Joint Far- and Near-end Speech and Listening Enhancement with a Minimum Processing Perspective

Fuglsig, Andreas Jonas

DOI (link to publication from Publisher):
10.54337/aau764598218

Publication date:
2024

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

JOINT FAR- AND NEAR-END SPEECH AND LISTENING ENHANCEMENT WITH A MINIMUM PROCESSING PERSPECTIVE

BY
ANDREAS J. FUGLSIG

PhD Thesis 2024

AALBORG UNIVERSITY
DENMARK

Joint Far- and Near-end Speech and Listening Enhancement with a Minimum Processing Perspective

PhD Dissertation

Andreas J. Fuglsig

PhD Thesis 2024

Submitted: August 2024

Main Supervisor: Professor Dr. Zheng-Hua Tan
Aalborg University, Aalborg, Denmark

Co-supervisors: Jens Christian Lindof
RTX A/S, Nørresundby Denmark

Professor Dr. Jan Østergaard
Aalborg University, Aalborg, Denmark

Lars Søndergaard Bertelsen
RTX A/S, Nørresundby, Denmark

Assessment: Associate Professor Jan Dimon Bendtsen (chairman)
Aalborg University, Denmark

Associate Professor Richard Christian Hendriks
Delft University of Technology, The Netherlands

Assistant Professor Aki Härmä
Maastricht University, The Netherlands

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

Department: Department of Electronic Systems

ISSN: 2446-1628
ISBN: 978-87-85239-55-6

Published by:
Aalborg University Open Publishing
Kroghstræde 1-3
DK – 9220 Aalborg Øst
aauopen@aau.dk

About the Author

Andreas J. Fuglsig

Andreas J. Fuglsig received the B.Sc. and M.Sc. degrees in Mathematical Engineering from Aalborg University in 2017 and 2019, respectively. He joined RTX A/S as an R&D engineer from 2019 to 2024, where since 2020, he has been working towards the PhD degree with the Centre on Acoustic Signal Processing Research (CASPR) at Aalborg University and RTX A/S. Since April 2024 he has been with the Audio Analysis Lab at the Department of Electronic Systems, Aalborg University, Denmark. His research interests include acoustic signal processing, speech enhancement and information theory.

Abstract

Speech communication across different environments presents challenges due to background noise affecting both the talker and listener. These disturbances can be highly annoying, leading to a decline in the perceived speech quality. Furthermore, they can also make it difficult for the listener to understand what was said, causing a drop in speech intelligibility. To address disturbances from the talker's environment (the so-called far-end), far-end speech enhancement algorithms are commonly employed to improve intelligibility and quality by reducing the noise in the recorded signals. Similarly, to overcome disturbances from the listener's environment (the so-called near-end), near-end listening enhancement algorithms are used to enhance intelligibility and quality by preprocessing signals before playback in the noisy near-end environment.

Traditionally, far- and near-end speech and listening enhancement systems have been developed independently, often overlooking the fact that noise can exist simultaneously at both ends. Treating these systems separately can result in reduced performance due to, e.g., conflicting processing goals, or excessive or insufficient processing because of an unawareness of the remaining noise and processing artifacts from the other end. Alternatively, by using joint far- and near-end speech and listening enhancement, which considers both environments and all processing steps simultaneously, it is possible to improve performance compared to the classic blind concatenation of far- and near-end systems.

While blind and joint far- and near-end speech and listening enhancement algorithms perform well in noisy conditions, they sometimes excessively prioritize noise suppression or the maximization of intelligibility at the near-end. This can lead to an exaggerated processing of the speech signals which can cause unwanted speech distortions and reduced quality, especially in quieter conditions where intelligibility is already high. However, using a minimum processing approach can help strike a balance between noise reduction, intelligibility improvement and high quality.

In this thesis, we explore joint far- and near-end speech and listening enhancement, along with minimum processing. Our approach aims to enhance intelligibility and quality beyond blind processing while minimizing speech

distortions compared to maximum processing. Particularly, we propose a joint far- and near-end speech intelligibility enhancement algorithm based on maximization of a speech intelligibility predictor. Additionally, we propose and study the use of a minimum processing formulation of near-end listening enhancement. Finally, we propose a combined joint far- and near-end minimum processing framework and comprehensively study and explore the effects of minimum versus maximum processing, joint versus blind processing, and their cross-combination.

Resumé

Talekommunikation på tværs af forskellige miljøer er udfordrende på grund af baggrundsstøj, der påvirker både taleren og lytteren. Disse forstyrrelser kan opleves meget irriterende og føre til et fald i den opfattede talekvalitet. Desuden kan de også gøre det vanskeligt for lytteren at forstå, hvad der bliver sagt, og dermed forringe taleforståeligheden. For at håndtere forstyrrelser fra talerens omgivelser (det såkaldte far-end-miljø) anvendes der ofte far-end-taleforbedringsalgoritmer til at forbedre forståeligheden og kvaliteten ved at reducere støjen i de optagede signaler. For at overvinde forstyrrelser fra lytterens miljø (det såkaldte near-end-miljø) bruges på samme måde near-end-lytteforbedringsalgoritmer til at forbedre forståelighed og kvalitet ved at præprocessere signaler før afspilning i det støjfyldte near-end-miljø.

Traditionelt er far- og near-end tale- og lytteforbedringssystemer blevet udviklet uafhængigt af hinanden, og man har ofte overset, at der kan være støj i begge miljøer på samme tid. Det at betragte disse systemer separat kan resultere i reduceret ydeevne på grund af f.eks. modstridende behandlingsmål eller overdreven eller utilstrækkelig behandling på grund af manglende bevidsthed om den resterende støj og behandlingsartefakter fra den anden ende. Alternativt er det muligt at forbedre ydeevnen sammenlignet med den klassiske blinde sammenkædning af far- og near-end-systemer ved at bruge kombineret far- og near-end tale- og lytteforbedring, hvor der tages højde for begge miljøer og alle behandlingstrin samtidigt.

Mens blinde og kombinerede far- og near-end tale- og lytteforbedringsalgoritmer fungerer godt under støjende forhold, prioriterer de nogle gange støjreduktion eller maksimering af forståeligheden i near-enden uforholdsmæssigt højt. Dette kan føre til en overdreven behandling af talesignalerne, som kan forårsage uønskede taleforvrængninger og reduceret kvalitet, især under mere stille forhold, hvor forståeligheden allerede er høj. Men ved at bruge en minimumsbehandlingstilgang kan man finde en balance mellem støjreduktion, forbedring af forståeligheden og høj kvalitet.

I denne afhandling undersøger vi kombineret far- og near-end tale- og lytteforbedring sammen med minimumsbehandling. Vores tilgang har til formål at forbedre forståeligheden og kvaliteten i forhold til den blinde ly-

dprocessering og samtidig minimere taleforvrængninger sammenlignet med maksimal lydprocessering. Vi foreslår især en kombineret far- og near-end taleforbedringsalgoritme baseret på maksimering af en prædiktor for taleforståelighed. Derudover foreslår og undersøger vi brugen af en minimumsbehandlingsformulering af near-end lytteforbedring. Endelig foreslår vi et samlet framework for kombineret far- og near-end-minimumsbehandling og undersøger grundigt effekterne af minimums- versus maksimumsbehandling og kombineret versus blind behandling samt deres krydskombination.

Contents

List of Abbreviations

List of Publications

The main body of this thesis consist of the following four papers:

[A] **Andreas Jonas Fuglsig**, Jan Østergaard, Jesper Jensen, Lars Søndergaard Bertelsen, Peter Mariager and Zheng-Hua Tan, "Joint Far- and Near-End Speech Intelligibility Enhancement based on the Approximated Speech Intelligibility Index," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7752–7756, 2022.

[B] **Andreas Jonas Fuglsig**, Jesper Jensen, Zheng-Hua Tan, Lars Søndergaard Bertelsen, Jens Christian Lindof and Jan Østergaard, "Minimum Processing Near-end Listening Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 2233–2245, 2023.

[C] **Andreas Jonas Fuglsig**, Jesper Jensen, Zheng-Hua Tan, Lars Søndergaard Bertelsen, Jens Christian Lindof and Jan Østergaard, "Joint Minimum Processing Beamforming and Near-End Listening Enhancement," *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 485–489, 2024.

[D] **Andreas Jonas Fuglsig**, Zheng-Hua Tan, Lars Søndergaard Bertelsen, Jesper Jensen, Jens Christian Lindof and Jan Østergaard, "Joint Far- and Near-end Speech and Listening Enhancement with Minimum Processing," *IEEE ACCESS*, p. 22, 2024.

In addition to the main papers, two patent applications have been filed in relation to the PhD project:

[E] **Andreas Jonas Fuglsig**, "JOINT FAR-END AND NEAR-END SPEECH INTELLIGIBILITY ENHANCEMENT," pub. no: WO2023057410A1, appl. date: 14. October 2022, pub. date: 13. April 2023.

[F] **Andreas Jonas Fuglsig**, " NEAR-END SPEECH INTELLIGIBILITY ENHANCEMENT WITH MINIMAL ARTIFACTS,", pub. no. EP4362015, appl. date: 28. October 2023, pub. date: 1. May 2024.

The patents are not included in the thesis.

Preface

This thesis documents the scientific work conducted as part of the PhD project "Joint Far- and Near-end Speech and Listening Enhancement with a Minimum Processing Perspective". The project was a collaboration between the Centre for Acoustic Signal Processing Research (CASPR), Aalborg University, Aalborg, Denmark and RTX A/S, Nørresundby, Denmark. The project was partially funded by Innovation Fund Denmark under case no. 9065-00204B. The thesis is submitted to the Technical Doctoral School of IT and Design at Aalborg University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

The thesis consists of two parts: Part I provides an extended introduction to fundamental concepts and essential fields relevant for understanding the research area, the objectives and the contributions of the project. Part II is the main body of the thesis and consists of a collection of peer-reviewed and published scientific papers. The papers present in details, the contributions, methods and discussion of the results of the PhD project.

