433–468.

[21] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *IWAENC*, Sept. 2014, p. 55–59.

[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, 2003.

[23] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

Paper F

Harmonic Beamformers for Non-Intrusive Speech Intelligibility Predictions

Charlotte Sørensen
Jesper Bünsow Boldt
Mads Græsbøll Christensen

The paper has been presented at the
20th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 4260–4264, Graz, Austria, 2019.

The layout has been revised.

Abstract

In recent years, research into objective speech intelligibility measures has gained increased interest as a tool to optimize speech enhancement algorithms. While most intelligibility measures are intrusive, i.e., they require a clean reference signal, this is rarely available in real-time applications. This paper proposes two non-intrusive intelligibility measures, which allow using the intrusive short-time objective intelligibility (STOI) measure without requiring access to the clean signal. Instead, a reference signal is obtained from the degraded signal using either a fixed or an adaptive harmonic spatial filter. This reference signal is then used as input to STOI. The experimental results show a high correlation between both proposed non-intrusive speech intelligibility measures and the original intrusively computed STOI scores.

1 Introduction

Speech intelligibility is an important property to consider, when developing signal processing for a wide range of applications, e.g., telecommunications [1, 2], and hearing aids [3]. As such, research into using objective measures of speech intelligibility as a tool to optimize speech enhancement algorithms has gained increased interest in recent years. There exists numerous different measures to estimate speech intelligibility with acceptable accuracy. The articulation index (AI) [4] and the speech transmission index [1] are some of the earliest measures to predict speech intelligibility of limited types of degradations such as linear filtering and additive noise. The speech-based envelope power spectrum model (sEPSM) [5] and the short-time objective intelligibility (STOI) [6] measure were recently introduced in order to increase the prediction accuracy for more complex degradation types. All the aforementioned measures are intrusive, i.e., in addition to the degraded signal they require access to a clean reference signal, which is rarely available in real-time applications.

This limitation has led to the proposal of non-intrusive speech intelligibility prediction measures, which do not require access to a clean reference signal. The speech to reverberation modulation energy ratio (SRMR) [7] provides an intelligibility prediction of reverberated speech based on the ratio between the energy of the low and high modulation frequency content. Similarly, the average modulation-spectrum area (ModA) [8] measure provides an intelligibility prediction based on the area of modulation spectrum of the degraded signal. Both these measures have been shown to perform well for conditions such as reverberation and additive noise compared to the previous non-intrusive measures [3, 7, 8].

Another approach to predict the speech intelligibility non-intrusively is to exploit a well-established and reliable intrusive metric, e.g. STOI [6], and

obtaining an estimate of the clean speech reference from degraded signal. Recently, different approaches to estimate the reference signal have been proposed using machine learning [9, 10], spectral codebooks [11, 12], principal component analysis [13] and neural network [14] methods. These approaches have been shown to outperform the existing non-intrusive speech intelligibility prediction measures and to have a comparable performance to the intrusive measures [9, 12–14]. However, since these methods are all single channel and non-intrusive, they have no way of determining which speech signal is the desired target if multiple speakers are present given that the model is not trained for the specific speaker.

Using a multi-channel approach such as spatial filtering, i.e. beamforming, offers the possibility to overcome this limitation with a non-intrusive approach given the direction of the desired speech signal as proposed in [15]. The advantage of this method is that it has a very low complexity such that it can run on applications with low computational power, e.g. a hearing aid. On the other hand, the performance deteriorates with increasing number of interferers and reverberation. Similarly, the pitch-based STOI (PB-STOI) [16] measure also exploits the spatial content but instead of a filtering approach it reconstructs the reference signal from estimates of the properties of the signal model of the clean signal. It is based on a spatio-temporal model, which assumes the desired signal to be a sum of sinusoids whose frequencies are integral multiples of the pitch. Combining the spatial and the temporal characteristics (i.e., the direction of the desired signal its pitch) makes it more robust to competing speakers and reverberation, since it is possible to follow the pitch of the desired speech signal. PB-STOI has been shown to have a high correlation with the intrusive STOI scores even under adverse conditions with multiple interferers. However, the method also requires more computational power than the beamforming-based approach.