Acknowledgment

First and foremost, I would like to express my sincere gratitude to my supervisors Prof. Jan Østergaard, Prof. Zheng-Hua Tan, Lars Søndergaard Bertelsen, and Jens Christian Lindof, for their unwavering support and guidance throughout this project. I deeply appreciate our numerous intriguing discussions, where their immense expertise and advice have significantly contributed to my professional and academic development. Without their support, neither this research nor I would be where we are today. A special thank you to Prof. Jesper Jensen for contributing his advice and massive expertise to the project. The project would not have been the same without his invaluable inputs. I am also grateful to my former manager and supervisor at RTX, Peter Mariager, for believing in me and initiating this project, and for his mentorship during our time at RTX. Additionally, I extend a great thanks to my former colleagues at RTX, especially Camilla S. Munk and the entire DSP and Audio team, for being wonderful office mates and for their inspiration, discussions, encouragement, and coffee breaks while working on this thesis. A great thank you to my colleagues at the section for AI and Sound (AIS) at Aalborg University, including professors, researchers, and fellow PhD students, especially my office mates Payam and Mohammad, for the enlightening discussions, support, many coffees, and great times. I also want to express my deep gratitude to my friends and family for their everlasting support and encouragement throughout my PhD journey. Finally, my most heartfelt thanks and gratitude go to Matilde for her unending support, comfort, and joy. Thank you for enduring with me, taking care of so many things that allowed me to focus on my PhD, and for being the best A-team mate I could ever wish for.

Andreas J. Fuglsig
Aalborg University, August 29, 2024

Part I

Introduction

Introduction

Speech and speech communication is central to human interaction, as it enables us to convey complex ideas, emotions, and intentions, and allows us to express our thoughts, share knowledge, and build connections. Consequently, we have used technology to enable speech communication over distances, and in diverse and challenging conditions, e.g., in hearing aids, public address systems, or two-way communication scenarios such as call centers, online meetings and (critical) intercom systems for, e.g., firefighters. In this thesis, we address a problem inherent in many of these communication scenarios illustrated in Fig. 1: the possible degradations of the *target speech* from a *target talker* captured in one noisy environment (called the *far-end*), that must be reproduced to a listener in a separate noisy environment (called the *near-end*). Here, the disturbances in the far- and near-end environments can lead to difficulties understanding the speech of interest and an overall bad experience for the listener, i.e, the disturbances can lead to deteriorations in both Speech Intelligibility (SI) and the perceived Speech Quality (SQ) for the listener.

A classic example of speech communication in a noisy environment, is the so-called *cocktail party problem* [1, 2], where the target talker and listener are having a conversation in a noisy environment with multiple disturbances, such as background music and conversations, that all limit both SQ and SI.

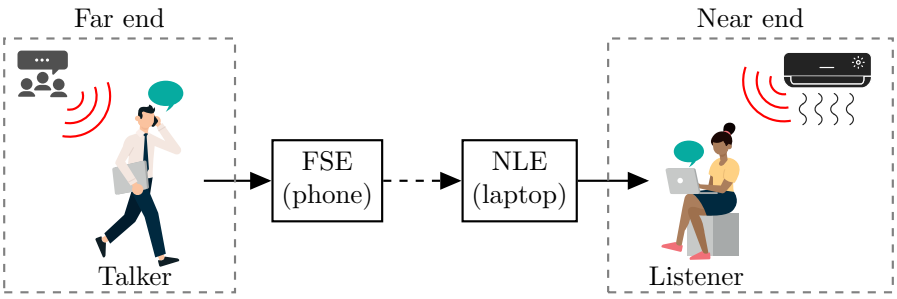


Fig. 1: Basic communication system with Far-end Speech Enhancement (FSE) and Near-end Listening Enhancement(NLE). Some images designed by pikisuperstar / Freepik.

However, here, the abilities of the human auditory systems enable the listener, with some effort, to pick up the correct speech stream from the otherwise noisy mixture of speech and background noise. Similarly, with the help of the human vocal system, the talker naturally alters their speech pattern such that it becomes more intelligible for the listener in the noisy background. This phenomenon is known as the *Lombard effect* [3, 4]. In addition to the Lombard effect, the natural redundancy of speech and the fine tuning of humans to the perception of speech provides humans with the ability to understand speech in a multitude of adverse conditions [2, 5].

Following the analogy of the cocktail party problem, what we consider in this thesis is the design of systems that automatically perform both the aspects of the auditory and vocal system in separate enhancements. That is, designing Far-end Speech Enhancement (FSE) systems that can pick up the correct stream from the noisy background (i.e., acting as the ears of the listener), and designing Near-end Listening Enhancement (NLE) systems that produce altered speech with increased SI in a noisy setting (i.e., acting as the talker).

Historically, due to the distinct technical challenges and optimization criteria involved in processing noisy microphone signals versus speech signals for playback, FSE and NLE system have been designed disjointly from each other, which leads to specialized and separate development efforts in each area. However, the disjoint approach ignores the bigger picture where noise is present at both ends simultaneously, and that the systems must be connected to each other in end-to-end communication scenarios. This can lead to performance degradations due to conflicting processing and optimization criteria [6–10]. In addition, there has been a focus on maximizing processing at either end by either only concentrating on removing all disturbances from the noisy recording at the far-end, or dedicating efforts to always maximizing the output SI at the near-end. However, this focus on maximizing noise reduction and output SI can lead to an overemphasis on these aspects in quieter conditions, potentially introducing artifacts and speech distortions compromising the naturalness and fidelity of the speech [11–16]. Hence, there is a trade-off in balancing effective noise reduction and SI improvement without sacrificing the inherent SQ.

Hypotheses

In this thesis, we explore two primary hypotheses throughout the four papers presented in Part II. These hypotheses are:

(H.1) Joint enhancement provides better objective and subjective SI and SQ than blindly concatenated FSE and NLE.

(H.2) Minimum processing provides better SQ than maximum processing in quieter conditions while preserving high SI.

That is, we envision leveraging knowledge from both the far-end and near-end simultaneously to enhance the performance of end-to-end speech and listening enhancement. Specifically, we consider the study, design, and evaluation of joint far- and near-end enhancement algorithms. Additionally, we focus on minimizing processing to better balance achieving high SI and SQ with reduced speech distortions. Hence, by integrating knowledge of the processing and environments of both ends and optimizing the overall processing jointly, our approach aims to improve and balance both high SQ and SI.

The rest of the Introduction provides a background for the papers presented in Part II, and serves to place the presented research in a broader perspective. In Section 1 we review basic concepts of speech science and communication. In Section 2, we introduce some basics of optimization problems. Then in Section 3, we introduce in more details FSE and NLE including some background on performance evaluations of enhancement algorithms and review of existing literature. Section 4 presents and discusses joint far- and near-end enhancement and reviews existing works. Similarly, Section 5 introduces the concept of minimum processing. In Section 6, we present some thoughts and observations regarding performance comparisons and listening tests in noisy near-end conditions. Finally, Sections 7 and 8 summarize the scientific contributions of our work and identify possible future research directions.

1 Speech Communication Model

In this section we consider the fundamentals of speech communication, consisting of three aspects; *speech production*, *speech transmission* and *speech perception* [17], as shown in the simple speech communication model in Fig. 2. Here a talker generates speech using their vocal system, the speech is then transmitted through a communication channel to a listener, who hears the output of the channel using their ears and attempts to interpret the speech. In this model, the communication channel may be purely acoustic or also consist of electronic processing parts, e.g., a headset. However, if the talker and listener are in the same environment or use, e.g., a video conferencing system, visual components of speech (body language or lib movements) may also be available to the listener. In this work we consider only the acoustic part of speech communication and refer to [18, 19] for an overview of audio-visual speech enhancement. Additionally, we note that the model assumes the communication is only one-way and no feedback is provided from the environment or the listener to the talker. However, this is sufficient for the scope of this thesis, as we are mainly concerned with speech enhancement scenarios where this limitation is fulfilled.

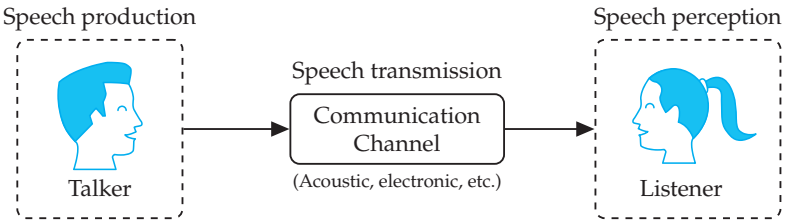


Fig. 2: A general block diagram showing the information transfer from a talker via a communication channel to a listener.

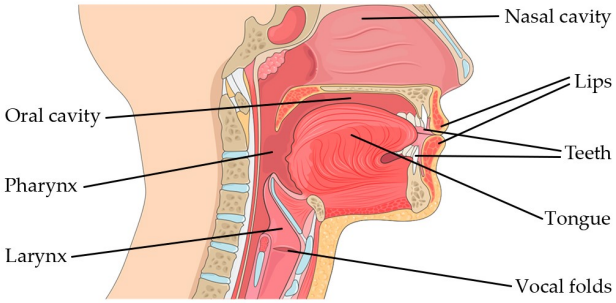


Fig. 3: Anatomy of the human vocal tract. Unlabelled art work provided by Servier Medical Art (`https://smart.servier.com/`), licensed under a Creative Commons Attribution 4.0 Unported License.

In the following, we introduce in more detail the three parts of the model in separation. We first consider the speech production and perception stages. Then, in Section 1.2, we expand and discuss the communication channel, as this is where the speech and listening enhancement of interest to this thesis occurs.

1.1 Speech Production and Perception

1.1.1 Speech Production

The very first stage of speech production happens when the brain converts the thoughts of the talker into a linguistic structure of sentences, cf. the areas of linguistics [20] and neurolinguistics [21]. However, in this thesis, we focus on the production of speech sounds that occurs after the brain sends electrical signals to the vocal muscles.

Fig. 3 shows an overview of the anatomy of the human vocal tract involved in speech production. Speech is produced when air is pushed from the lungs passed the *vocal folds* located in the *larynx*. To produce *voiced* sounds, such as *vowels*, the vocal folds close off to restrict the airflow from the lungs, causing a

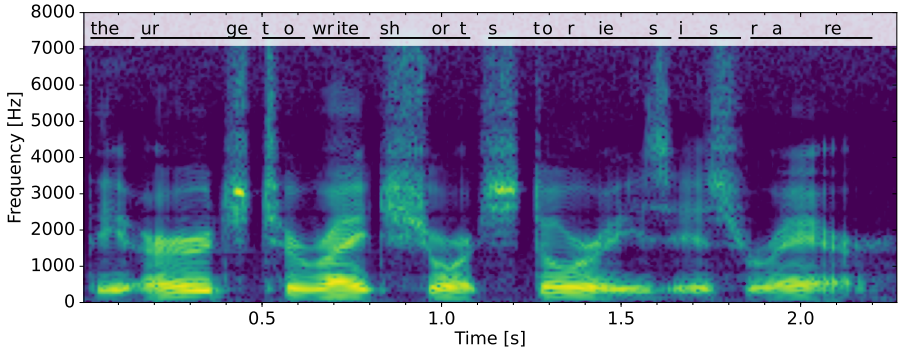


Fig. 4: Spectrogram of a male saying "*the urge to write short stories is rare*".

periodical release of bursts of air with a frequency of about 60 Hz to 400 Hz [22, 23]. This frequency is known as the *fundamental frequency*, and is the most contributing factor to the perceived *pitch* in speech [22]. As the sound moves from the vocal folds, muscle activity in the laryngeal cavity, the *pharynx*, the oral cavity, and the nasal cavity, causes a time-varying filtering of the excitation signal and thereby produce distinctly different sounds [23–25].

In addition to vowel sounds, speech also includes *consonant* sounds, which are produced by various constrictions in the vocal tract. For example, *fricatives* (e.g. /s/, /f/) are produced by very narrow constrictions leading to a noise-like burst of sound [24], and *plosives* (e.g. /g/, /t/ and /k/) are produced when the airflow is blocked by a complete constriction or released by the removal of a constriction [24]. Consonants may be either voiced or *unvoiced*, depending on if the vocal folds vibrate or not at the time of constriction [25].

To illustrate properties of a speech signal it is beneficial to look at a time-frequency representation of the signal called a *spectrogram*, as seen in Fig. 4. We see that the vowels in the spectrum consist of a series of harmonics and broad horizontal peaks. These peaks are called *formants* with corresponding *formant frequencies* [23]. Vowels are characterized by their different formant frequencies, corresponding to resonances in the vocal tract [25]. In free flowing-speech the articulators are continuously moving between the positions required to produce the latest and the next sound [25]. Therefore, as is apparent from Fig. 4, there is not a stable spectrum for each vowel and consonant sound. It can be seen, that consonants have a relatively lower energy compared to vowels [23]. Furthermore, the voiced segments (e.g. at 0.9 s and 1.3 s) have more energy at the lower frequencies, compared to the fricatives (/sh/ at 0.8 s and /s/ at 1.1 s) with less powerful, noise-like characteristics and more energy at the higher frequencies. Finally, plosives (/g/, /t/) are associated with a brief silence followed by a noise-like burst of air (e.g. at 0.4 s and 0.5 s).

For more details regarding speech and spoken language production we

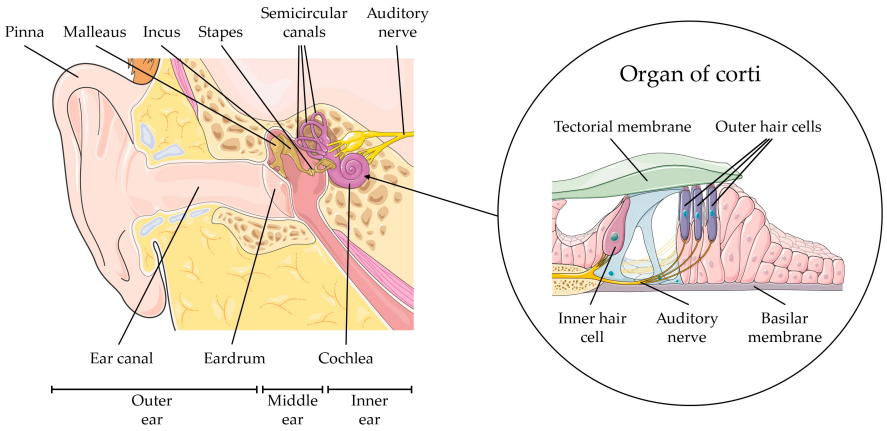


Fig. 5: Anatomy of the auditory system. Figure adapted from [19]. Unlabelled art work provided by Servier Medical Art (`https://smart.servier.com/`), licensed under a Creative Commons Attribution 4.0 Unported License.

refer to [22–25].

1.1.2 Speech Perception

We consider the final step in the speech communication model (Fig. 2) where sound signals enter the ear. Fig. 5 shows an overview of the anatomy of the human ear. Incoming sound waves initially hit the *pinna* (the external visible part of the ear), which causes multiple reflections to enter the ear canal. The reflections vary according to the direction of the incoming sound, hence these reflections help in determining the position of the speaker [25, 26]. The sound waves then propagate along the ear canal ending at the *eardrum*, which separates the outer ear from the middle ear. The incoming acoustic waves causes the eardrum to vibrate, this converts the acoustic waves to mechanical waves, which are transmitted to the inner ear by the *ossicles*; the three tiny bones in the middle ear called the *malleus (hammer), incus (anvil)* and *stapes (stirrup)*. The inner ear are made up of the spiral-shaped *cochlea*, and a set of semicircular canals that are related to the sense of balance [27]. The cochlea is the most important part of the auditory system, since this is where the mechanical vibrations coming from the stapes are converted to electrical neural activity exiting through the auditory nerve [25]. The cochlea is a coiled-up tube, which is separated along its length into three fluid-filled champers by *Reissner's membrane* and the *basilar membrane* [25]. When the mechanical vibrations of the stapes enters the cochlea the vibrations propagate along the

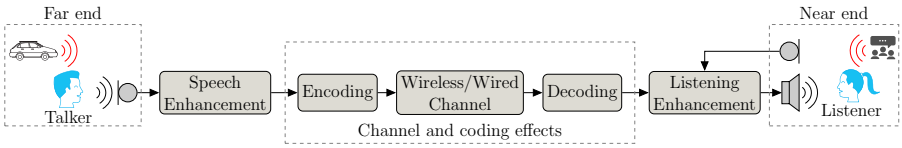


Fig. 6: Model of speech communication with expanded communication channel.

basilar membrane from the base to the apex. The basilar membrane is tuned to resonate at higher frequencies towards the base and lower frequencies towards the apex [25]. Since the properties of the basilar membrane vary continuously from base to apex, the basilar membrane performs a mechanical frequency analysis of the incoming sound and thus acts as a bank of overlapping *bandpass filters* [25, 28]. On top of the basilar membrane lies the *organ of Corti*. The organ of Corti consists of multiple rows of outer hair cells and a row of inner hair cells, which are all covered by several *stereocilia* ("hairs"). Above the hair cells lies the *tectorial membrane*. The tallest tips of the outer hair cells make contact with the tectorial membrane. As the basilar membrane vibrates the tectorial membranes causes the stereocilia to displace. In the inner hair cells, this displacement triggers a chemical reaction and a signal is sent along the auditory nerve [28]. The outer hair cells are not directly involved in the signals sent along the auditory nerve. Instead they are believed to actively influence the mechanical properties of the basilar membrane by increasing the sensitivity and frequency resolution [25, 28]. The signals sent along the auditory nerve are then processed through a complex network of neuronal structures ending in the primary auditory cortex [25]. For more introduction to hearing and the ear we refer to [25, 28]

The above described elements allow humans with healthy auditory system to perceive acoustic stimuli within a frequency range of 20 Hz to 20 kHz and a large dynamic range of around 120 dB [25]. However, damage or problems with parts of the auditory system can impair sound perception causing a *hearing loss*. Since we are mainly concerned with the scenario, where the talker and listener are in separate environments and have healthy hearing, the various aspects and causes of hearing loss are beyond the scope of this thesis and we refer to [25]. Although, we note that many of the existing speech enhancement techniques considered in this thesis are applicable in hearing aids for the treatment of hearing loss [29].

1.2 Speech Transmission

We now turn our attention to the communication channel in speech communication model of Fig. 2. To give a better understanding of the many components in the speech communication channel and their effects, we expand and update

the model as shown in Fig. 6. Here the previously described speech production and speech perception steps are included as inherent processes in the far-end talker and near-end listener. We note that not all blocks in this model are applicable to all communication scenarios.

Before the speech, produced by the talker, reaches the listener it must travel through an acoustic or partially acoustic communication channel to the ears of the listener. As the signal travels through the communication channel it may be effected by several factors. In the simplest case, where the speaker and listener are located close to each other such that the signal travels across a purely acoustic channel the speech signal is only effected by environmental noise and possibly reverberation. In many cases, it is desired to improve the experience of the listener by introducing steps in the chain that modify the acoustic signal. For example, in the slightly more advanced scenario where the listener may be wearing a hearing assistive device. Here the the speech signal travels through both a purely acoustic path, and also a more advanced path including the microphone(s), loudspeaker and signal processing of the hearing assistive device along with analog-to-digital- and digital-to-analog-conversion. Finally, in the scenario of interest to this thesis, the speaker and listener are physically separated. Here the situation is even more complex as communication is facilitated through, e.g., mobile phones or headsets, and the speech signal enters a complex processing chain, including analog-to-digital conversion, multiple cases of digital speech enhancement, wired/wireless transmission(s) with encoding and decoding, and digital-to-analog conversion. All these steps may introduce a number of distortions in the resulting signal [23, 30–34], in addition to the environmental noise and reverberation in the acoustic environments at both the far- and near-end.