The present paper proposes new solutions to non-intrusive speech intelligibility prediction using, respectively, a fixed and an adaptive harmonic spatial filter based on a combination of the principles in [15, 16]. More specifically, the reference signal to be used as input to the intrusive framework STOI is obtained using model-based harmonic beamforming that resembles a filterbank designed for the given spatial and spectral characteristics of the desired signal. The rationale behind this approach is that the most energetic spectro-temporal regions, i.e. glimpses, occur during the voiced, i.e. harmonic, parts of speech. According to the glimpses model, intelligibility is related to the presence of such glimpses in which the most energetic regions are most important for speech intelligibility [17]. It is shown that the number of such glimpses correlates well with measured intelligibility and, thus, is a promising predictor for speech intelligibility [17].

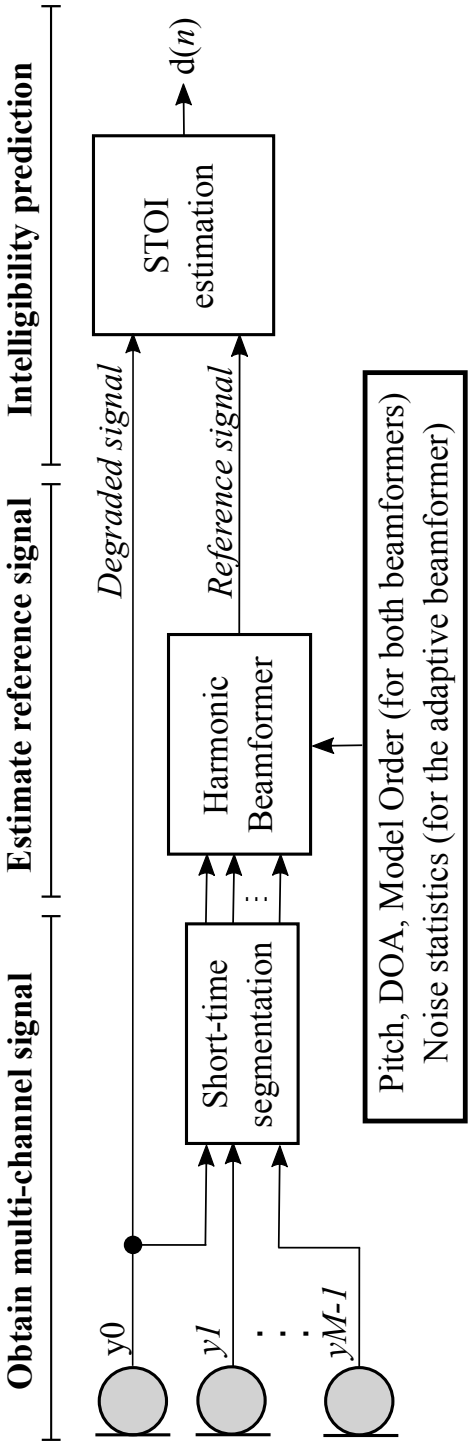


Fig. F1: Block diagram of the proposed harmonic beamforming-based non-intrusive intelligibility measures in which a reference signal is obtained using a harmonic spatial filter and compared with the output of an omnidirectional microphone with the intrusive STOI measure.

2 Methods

This section presents the principles behind the proposed non-intrusive speech intelligibility predictions measures based on a fixed harmonic beamformer, dubbed the harmonic delay-and-sum beamformer-based STOI (HDSB-STOI), and an adaptive harmonic beamformer, dubbed the harmonic wiener beamformer-based STOI (HWB-STOI). Both HDSB-STOI and HWB-STOI allows predicting the speech intelligibility non-intrusively, i.e., without requiring access to the clean reference signal, by obtaining a reference signal from the degraded signal using a harmonic spatial filter and use this as input to STOI. Figure F.1 depicts the general structure of both methods, which consists of three main steps: 1) Obtain a multi-channel signal using a microphone array, 2) Estimate a reference signal with the harmonic spatial filters and 3) Predict the speech intelligibility within the STOI framework.

2.1 Fundamentals

In the following, the signal model and the associated assumptions of the proposed methods are presented based on [18] in which a more thorough description of the theory behind the harmonic beamformers is available. In the proposed methods it is assumed that a uniform linear array (ULA) consisting of M microphones obtains the desired speech signal added to a mixture of interfering background noise and reverberation such that the samples of the mth microphone observations in a vector of frame length L is given by:

$$\mathbf{y}_m(t) = \mathbf{x}_m(t) + \mathbf{v}_m(t), \tag{F.1}$$

where $\mathbf{x}_m(t)$ and $\mathbf{v}_m(t)$ are vectors containing samples of the desired signal and the inference at the mth microphone, respectively.