The model in Fig. 6 does not cover all communication scenarios. In particular the order and number of blocks may change depending on a particular scenario, and also affect the signal presented to the listener's ears [8, 35]. That is, transmissions may occur across several channels and include multiple signal enhancement steps at various points in the chain.

1.2.1 Speech Processing as Mappings

It is useful for us to consider each processing step in the speech communication model of Fig. 6 as a mapping, $f : \mathcal{X} \rightarrow \mathcal{Y}$, between an input vector, $\boldsymbol{x} \in \mathcal{X}$, and an output vector, $\boldsymbol{y} \in \mathcal{Y}$, as illustrated in Fig. 7. Often the mapping is parameterized by a set of parameters, $\boldsymbol{w}$. For example, $\boldsymbol{w}$ could be a simple gain, such that $\boldsymbol{y} = w\boldsymbol{x}$, or $\boldsymbol{w}$ could be the parameters of a Deep Neural Network (DNN).

The goal of speech and listening enhancement is then to determine the optimal mapping, f^*, in some sense that improves the experience of the listener. This thesis centers on algorithms designed for implementation as such

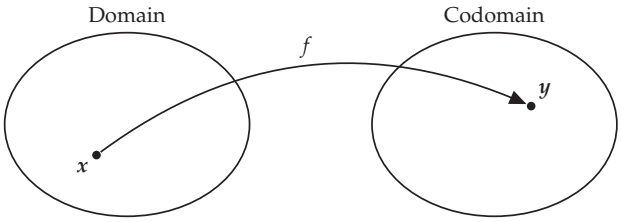


Fig. 7: We consider each block in speech communication model as a mapping, f, between an input vector $\boldsymbol{x}$ to an output vector $\boldsymbol{y}$.

mappings in communication channel devices, aiming to diminish the effects of background noises from both the far- and near-end to enhance SQ and SI. Hence, we limit ourselves from considerations of the channel and coding effects. However, we do comment on the necessity to consider the channel and coding when implementing algorithms and in future work cf. Sections 4 and 8. In Section 3, we return to how different speech and listening enhancement mappings have been derived in the literature.

2 Optimization Problems

As mentioned in Section 1.2, the derivation of speech and listening enhancement algorithms often requires some form of optimization to determine the optimal mapping between an input and output vector. Therefore, in this Section, we introduce some basic concepts of *mathematical optimization problems* and in particular *convex optimization problems*, which have been used extensively in our work in Part II.

Letting $\mathcal{W}$ be a subset of a vector space over the real numbers, a general constrained optimization problem can be formulated as [36]

$$\begin{aligned} \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \quad & f_0(\boldsymbol{w}) \\ \text{subject to} \quad & f_i(\boldsymbol{w}) \leq b_i, \quad i = 1, \ldots, m, \end{aligned} \tag{1}$$

where the vector, $\boldsymbol{w} \in \mathcal{W}$, is called the *optimization variable*, the function $f_0 : \mathcal{W} \to \mathbb{R}$ is the *objective-* or *cost*-function, and the functions $f_i : \mathcal{W} \to \mathbb{R}$ are the *constraint functions*, where the constants b_i are the limits or bounds for the constraints [36]. A vector $\tilde{\boldsymbol{w}}$ is said to be a *feasible solution* to the optimization problem if it satisfies all constraints of the problem, i.e., $f_1(\tilde{\boldsymbol{w}}) \leq b_1, \ldots, f_m(\tilde{\boldsymbol{w}}) \leq b_m$ [36]. The set of all feasible points is called the *feasible set*. We say a vector, $\boldsymbol{w}^*$ is an *optimal solution* to the optimization problem, if it achieves the smallest value of the objective function among all feasible solutions, i.e., $f_0(\boldsymbol{w}^*) \leq f_0(\tilde{\boldsymbol{w}})$ for all feasible $\tilde{\boldsymbol{w}}$ [36]. Finally, if there are no constraints, the problem is said to be *unconstrained*.

We limit the discussion of optimization to the perspective of minimization, since maximizing f_0 is equivalent to minimizing $-f_0$ subject to the same constraints [36].

2.1 Convex Optimization Problems

An optimization problem of the form

$$\begin{aligned} \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \quad & f_0(\boldsymbol{w}) \\ \text{subject to} \quad & f_i(\boldsymbol{w}) \leq 0, \quad i = 1, \ldots, m, \\ & h_i(\boldsymbol{w}) = 0, \quad i = 1, \ldots, p, \end{aligned} \tag{2}$$

is said to be a *convex optimization problem*, if the objective function, f_0 and all the inequality constraint functions, $f_1, \ldots, f_m$, are *convex*, and the equality constraint functions, $h_1, \ldots, h_p$, are *affine* [36].

Convex optimization problems possess a number of desirable properties [36], where the most important one is *global optimality*. That is, any local minimum of a convex optimization problem is also a global minimum. Furthermore, they can be efficiently solved using various algorithms, such as interior-point methods, gradient descent, and subgradient methods [36]. These algorithms exploit the convexity of the problem to converge to the optimal solution efficiently. Furthermore, convex optimization plays an important role in problems that are not convex [36]. For example, by formulating and solving an approximate convex problem, we can find a good initial guess for a local optimization method applied to a nonconvex problem. Additionally, convex optimization heuristics can be applied to nonconvex optimization, or as lower bounds for global optimization using, e.g., *Lagrangian* relaxation.

2.2 Lagrangian Duality and Karush-Kuhn-Tucker Conditions

In the following, we will cover part of *Lagrangian duality*, which has a central role in convex optimization, and provide the classical *Karush-Kuhn-Tucker conditions* for optimality.

For an optimization problem of the form (2), which may not necessarily be convex, the *Lagrangian* $L : \mathcal{W} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as [36]

$$L(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq f_0(\boldsymbol{w}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{p} \nu_i h_i(\boldsymbol{w}). \tag{3}$$

Here λ_i and ν_i are called the *Lagrange multipliers* associated with the ith inequality and equality constraint, respectively, and the vectors $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\nu} \in \mathbb{R}^p$ are the *dual variables* [36].

Letting $\mathcal{D} = \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \cap \bigcap_{i=1}^{p} \mathbf{dom}\, h_i$ be the domain of the optimization problem, which we assume to be non-empty, the *Lagrange dual function* $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as the minimum of the Lagrangian over $\boldsymbol{w}$ for $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\nu} \in \mathbb{R}^p$, i.e., [36]

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq \inf_{\boldsymbol{w} \in \mathcal{D}} L(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}). \tag{4}$$

The dual function is a concave function because it is a pointwise infimum of a family of affine functions [36]. It can be shown that the dual function provides a lower bound on the optimal value, p^*, of the problem (2). That is, for any $\boldsymbol{\lambda} \succeq 0$ and any $\boldsymbol{\nu}$, then [36]

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*. \tag{5}$$

This lower bound depends on the Lagrange multipliers, $(\boldsymbol{\lambda}, \boldsymbol{\nu})$. Therefore, to find the best lower bound we can solve the *Lagrange dual problem* [36]

$$\begin{aligned} \underset{\boldsymbol{\lambda}, \boldsymbol{\nu}}{\text{maximize}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} \quad & \lambda_i \geq 0, \quad i = 1, \ldots, m. \end{aligned} \tag{6}$$

In a similar manner the original optimization problem (2) is often called the *primal problem*. The pair of Lagrange multipliers, $(\boldsymbol{\lambda}, \boldsymbol{\nu})$, are said to be *dual feasible* if $\boldsymbol{\lambda} \succeq 0$ and $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) > -\infty$, i.e., they are feasible for the dual problem [36]. Similarly, $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ are called *dual optimal* or *optimal Lagrange multipliers* if they are optimal for (6). The dual problem is a convex optimization problem, because it is a maximization problem where the objective function, i.e., the dual function, is a concave function, and the constraints are convex.

We denote the optimal value of the dual problem by d^*. By definition, we have that $d^* \leq p^*$, and this property is called *weak duality*. The difference $p^* - d^*$ is known as the *optimal duality gap*, as it provides the difference between the optimal value of the primal problem and the greatest lower bound obtained from the Lagrange dual functions [36]. If the duality gap is zero, i.e., if $d^* = p^*$, then *strong duality* holds [36]. In this case, the greatest lower bound obtained by the Lagrange dual function is tight.

Various results provide regularity conditions (or constraint qualifications) for which strong duality holds. If the optimization problem (2) is convex, then a simple regularity conditions is *Slater's condition*, which states that there exist a point $\boldsymbol{w}$ such that [36]

$$f_i(\boldsymbol{w}) < 0, \quad i = 1, \ldots, m, \tag{7}$$

$$h_i(\boldsymbol{w}) = 0, \quad i = 1, \ldots, p. \tag{8}$$

This point is often said to be *strictly feasible*. If the primal optimization problem is convex, then Slater's theorem says, that strong duality holds if Slater's condition holds [36].

2.2.1 KKT Optimality Conditions

We now present the commonly used *Karush-Kuhn-Tucker* (KKT) conditions, for which a solution to an optimization problem is optimal, given that the objective and constraint functions, $f_0, \ldots, f_m, h_1, \ldots, h_p$ are differentiable [36].

The KKT conditions for the triplet of dual and primal variables $\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}$ for an optimization problem of the form (2) are [36]

$$f_i(\boldsymbol{w}) \leq 0, \quad i = 1, \ldots, m \tag{9a}$$

$$h_i(\boldsymbol{w}) = 0, \quad i = 1, \ldots, p \tag{9b}$$

$$\lambda_i \geq 0, \quad i = 1, \ldots, m \tag{9c}$$

$$\lambda_i f_i(\boldsymbol{w}) = 0, \quad i = 1, \ldots, m \tag{9d}$$

$$\nabla f_0(\boldsymbol{w}) + \sum_{i=1}^{m} \lambda_i \nabla f_i(\boldsymbol{w}) + \sum_{i=1}^{p} \nu_i \nabla h_i(\boldsymbol{w}) = 0. \tag{9e}$$

The first two conditions state that $\boldsymbol{w}$ must be primal feasible and satisfy all the constraints of (2). The third condition states that the dual variables must be dual feasible. The fourth condition is called *complimentary slackness*, and states that at point $\boldsymbol{w}$, either $f_i(\boldsymbol{w}) = 0$, $\lambda_i = 0$ or both are equal to zero. Finally, the fifth condition is the *stationarity* condition and states that for the dual variable pair $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ the point $\boldsymbol{w}$ minimizes the Lagrangian, $L(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu})$. We now state some important results regarding the KKT conditions.

Nonconvex Problems For *any* optimization problem of the form (2) with differentiable constraint and objective functions, any pair of primal and dual optimal points where strong duality holds must necessarily satisfy the KKT conditions (9) [36].

Convex Problems If the primal problem (2) is convex, the KKT conditions are also sufficient conditions for primal and dual optimality. That is, for any convex optimization problem of the form (2) with differentiable constraint and objective functions, any pair of points, $\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}$, that satisfy the KKT conditions (9) are primal and dual optimal with zero duality gap [36]. Particularly, if Slater's condition is satisfied for a convex optimization problem with differentiable objective and constraint functions, the KKT conditions are both necessary and sufficient conditions for optimality [36].

The KKT conditions play a crucial role in optimization, either as a backbone in algorithms solving convex and nonconvex optimization problems, or in the few cases where they can be solved analytically to find an optimal point [36]. In our work, we have analytically solved several optimization problems using the KKT conditions to find optimal speech and listening enhancement procedures, cf. Papers A, B, and D.

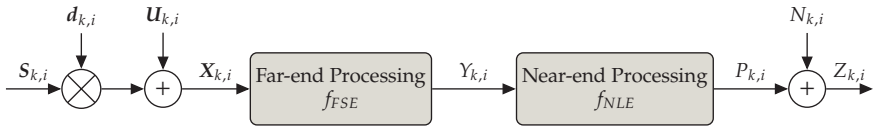


Fig. 8: Signal model with far-end speech enhancement and near-end listening enhancement.

For more details on the above topics and convex optimization in general including numerical methods we refer to the work of [36].

3 Speech and Listening Enhancement

One of the main topics of this thesis revolve around joint Far-end Speech Enhancement (FSE) and Near-end Listening Enhancement (NLE), cf. Papers A, C, and D. However, before we can discuss joint far- and near-end speech and listening enhancement, we first present some background on FSE and NLE in separation in this Section, which also sets the scene for our NLE work in Paper B. We start by first introducing a signal model, and introduce metrics for estimating SI and SQ performance. We then formulate the FSE and NLE problems along with an overview of existing methods for FSE and NLE.

3.1 Signal Model

We initially introduce the multi-microphone signal model for the signal processing chain of speech in noise, cf. Fig. 8. Here $\boldsymbol{X}_{k,i} \in \mathbb{C}^M$ denotes the noisy multi-microphone signal recorded by M microphones in time-frequency domain, where k denotes frequency index and i is time index, and is given as

$$\boldsymbol{X}_{k,i} = \boldsymbol{d}_{k,i} S_{k,i} + \boldsymbol{U}_{k,i}. \tag{10}$$

Here $S_{k,i} \in \mathbb{C}$ is the clean target speech signal measured at the source location, and $\boldsymbol{d}_{k,i} \in \mathbb{C}^M$ denotes the vector of Acoustic Transfer Functions (ATFs) from the source to the microphones. The vector $\boldsymbol{U}_{k,i} \in \mathbb{C}^M$ is the additive environmental noise recorded at the M microphones, which could be any undesired sound signal, such as ambient noise, reverberation and competing talkers, etc. The ATFs, $\boldsymbol{d}_{k,i}$, are considered as deterministic, whereas the speech and noise signals, $S_{k,i}$, $\boldsymbol{U}_{k,i}$, are considered as random vectors.

The first processing stage encountered in this model is the FSE, where the multi-microphone signal $\boldsymbol{X}_{k,i}$ is mapped into a new signal $Y_{k,i}$, i.e.,

$$Y_{k,i} = f_{FSE}(\boldsymbol{X}_{k,i}; \boldsymbol{w}_{k,i}, \mathcal{S}_S, \mathcal{S}_U), \tag{11}$$

where the signal $Y_{k,i}$ is an estimate of the target speech signal, $S_{k,i}$. The task of the FSE mapping, f_{FSE}, parameterized by $\boldsymbol{w}$, is to optimally estimate the target

speech signal, $S_{k,i}$, from the noisy input, $\boldsymbol{X}_{k,i}$. Furthermore, the mapping might take as input some relevant *side-information*, $\mathcal{S}$, of the clean speech and the far-end noise, e.g., second-order statistics. See Section 3.3 below for more details on FSE.

As mentioned in Section 1.2, we do not consider the channel and coding effects of the speech communication model explicitly in this thesis. Therefore, we assume the signal from the far-end is transmitted to the near-end without any degradations or delays.

The signal received from the far-end, $Y_{k,i}$, is then mapped into a new signal, $P_{k,i}$, by the NLE, before it is played out in the near-end environment, where it is contaminated by the near-end environmental noise, $N_{k,i}$, i.e.,

$$P_{k,i} = f_{NLE}(Y_{k,i}; \boldsymbol{\theta}_{k,i}, \mathcal{S}_Y, \mathcal{S}_N) \tag{12}$$

$$Z_{k,i} = P_{k,i} + N_{k,i}, \tag{13}$$

where $Z_{k,i}$ is the noisy near-end signal heard by the listener. Similar to the far-end signals, the near-end noise, $N_{k,i}$, is also considered to be a random process. The task of the NLE mapping, parameterized by $\boldsymbol{\theta}$, is to optimally map the signal $Y_{k,i}$ into $P_{k,i}$, such the final signal received by the listener, $Z_{k,i}$, has the best possible performance (SI and/or SQ). The NLE mapping may also take some relevant side-information of the received signal and the near-end noise as input. We give more details on NLE in Section 3.4 below. We note, that we limit ourselves from considering any spatial aspects at the near-end, as this is beyond the scope of this thesis.

3.1.1 Second-order Statistics

Many multi-channel FSE methods, especially those considered in our work, fall in the class of statistical model based algorithms minimizing a Mean-Squared-Error (MSE) criterion [37]. Therefore, particularly the Linear Minimum Mean-Squared-Errors (LMMSEs) estimators of $S_{k,i}$, that we consider in the following, are typically functions of the first- and second-order statistics of the speech and noise signals. The first-order statistics of the speech and noise signals are commonly assumed to be zero-mean [23]. The second-order statistics are the Power Spectral Density (PSD) of $S_{k,i}$ and $N_{k,i}$, and the (spatial) Cross-Power Spectral Density (CPSD) matrix of $\boldsymbol{X}_{k,i}$ and $\boldsymbol{U}_{k,i}$ [29, 37]. The PSD of a signal, $S_{k,i}$, is defined as

$$\sigma^2_{S_{k,i}} \triangleq \mathrm{E}\left[|S_{k,i}|^2\right], \tag{14}$$

and similarly for $N_{k,i}$. The CPSD matrix of $\boldsymbol{X}_{k,i}$ is defined as

$$C_{\boldsymbol{X}_{k,i}} \triangleq \mathrm{E}\left[\boldsymbol{X}_{k,i}\boldsymbol{X}^H_{k,i}\right], \tag{15}$$

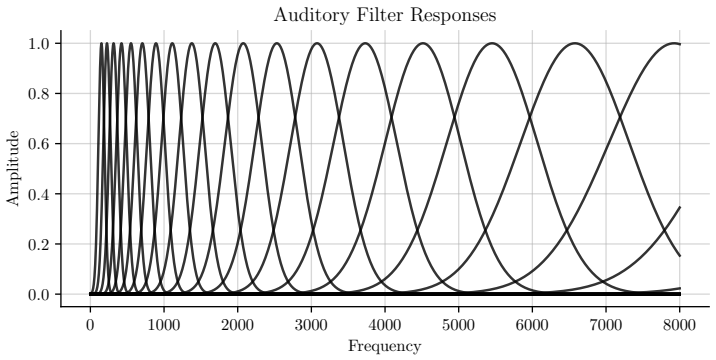


Fig. 9: Example of auditory subband filters.

and is often referred to as the (spatial) covariance matrix [37]. Using the common assumption, that speech and noise signals are uncorrelated, the covariance matrix of $\boldsymbol{X}_{k,i}$ is given as

$$C_{\boldsymbol{X}_{k,i}} = C_{\boldsymbol{S}_{k,i}} + C_{\boldsymbol{U}_{k,i}}, \tag{16}$$

where $C_{\boldsymbol{S}_{k,i}}$ is the target speech covariance matrix and $C_{\boldsymbol{U}_{k,i}} \triangleq \mathrm{E}\left[\boldsymbol{U}_{k,i}\boldsymbol{U}_{k,i}^H\right]$ is the far-end noise covariance matrix. Under the assumption of deterministic ATFs, $\boldsymbol{d}_{k,i}$, the target speech covariance matrix can be expressed as $C_{\boldsymbol{S}_{k,i}} = \sigma^2_{S_{k,i}}\boldsymbol{d}_{k,i}\boldsymbol{d}_{k,i}^H$.

In practice, the statistics of the speech and noise signal need to be estimated using signals recorded by microphones in the far- and near-end environment. Throughout our work, we assume that the relevant statistics are available and hence do not consider further the details of estimating noise and speech statistics and refer to, e.g., [23, 37, 38].