The desired speech signal, $x_m(t)$, is modeled as a sum of sinusoids, i.e., a harmonic signal model, which is a good model for the voiced speech segments. Furthermore, using the harmonic model to obtain the desired signal does not only reduce the interfering sources but also reverberation, since spectral and temporal smearing of the signal source due to reverberation is not included in the harmonic model. As such, the desired speech signal is modeled as [18, 19]:

$$\mathbf{x}_m(t) = \mathbf{D}_{m,N}(\theta, \omega_0)\mathbf{a}(t, \omega_0), \tag{F.2}$$

where $\mathbf{D}_{m,N}(\theta, \omega_0)$ is a $L \times 2N$ matrix with the nth column being a vector of length L given by

$$\mathbf{d}_{m,n}(\theta, \omega_0) = e^{-jn\omega_0 f_s \tau_m(\theta)} \times \left[1 \;\; e^{-jn\omega_0} \;\; \dots \;\; e^{-jn\omega_0(L-1)} \right]^T, \tag{F.3}$$

where the superscript ^{T} is the transpose operator, N is the model order, $\jmath = \sqrt{-1}$ is the imaginary unit, ω_0 is the pitch or fundamental frequency, f_s is the sampling frequency and $\tau_m(\theta)$ is the relative delay of the desired source on the ULA. Furthermore, the complex amplitude $\mathbf{a}(t, \omega_0)$ is a vector of length $2N$ given by:

$$\mathbf{a}(t, \omega_0) = [\ a_{-N} e^{-\jmath N \omega_0 t} \quad a_{-N+1} e^{-\jmath (N-1) \omega_0 t} \quad \cdots \quad a_N e^{\jmath N \omega_0 t}\ ]^T, \tag{F.4}$$

and the correlation matrix of $\mathbf{a}$ (of size $2N \times 2N$) is

$$\mathbf{R_a} = \text{diag}\left(E\left[|a_{-N}|^2\right], E\left[|a_{-N+1}|^2\right], \ldots, E\left[|a_N|^2\right]\right), \tag{F.5}$$

and

$$\mathbf{R}_{\underline{\mathbf{v}}} = E\left[\underline{\mathbf{v}}(t)\underline{\mathbf{v}}^H(t)\right], \tag{F.6}$$

where $E[\cdot]$ is the mathematical expectation, and the superscript ^{H} is the conjugate-transpose operator.

Concatenating all the microphone signal vectors gives the vector of length ML:

$$\underline{\mathbf{y}}(t) = \underline{\mathbf{D}}_N(\theta, \omega_0)\mathbf{a}(t, \omega_0) + \underline{\mathbf{v}}(t), \tag{F.7}$$

where $\underline{\mathbf{y}}(t) = [\ \mathbf{y}_1^T(t) \quad \mathbf{y}_2^T(t) \quad \cdots \quad \mathbf{y}_M^T(t)\ ]^T$,
$\underline{\mathbf{v}}(t) = [\ \mathbf{v}_1^T(t) \quad \mathbf{v}_2^T(t) \quad \cdots \quad \mathbf{v}_M^T(t)\ ]^T$ and

$$\underline{\mathbf{D}}_N(\theta, \omega_0) = \begin{bmatrix} \mathbf{D}_{1,N}(\theta, \omega_0) \\ \mathbf{D}_{2,N}(\theta, \omega_0) \\ \vdots \\ \mathbf{D}_{M,N}(\theta, \omega_0) \end{bmatrix}. \tag{F.8}$$

2.2 Harmonic delay-and-sum beamformer-based STOI (HDSB-STOI)

The harmonic delay-and-sum beamformer (DSB) is a fixed beamformer, which cannot adjust to the spatial characteristics of the interfering noise. It is advantageous for applications such as hearing aids, since it only requires low computational power and does not require estimates of the noise statistics but, at least in theory, comes at a cost in performance [18]. The DSB can be deduced by maximizing the white noise gain subject to the distortionless constraint:

$$\min_{\underline{\mathbf{h}}} \underline{\mathbf{h}}^H \underline{\mathbf{h}} \quad \text{subject to} \quad \underline{\mathbf{h}}^H \underline{\mathbf{D}}_N(\theta_0, \omega_0) = \mathbf{1}_{2N}^T, \tag{F.9}$$

where $\mathbf{1}_{2N} = [\,1\ \ 1\ \cdots\ 1\,]^T$ is a vector of length $2N$.