3.1.2 Subband Model

An important part of sound perception by the human ear is the mechanical filtering of the incoming sounds into a set of auditory subbands [25]. Therefore, many SQ and SI estimators mimic this aspect of human sound perception, by analyzing signals in perceptually relevant subbands, e.g., octave bands, fractional octave bands or critical bands [39]. Figure 9 shows an example of a set of auditory subband filters, where we see that the filters overlap such that multiple frequencies may contribute to the same and/or more subbands. Furthermore, the filters get wider at higher frequencies to mimic the behavior of the human ear [25]. In our work, we denote subbands with the index j and frequencies with the index k. For the j'th subband, we model the subband filters with the non-negative filter weights $\omega_{j,k} \in \mathbb{R}_+$. Thus, the clean speech

power spectrum level within one subband, j, and time-frame, i, is given as

$$\sigma^2_{S_{j,i}} \triangleq \sum_{k \in \mathbb{B}_j} \omega_{j,k} \sigma^2_{S_{k,i}}, \tag{17}$$

where $\mathbb{B}_j$ is the set of frequency bins that contribute to the j'th subband. Similar definitions hold for noise signals in each subband.

3.2 Speech Intelligibility and Quality Estimators

When deriving speech and listening enhancement algorithms it is important to evaluate their performance both during and after development. The most popular factors to evaluate in speech enhancement are SQ and SI. SQ is measures "how" speech is perceived, whereas SI measures "what" was said [23].

SQ and SI are perceptual attributes. Therefore, it would be ideal to conduct *subjective listening tests* with human listeners to properly evaluate performance [23, 40]. In the literature, numerous ways to conduct subjective listening tests have been proposed [23, 40]. In some of our works we conduct subjective listening tests to evaluate the performance of the proposed algorithms, cf. Papers B and D. Later on in Section 6, we discuss some thoughts behind conducting these tests. However, subjective listening tests are often complex and time consuming to prepare and conduct. Hence, it is impractical to conduct listening tests each time there is a desire to evaluate SI or SQ, for example when iteratively developing algorithms or devices for speech enhancement. Therefore, it is common to use *objective measures* to estimate SQ and SI. In addition to their use in evaluating algorithms, these performance estimators can also serve as optimization targets for deriving FSE and NLE algorithms. Below we present some estimators that have been used in our works. For a more comprehensive review and comparisons of SI and SQ estimators we refer to [23, 41–45].

3.2.1 Signal-to-Noise and Distortion Ratios

It is very common in speech and listening enhancement literature to encounter metrics, discussions or observations regarding Signal-to-Noise Ratio (SNR) and Signal-to-noise and Distortion Ratio (SDR). Therefore, it is important to clearly distinguish between SNR and SDR. Especially, as we may sometimes see SDR referred to as SNR, particularly in some FSE work [16, 23], and also in some performance metrics [23].

We define the subband SDR at the far- and near-end, respectively, as

$$\psi_{j,i}^{F} \triangleq \frac{\sigma_{S_{j,i}}^2}{\sum_{k\in\mathbb{B}_j} \omega_{j,k}\, \mathrm{E}\left[|S_{k,i} - Y_{k,i}|^2\right]} \tag{18}$$

$$\psi_{j,i}^{N} \triangleq \frac{\sigma_{S_{j,i}}^2}{\sum_{k\in\mathbb{B}_j} \omega_{j,k}\, \mathrm{E}\left[|S_{k,i} - Z_{k,i}|^2\right]}. \tag{19}$$

That is, SDR is the ratio between the clean speech power and the MSE between clean speech and the processed signal. Thus, it is the ratio between speech and the combined power of both noise and distortion. Hence, all differences between the signal received by the listener and the clean speech are undesirable if maximizing the SDR.

The subband SNR for the processed signals is defined as the ratio between processed speech, $\widetilde{S}$, and the total processed noise powers, $\widetilde{U}$ and $\widetilde{N}$, for the far- and near-end, respectively, as

$$\xi_{j,i}^{F} \triangleq \frac{\sigma_{\widetilde{S}_{j,i}}^2}{\sigma_{\widetilde{U}_{j,i}}^2} \tag{20}$$

$$\xi_{j,i}^{N} \triangleq \frac{\sigma_{\widetilde{S}_{j,i}}^2}{\sigma_{\widetilde{U}_{j,i}}^2 + \sigma_{\widetilde{N}_{j,i}}^2}. \tag{21}$$

The unprocessed SNR is defined in a similar manner using the unprocessed speech and noise powers. Hence, only the environmental noise is considered as the undesired part in SNR maximization, and any changes to the speech signal (e.g., from an NLE gain) also help to increase or reduce the subband SNR. We note, if no distortion is present in the system (e.g., if no processing has occurred), the SDR and SNR are equivalent. Furthermore, it can be difficult or impossible to compute the SNR in the presence of non-linear processing, as the signal and noise components not easily separable. Therefore, the SDR provides a good alternative in the presence of non-linear processing.

3.2.2 Speech Intelligibility Estimators

The Speech Intelligibility Index (SII) [39] and particularly the Approximated Speech Intelligibility Index (ASII) [46] are important for the FSE and NLE optimization in Part II of this thesis. Both SII and ASII are SNR based measures [42], that are build on the concept, that audibility of speech in a narrow frequency bands is the deciding factor for SI [47]. That is, the SI is computed as [39, 46],

$$(A)SII \triangleq \sum_{j=1}^{J} \gamma_j I(\xi_j), \tag{22}$$

where J is the total number of subbands, and γ_j are the so-called *band importance functions* which constitute a weighting on of the relative importance of the j'th subband. The *audibility function*, $I(\xi_j)$, of the subband SNR, ξ_j, measures the intermediate speech SI in the j'th subband and is defined as [39, 46],

$$I(\xi_j) = \begin{cases} \frac{\max(\min(10\log_{10}(\xi_j),15),-15)}{30} + \frac{1}{2} & \text{for SII [39]} \\ \frac{\xi_j}{\xi_j+1} & \text{for ASII [46].} \end{cases} \tag{23}$$

The ASII [46] introduces a linear approximation of the logarithm and clipping of the SII, which provide better SI prediction for lower and higher SNRs and is more mathematically simple [46]. The ASII and SII both require that the speech and noise components are available separately, in order to be able to compute the SNR. The SII and ASII are simple metrics with good but also limited performance [41, 43]. Particularly, they are difficult to use when signal and noise components are not easily separable as in the case of non-linear processing. Therefore, more advanced measures, that utilize other statistics than SNR, e.g., cross correlation and coherence, which can be computed for non-linearly processed signals have been developed with better SI prediction capabilities [41, 42, 42, 43].

The Extended Short-Time Objective Intelligibility (ESTOI) measure [48], which is an extension of Short-Time Objective Intelligibility (STOI) [49], is used extensively in Part II of the thesis for performance evaluation. STOI computes a correlation coefficient between the short-time overlapping temporal envelope segments of a clean and processed speech signal [49]. This is done by using a model of the human auditory system to obtain temporal envelopes from both clean and distorted speech across various frequency bands, and then calculating Pearson's correlation coefficient for each short-time frame and frequency band, averaging these values for the final SI score. To address STOI's limitations with highly modulated noise sources, ESTOI improves upon STOI by computing correlations between clean and distorted spectra instead of just temporal envelopes enabling the detection of 'glimpses' of clean speech [43, 48].

3.2.3 Speech Quality Estimators

The SDR and SNR as standalone measures are not well suited for SQ estimation [50]. However, the Segmental Signal-to-Noise Ratio (Seg-SNR) and especially its frequency weighted variations [23], are simple metrics of SQ with good performance [23]. The Seg-SNR is defined as [23],

$$\text{Seg-SNR} \triangleq \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\left\|\boldsymbol{s}_{Nm}^{Nm+N-1}\right\|_2^2}{\left\|\boldsymbol{s}_{Nm}^{Nm+N-1}-\widehat{\boldsymbol{z}}_{Nm}^{Nm+N-1}\right\|_2^2}, \tag{24}$$

where $\boldsymbol{s}_{Nm}^{Nm+N-1} = [s[Nm],\ldots,s[Nm+N-1]]^T$ is a frame of N time-domain samples, and M is the number of frames in the signal. When computing the

Seg-SNR each term in the sum is limited to the range $-10\,\text{dB}$ to $35\,\text{dB}$, since SNRs above $35\,\text{dB}$ do not show large perceptual differences, and silent frames may have very low SNRs not reflecting the proper perceptual contribution [50]. We notice and emphasize, that each term in the sum of the Seg-SNR is in fact *not an SNR but an SDR*. This illustrates, why some caution must be taken when referring to SNR and SDR.

The Perceptual Evaluation of Speech Quality (PESQ) measure [51], is the most widely used estimator of SQ and was originally designed to evaluate speech coding in telephone networks. PESQ approximates the results of a Mean Opinion Score (MOS) test, where human subjects rate speech signals on a five-point scale [40], by considering psychoacoustic principles such as the non-uniform frequency resolution of the human auditory system, non-linear loudness perception, and masking effects that prevent the perception of weak sounds [25, 28]. The PESQ algorithm involves time-aligning the reference and processed speech signals, mapping them to an auditory representation using power distributions over time-frequency and compressive loudness scaling, and then calculating differences between them [45]. These disturbances are scaled asymmetrically and symmetrically, and their combination provides the final PESQ SQ score [45]. PESQ has been shown to reliably assess the overall SQ of speech signals processed with common speech enhancement systems [23, 45]. However, because PESQ was not developed for speech enhancement, care must be taken with PESQ in the presence of environmental noise, since the noise easily dominates the PESQ score and hides the effect of distortions [16].

3.3 Far-end Speech Enhancement

In this Section, we introduce in more detail the FSE problem and some relevant multi-channel noise reduction methods. For more details on various FSE methods we refer to the comprehensive reviews and books of [18, 23, 29, 37, 52–56] as well as the results of [57–61].

3.3.1 Problem Formulation

The general idea behind the FSE task is to determine a mapping that provides the best possible estimate of the target speech signal, $S_{k,i}$, from the noisy microphone input signal, $\boldsymbol{X}_{k,i}$. That is, the task is to solve the following optimization problem,

$$\begin{aligned} \boldsymbol{w}^*_{k,i} = \underset{\boldsymbol{w}_{k,i}}{\arg\min} \quad & \mathcal{D}\left(f_{FSE}(\boldsymbol{X}_{k,i}; \boldsymbol{w}_{k,i}, \mathcal{S}_S, \mathcal{S}_U), S_{k,i}\right) \\ \text{subject to} \quad & \mathcal{I}_l\left(f_{FSE}(\boldsymbol{X}_{k,i}; \boldsymbol{w}_{k,i}, \mathcal{S}_S, \mathcal{S}_U), S_{k,i}\right) \geq 0, \quad l = 1, \ldots, m, \\ & h_l\left(\boldsymbol{w}_{k,i}\right) \leq 0, \qquad\qquad\qquad\qquad\quad l = 1, \ldots, p. \end{aligned} \tag{25}$$

Here $\mathcal{D}$ is a cost function indicating the performance of the FSE noise reduction, $\mathcal{I}_l$ are additional possible performance constraints on the FSE mapping, and h_l are possible constraints on the FSE mapping parameters. The cost and constraint functions could for example be SI or SQ estimators [23]. Depending on the choice of cost and constraint functions, performance might be evaluated in relation to the target speech signal [41].

Although we formulate the FSE task as an optimize problem, such an optimization problem might not be directly solved in order to determine an FSE mapping with more heuristic methods. However, for heuristic methods the idea is still the same, i.e., to determine a mapping that results in high SQ and/or SI possibly subject to some constraints.

3.3.2 Multi-channel Noise Reduction / Beamforming

In general, the mapping constituting the FSE can be either linear or non-linear. For example if the FSE method is a DNN, the mapping is highly non-linear. However, in our works in Part II, and in many existing works, the proposed FSE method is a linear process. Therefore, we now introduce the following typical linear noise reduction system, where the FSE mapping consists of a *beamformer*, $\boldsymbol{w}_{k,i} \in \mathbb{C}^M$, and a post-filter, $g_{k,i} \in \mathbb{R}$ [37]. That is,

$$Y_{k,i} = f_{FSE}(\boldsymbol{X}_{k,i}; \boldsymbol{w}_{k,i}, g_{k,i}) = g_{k,i} \boldsymbol{w}_{k,i}^H \boldsymbol{X}_{k,i}, \tag{26}$$

where H denotes Hermitian transpose. Multi-microphone methods are able to utilize the spatial diversity, i.e., the different acoustic properties at the different microphones, to suppress and enhance sounds in certain directions [37]. Therefore, multi-microphone speech enhancers are also often known as *multi-channel spatial filters* or beamformers.

Following the spatial filtering, a single-channel post-filter might be applied to remove residual noise. If only a single-microphone is available, the beamformer is not applied and the post-filter acts as the single-channel FSE. The post-filter / single-microphone methods lead to improvements in SQ [23, 62]. However, they rarely improve SI [23, 62]. Whereas the multi-microphone speech enhancement leads to improvements in both SQ and SI [58, 63].

We introduce two beamformers that appear in our work in Part II: The Minimum Variance Distortionless Response (MVDR) beamformer [29, 64], and the Multi-channel Wiener Filter (MWF) [65] in particular the Speech-Distortion-Weighted Multi-channel Wiener Filter (SDW-MWF) [66].

MVDR The MVDR beamformer is the solution to the optimization problem [37],

$$\boldsymbol{w}_{k,i}^{MVDR} = \arg\min_{\boldsymbol{w}} \ \boldsymbol{w}^H C_{\boldsymbol{U}_{k,i}} \boldsymbol{w}, \quad \text{s.t. } \boldsymbol{w}^H \boldsymbol{d}_{k,i} - 1 = 0, \tag{27}$$

where, as introduced above, $\boldsymbol{d}_{k,i}$ is the ATFs from the source to the microphones and $C_{\boldsymbol{U}_{k,i}}$ is the far-end noise covariance matrix. Here the constraint ensures that the beamformer does not distort the target speech coming from the direction of the source, and that the solution is non-trivial [38]. The MVDR beamformer is then given as [37]

$$\boldsymbol{w}_{k,i}^{MVDR} = \frac{C_{\boldsymbol{U}_{k,i}}^{-1}\boldsymbol{d}_{k,i}}{\boldsymbol{d}_{k,i}C_{\boldsymbol{U}_{k,i}}^{-1}\boldsymbol{d}_{k,i}}. \tag{28}$$

MWF The MWF minimizes the MSE between the beamformer output, $Y_{k,i}$, and the target speech $S_{k,i}$, and thus maximizes the SDR. Hence, the MWF is the LMMSE estimator of the clean speech spectrum and is the solution to the optimization problem [37],

$$\boldsymbol{w}_{k,i}^{MWF} = \arg\min_{\boldsymbol{w}} \mathrm{E}\left[\left|S_{k,i} - \boldsymbol{w}^H\boldsymbol{X}_{k,i}\right|^2\right], \tag{29}$$

which contrary to the MVDR and other common beamformers does not have a distortionless constraint [37]. For uncorrelated speech and noise components the MWF is then [29, 37],

$$\boldsymbol{w}_{k,i}^{MWF} = C_{\boldsymbol{X}_{k,i}}^{-1}\sigma_{S_{k,i}}^2\boldsymbol{d}_{k,i} \tag{30}$$

$$= \boldsymbol{w}_{k,i}^{MVDR} \cdot \frac{\sigma_{S_{k,i}}^2}{\sigma_{S_{k,i}}^2 + \sigma_{\widetilde{U}_{k,i}}^2}. \tag{31}$$

The second equation shows, the MWF can be decomposed into an MVDR beamformer and a Wiener post-filter, where $\sigma_{\widetilde{U}_{k,i}}^2$ is the residual noise power after the beamformer [37]. The gain following the MVDR beamformer is called a Wiener post-filter because it is equivalent to applying a single-channel Wiener filter [23] at the output of the MVDR beamformer.

SDW-MWF Writing out the cost function for the MWF, we see that it equally punishes both speech distortions and the residual noise after the beamformer, i.e.,

$$\mathrm{E}\left[\left|S_{k,i} - \boldsymbol{w}^H\boldsymbol{X}_{k,i}\right|^2\right] = \left|1 - \boldsymbol{w}^H\boldsymbol{d}_{k,i}\right|^2\sigma_{S_{k,i}}^2 + \boldsymbol{w}^H C_{\boldsymbol{U}_{k,i}}\boldsymbol{w}. \tag{32}$$

To better control the compromise or tradeoff between these two terms, it was proposed in [66] to generalize the cost function with a weight factor $\mu \in R_+$, and solve the optimization problem,

$$\boldsymbol{w}_{k,i}^{\mu MWF} = \arg\min_{\boldsymbol{w}} \left|1 - \boldsymbol{w}^H\boldsymbol{d}_{k,i}\right|^2\sigma_{S_{k,i}}^2 + \mu\boldsymbol{w}^H C_{\boldsymbol{U}_{k,i}}\boldsymbol{w}. \tag{33}$$

The parameter μ, is known as the *speech distortion weight*, and the solution to this optimization problem is the SDW-MWF given as [29],

$$\boldsymbol{w}_{k,i}^{\mu MWF} = \left(C_{S_{k,i}} + \mu C_{\boldsymbol{U}_{k,i}}\right)^{-1} \sigma_{S_{k,i}}^2 \boldsymbol{d}_{k,i} \tag{34}$$

$$= \boldsymbol{w}_{k,i}^{MVDR} \cdot \frac{\sigma_{S_{k,i}}^2}{\sigma_{S_{k,i}}^2 + \mu \sigma_{\widetilde{U}_{k,i}}^2}. \tag{35}$$

We see that for $\mu = 1$ the SDW-MWF reduces to the standard MWF (31). When $\mu < 1$, the post-filter introduces less speech distortion compared to standard MWF but has more residual noise. For the case of strictly no speech distortion when $\mu \to 0$, the SDW-MWF reduces to the MVDR beamformer (28). Finally, when $\mu > 1$, the post-filter has a more aggressive noise reduction than the standard MWF but at the cost of an increased speech distortion.

3.4 Near-end Listening Enhancement

We now turn to the NLE problem. We introduce the general idea and problem formulation behind NLE, and some background on existing methods.

We initially note that, in the NLE literature, it is common to assume the signal coming from the far-end is clean, i.e., that $Y_{k,i} = S_{k,i}$. This might be possible if the FSE has a very high performance or if the signal to be played out in the near-end environment is previously recorded in a silent environment. We also make this assumption in our NLE work in Paper B, as it provides the ability to focus on only the NLE scenario and its effects. However, we also note, that this assumption is often not valid, and the problem of dealing with both a noisy far- and near-end is a main part of our work, cf. Papers A, C, and D, we return to this problem in Section 4.

Since early contributions to NLE, e.g. [67–69], it is only in the recent decades that the problem has seen a significant revival and is a very active research area [14, 46, 70–111], see also the reviews in [112–114] and the contributions to the Hurricane challenges [115, 116], as well as the performance studies of [7, 11, 13, 117].