Then, the DSB is derived as the optimal solution given by:

$$\underline{\mathbf{h}}_{\text{HDSB}} = \underline{\mathbf{D}}_N(\theta_0, \omega_0)\left[\underline{\mathbf{D}}_N^H(\theta_0, \omega_0)\underline{\mathbf{D}}_N(\theta_0, \omega_0)\right]^{-1}\mathbf{1}_{2N}. \tag{F.10}$$

2.3 Harmonic Wiener beamformer-based STOI (HWB-STOI)

The harmonic Wiener beamformer is an adaptive beamformer that can adapt to the spatial characteristics of the interfering noise, which in theory should give a better performance than fixed beamformers. However, it also needs access to the noise statistics and requires more computational power than the fixed beamformer.

The harmonic Wiener beamformer can be derived using the mean square error (MSE), which is given by [18]:

$$\begin{aligned} J(\underline{\mathbf{h}}) = &\ E\left[|e(t)|^2\right] && \text{(F.11)}\\ = &\ \mathbf{1}_{2N}^T\mathbf{R}_{\mathbf{a}}\mathbf{1}_{2N} && \text{(F.12)}\\ &+ \underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta_0, \omega_0)\underline{\mathbf{h}}\\ &- \underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\mathbf{1}_{2N}\\ &- \mathbf{1}_{2N}^T\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta_0, \omega_0)\underline{\mathbf{h}} + \underline{\mathbf{h}}^H\mathbf{R}_{\underline{\mathbf{v}}}\underline{\mathbf{h}}, \end{aligned}$$

where the error signal between the estimated and desired signal, $e(t) = \left[\underline{\mathbf{h}}^H\underline{\mathbf{D}}_N(\theta_0, \omega_0) - \mathbf{1}_{2N}^T\right]\mathbf{a}(t, \omega_0) + v_{\text{rn}}(t)$, is the sum of the signal distortion and the residual noise.

Finally, the optimal solution for the harmonic Wiener beamformer can be found by differentiating the MSE, $J(\underline{\mathbf{h}})$ [eq. (F.11)], with respect to $\underline{\mathbf{h}}$ and setting the result equal to zero:

$$\begin{aligned} \underline{\mathbf{h}}_{\text{HWB}} = &\left[\underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\underline{\mathbf{D}}_N^H(\theta_0, \omega_0) + \mathbf{R}_{\underline{\mathbf{v}}}\right]^{-1} \times\\ &\underline{\mathbf{D}}_N(\theta_0, \omega_0)\mathbf{R}_{\mathbf{a}}\mathbf{1}_{2N}. \end{aligned} \tag{F.13}$$

2.4 Pitch-based STOI (PB-STOI)

The results of the proposed HDSB-STOI and HWB-STOI are compared with the non-intrusive PB-STOI measure proposed in [16], where a more detailed description is available. Similar to the proposed methods in this paper, PB-STOI is based on a harmonic model that takes the spatial input into account:

$$\mathbf{y}_m = \beta_m\mathbf{Z}\mathbf{D}(m)\boldsymbol{\alpha} + \mathbf{v}_m, \tag{F.14}$$

where β_m is the attenuation of the desired source at the m'th microphone, $\mathbf{Z} = [\mathbf{z}(\omega_0)\ldots\mathbf{z}(L\omega_0)], \mathbf{z}(l\omega_0) = [1\ldots e^{jl\omega_0(N-1)}]$, $\mathbf{D}(m) =$

$\text{diag}([e^{-j\omega_0 f_s \tau_k} \dots e^{-jL\omega_0 f_s \tau_m}])$ for $l = 1, \dots, L$ with all other entries equal to zero and $\mathbf{v}_m$ is the sum of the recorded noise and interference.

Based on the signal model, the attenuation factor, the complex amplitude, the variance and the pitch is estimated in an iterative manner. These parameters are then used to directly to reconstruct a reference signal from the signal model. Finally, the reconstructed reference signal is applied as input to STOI instead of the clean signal.

3 Experimental results

The proposed measures are evaluated using a broadside ULA setup consisting of $M = 10$ omnidirectional microphones with a microphone spacing of $d = c/f_s$, where the speech of sound in air is $c = 343$ m/s and the sampling frequency is $f_s = 8$ kHz. The direction of arrival (DOA) of the desired source was $\theta = 0°$ resulting in a $\tau_m = 0$. The pitch is efficiently estimated using the multi-channel maximum likelihood pitch estimator proposed in [20] as the sum over the squared magnitude of the FFT of $y_m(t)$, denoted $Y_m(\omega_0)$, evaluated at a set of candidate harmonics, Ω_0, which is given by $\hat{\omega}_0 = \arg\max_{\omega_0 \in \Omega_0} \sum_{l=1}^{L} \sum_{m=1}^{M} |Y_m(\omega_0 l)|^2$ when assuming that the DOA is coming from the front, the noise variance is known and the same for all channels. The pitch is evaluated in the range $\Omega_0 = 80 - 400$ Hz and the model order, $L = 10$. In the experimental evaluation, a set of 50 English sentences (both male and female) from the EUROM_1 database [21] is used for both the desired source and interfering speakers. The sentences contain both voiced and unvoiced segments. The signals are 5.0 s long and are processed in segments of 20 ms with 50 % overlap. The toolbox McRoomSim [22] is used to create the simulations of a complex multi-talker scenario with 8 interfering speakers in a room with dimensions of 10x6x4 m similar to the evaluation setup in [16]. The simulations are carried out at three different levels of reverberation ranging from low to high (RT60 = 0.3 s, RT60 = 0.6 s and RT60 = 1.5 s) at signal-to-noise ratios (SNRs) ranging from -15 to 5 dB. A white Gaussian noise is added to each microphone channel at a SNR of 20 dB.

The performances of the proposed non-intrusive intelligibility measures are evaluated against the original intrusively computed STOI scores as the ground truth. The results are shown in Figures F.2(a), F.2(b) and F.2(c) for the low, medium and high reverberation scenarios, respectively. The results of the PB-STOI (red squares), HWB-STOI (yellow diamonds) and HDSB-STOI (purple triangles) are plotted together with the intrusively computed STOI scores indicated by the blue circles. At low reverberation all of the three non-intrusive measures show a good performance even though the harmonic beamforming-based non-intrusive speech intelligibility measures both out-

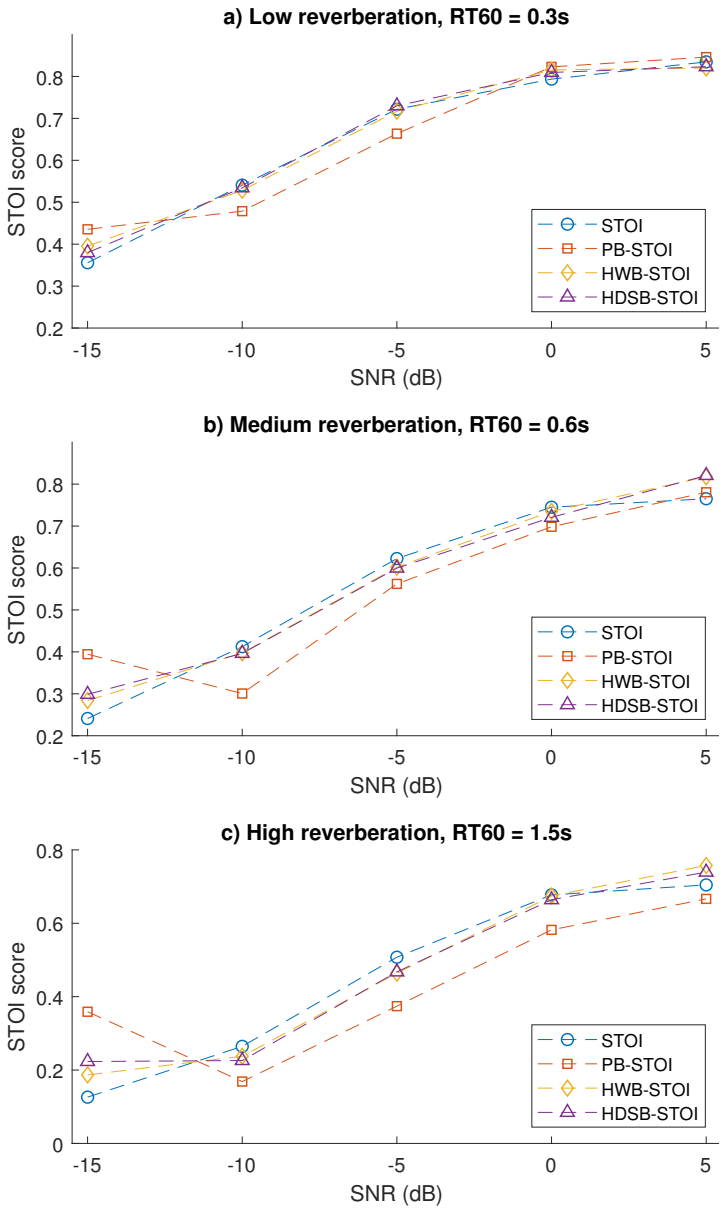


Fig. F.2: Performance shown in terms of estimated STOI score as function of the SNR in dB for a) Low reverberation with RT60 = 0.3 s, b) Medium reverberation with RT60 = 0.6 s and c) High reverberation with RT60 = 1.5 s. The results of STOI, PB-STOI, HWB-STOI and HDSB-STOI is given by the blue circles, red squares, yellow diamonds and purple triangles, respectively.

perform the PB-STOI measure. As reverberation increases, the performance of PB-STOI deteriorates and especially at low SNRs it predicts an increase in speech intelligibility rather than a decrease in intelligibility as predicted by the intrusive STOI measure. It is obvious that both of the proposed HDSB-STOI and HWB-STOI measures yield more suppression of interfering speakers and reverberation compared with the PB-STOI measure even though there is a slight decrease in performance at low SNRs with increasing reverberation levels for both measures. While the PB-STOI measure is also based on a harmonic model and should, thus, also perform well in reverberation, the difference in performance is likely due to PB-STOI being more sensitive to errors in the pitch estimate, since it is based on a reconstruction of the reference signal rather than a filtering approach.

Notably, the measures based on the fixed and adaptive approach perform almost equally well. The performance of the adaptive Wiener beamformer is only slightly better at low SNRs at high reverberation levels. Even though the adaptive beamformer in theory should have a better performance this is not necessarily the case in practical performance, since it relies on estimates of the noise statistics. This is also supported by the findings in [18], where the adaptive beamformers provide a slightly lower SNR gain compared to the harmonic DSB. As such, due to being computational efficient and simple, i.e. not requiring noise statistics, the HDSB-STOI measure might be the best choice depending on the applications, e.g. hearing aids, given the comparable performance to the HWB-STOI measure.

4 Conclusions

This paper proposes two approaches, the harmonic delay-and-sum beamformer-based STOI (HDSB-STOI) and the harmonic wiener beamformer-based STOI (HWB-STOI), for non-intrusive prediction of speech intelligibility. The HDSB-STOI measure and the HWB-STOI measure estimate a reference signal from the degraded signal using a fixed and an adaptive harmonic spatial filter, respectively. The estimated reference signal is then used as input to the established and thoroughly evaluated intrusive measure STOI, which requires a clean reference signal. Both of the proposed non-intrusive measures have a high correlation with the original intrusively computed STOI scores.

References

[1] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[2] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acustica United with Acta Acustica*, vol. 101, pp. 1016–1025, 2015.

[3] 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 Process. Mag.*, vol. 32, no. 2, pp. 114–124, 2015.

[4] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[5] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[7] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[8] 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.

[9] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-hmm-based non-intrusive speech intelligibility prediction," in *ICASSP*, March 2016, pp. 624–628.

[10] 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, 2016.

[11] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Non-intrusive intelligibility prediction using a codebook-based approach," in *EUSIPCO*, August 2017, pp. 226–230.

[12] ——, "Non-intrusive codebook-based intelligibility prediction," *Speech Communication*, vol. 101, pp. 85 – 93, 2018.

[13] A. Heidemann Andersen, J. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *ICASSP*, March 2017, pp. 5085–5089.

[14] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 26, no. 10, pp. 1925–1939, 2018.

[15] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *EUSIPCO*, August 2016, pp. 1358–1362.

[16] C. Sørensen, A. Xenaki, J. Boldt, and M. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *ICASSP*, March 2017, pp. 386–390.

[17] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[18] J. R. Jensen, S. Karimian-Azari, M. G. Christensen, and J. Benesty, "Harmonic beamformers for speech enhancement and dereverberation in the time domain," *Speech Communication*, vol. 116, pp. 1–11, 2020.

[19] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 21, no. 10, pp. 2042–2056, 2013.

[20] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 409–412.

[21] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *EUROSPEECH*, vol. 1, 18-21 September 1995, pp. 867–870.

[22] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, 2010, pp. 1–6.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-560-4

AALBORG UNIVERSITY PRESS