3.4.1 Problem Formulation

In contrast to the FSE scenario, at the near-end the interfering noises are physically present in the environment, and are mixed with the target speech after processing. Therefore, in NLE, the clean speech signal cannot be estimated by removing noise from a noisy input signal as done in FSE. Instead, before playback in the noisy environment, the goal of NLE is to increase SI and SQ by adaptively pre-processing the signal received from the far-end before playback while exploiting knowledge about the near-end environment. Hence, the

idea behind the NLE task is to determine a mapping, that alters the incoming signal, $Y_{k,i}$, such that the output signal $P_{k,i}$ has the maximal performance, i.e., SI, when being listened to in the near-end noise, $N_{k,i}$. That is, the task is to solve the following maximization problem,

$$\begin{aligned} \boldsymbol{\theta}^*_{k,i} = \underset{\boldsymbol{\theta}_{k,i}}{\arg\max} \quad & \mathcal{I}\left(f_{NLE}(Y_{k,i};\boldsymbol{\theta}_{k,i},\mathcal{S}_Y,\mathcal{S}_N),S_{k,i}\right) \\ \text{subject to} \quad & \mathcal{D}_l\left(f_{NLE}(Y_{k,i};\boldsymbol{\theta}_{k,i},\mathcal{S}_Y,\mathcal{S}_N),S_{k,i}\right) \leq 0, \quad l=1,\ldots,m, \\ & h_l\left(\boldsymbol{\theta}_{k,i}\right) \leq 0, \qquad\qquad\qquad\quad l=1,\ldots,p. \end{aligned} \tag{36}$$

Similar to the FSE case, $\mathcal{I}$ is an optimization target measure of the NLE performance, $\mathcal{D}_l$ are additional possible performance constraints on the NLE, and h_l are possible constraints on the NLE mapping parameters. These may also evaluate performance in relation to the target speech signal and the near-end noise.

An obvious approach to increasing SI in additive noise environments is to increase the playback level of the output speech, similar to the talker raising their voice or shouting in the cocktail party scenario. However, beyond a certain point increasing the playback level may not be possible due to loudspeaker overloading or unpleasantly high sound levels [92]. Therefore, many NLE approaches take a power constraint into account, such that equal power is maintained between the unprocessed and processed signals, thus avoiding the trivial solution of applying an infinite gain.

It also applies to the NLE optimization, that we might not directly solve (36) in order to determine the NLE mapping with more heuristic methods, but the idea remains the same. That is, determine a mapping that results in high SQ and/or SI for the output signal possibly subject to some constraints.

3.4.2 Existing Works

Existing NLE techniques can be categorized into two main classes: Heuristic methods and mathematical optimization based methods. We provide a short overview of each class in the following.

Heuristic Methods The natural redundancy of speech and the Lombard effect have been utilized in many heuristic NLE techniques resulting in great increases in SI. Many heuristic / expert-driven approaches are based on considerations of the second formant; high frequency and transient components; and consonants being important for SI, e.g., [67, 68, 71, 75–77, 81, 82, 86, 108, 118]. Particularly, the use of Spectral Shaping and Dynamic Range Compression (SSDRC) proposed in [82] to reallocate energy to the higher frequencies and voiced offsets and onsets (bursts and fricatives) has shown very good results [113, 116]. Furthermore, several works in NLE increase SI by mimicking the Lombard effect [73, 89, 101–103].

By considering SI to increase with audibility the early works of Sauert [72, 74] raise the average speech spectrum above the average noise spectrum. The speech spectrum is raised above the noise by redistributing energy across time-frequency to boost the SNR using a heuristic weighting of the time-frequency bins. More recently, it also been proposed to increase audibility by scaling the speech spectrum proportionally [99] or inverse proportionally to the near-end noise spectrum [100].

Mathematical Optimization Methods The second class of algorithms, where our work also lies, considers the direct solving of the NLE optimization in (36), where a target SI metric is maximized, usually subject to an equal power constraint.

Considering audibility as the decisive factor of SI, the earliest attempt at mathematical optimization based NLE of [69] optimized NLE according to the Articulation Index (AI) [47, 119]. The related SII [39], stands out as one of the most frequently utilized optimization targets for NLE in numerous studies, e.g., [78–80, 92, 93, 120]. An important measure used in our work is the ASII which was introduced for optimal NLE in [46]. In [84, 87, 91, 95] near-end SI is increased by optimizing the glimpse proportion metric [121], which measures the proportion of spectro-temporal regions where the SNR is greater than a pre-determined threshold. Hence, similar to the SII the glimpse proportion metric assumes that audibility is the decisive factor of SI [43]. Although the SII focuses on long-term SNR of each frequency band, and the glimpse proportion metric is concerned with the proportion of short-time frequency SNRs above a given threshold [43].

The approaches of [14, 83, 96, 97] use a perceptual distortion measure [122] to derive optimal reallocation of speech energy in the time-frequency domain. In [88] this is expanded to also include reverberation. Extending on this, [90] considers multizone NLE using a general smooth distortion measure in the form of the ℓ_2-norm in the time-frequency domain.

Finally, recent studies have focused on the use of mutual information [123] as optimization criteria for SI [6, 94, 124]. Here, the idea is an information theoretic consideration of speech communication, where SI measures the amount of information conveyed from talker to listener. Thus, from an information theoretic perspective, maximizing SI is equivalent to maximizing the mutual information, i.e., conveying the largest possible amount of information across the speech communication channel.

Active Noise Cancellation The above NLE methods, perform various speech modifications to the input signal in order to make it more intelligible within the near-end noise. However, as an alternative to the speech modification based methods, it is possible to use Active Noise Cancellation (ANC) [105, 125–127],

where the goal is to cancel the near-end noise by adding an anti noise-phase component to the speech signal before it is played out in the noisy environment. In this thesis we focus purely on speech modification based methods, hence a thorough review of ANC is beyond the scope of this work, but we note that ANC provides a viable alternative or addition to NLE if the near-end listening environment can facilitate it with, e.g., headphones.

3.5 DNN based Methods

Although this work does not use DNN based approaches, it is important to mention them, since DNN based FSE and NLE have become very active research branches in the last decade [54, 56, 61, 104–107]. These DNN based approaches show great performance when optimizing for advanced SQ and SI predictors such as ESTOI [48]. Furthermore, even though many DNN based FSE methods focus on SQ in the optimization, they also show an increase in SI [56, 128]. However, recent results [60] have shown that the performance increases, obtained by DNNs, in the advanced predictors, do not always translate to performance increases in subjective listening tests. In addition, many state-of-the-art DNNs consume large amounts of memory and may struggle to generalize to new acoustic scenarios. Finally, the processing conducted by DNNs can be very difficult to interpret.

4 Joint Far- and Near-end Enhancement

In this Section we formally introduce the concept of joint far- and near-end speech and listening enhancement that we consider in this thesis. Initially, we present a categorizing of processing scenarios that help facilitate the discussion of processing requirements in various enhancement approaches. We then explicitly define the joint far- and near-end enhancement problem and relate it to the previously seen separated FSE and NLE approach, and then introduce the motivation behind the joint approach. To conclude this section, we conduct a literature review of existing studies on joint far- and near-end enhancement.

4.1 Informed Speech and Listening Enhancement

Before defining in the detail the joint far- and near-end speech enhancement problem, we first present a categorization of processing scenarios depending on which side information is available to the different parts of the system.

The idea behind joint processing is to optimize for multiple scenarios simultaneously by taking more things into account in the optimization. However, to take something into account you have to be "aware" of or "informed" about

it first. We refer to any such additional knowledge as side information. For example, side information could be the noise or signal CPSD matrix from the far- or near-end. It could also be knowledge about what processing has ocurred or will occur at the other end, and if the processing is invertible. Furthermore, it could be important information about the output of the processing, e.g., remaining speech and noise power in parts of the spectrum.

The opposite of joint processing is *blind processing*, where no side information is shared between the far- and near-end, cf. Section 4.2.1 for a detailed definition of blind processing. However, in addition to the two opposing scenarios of joint and blind processing, we can also consider a third scenario: *informed processing*, where only some side information is shared between the far- and near-end. That is, we can have awareness of side information in only parts of the processing chain. We illustrate four possible scenarios of side-information exchange between the far- and near-end in Fig. 10.

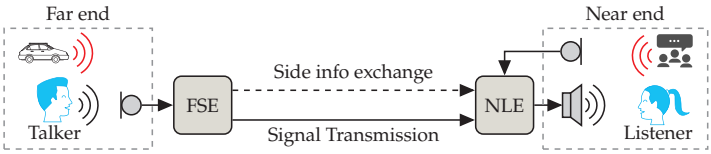
First, in Fig. 10a no side information is exchanged between the far- and near-end and both ends are unaware of what happens at the other end. This is the classic blind processing scenario, cf. Fig 8. Here we can only do the best possible at either end while assuming the other end is perfect, or if necessary estimate any missing information with only the knowledge and signals available at a single end.

In the second scenario, Fig. 10b, the near-end is informed of side information about the far-end and can optimize the NLE in accordance. For example, the near-end could receive information about the remaining far-end noise after FSE and correct speech levels in parts of the spectrum. However, the FSE is blind towards the NLE and cannot optimize its processing for the near-end and must optimize the processing blindly. Thus, this scenario does not always allow for fully joint processing, for example if the FSE is not optimal for the near-end scenario and the NLE cannot invert the FSE. However, if the FSE is invertible and sufficient information is exchanged (and we assume a lossless signal and information transfer) then the NLE could act as jointly optimized processing.

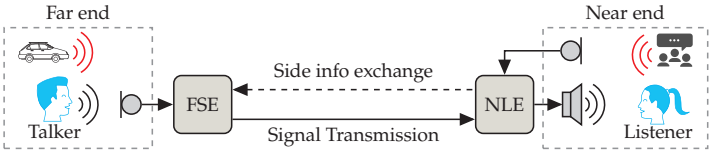
In the third scenario, Fig. 10c, the far-end is informed of side information about the near-end and can optimize the FSE in accordance. However, the near-end is blind towards the far-end. For example, this can occur in a broadcast scenario where the FSE is knowledgeable about multiple near-ends and optimizes its processing to be best on average, and each NLE compensates blindly for each user. Similar to the above case, fully joint processing is not possible because the NLE can only assume the signal coming from the far-end is clean speech. Alternatively, the NLE could assume the received signal is noisy and apply a second instance of noise reduction, however, this also leads to a loss in overall performance [8].

Finally, in the last scenario, Fig. 10d, side information is shared in both directions between the far- and near-end. Because both ends are fully aware

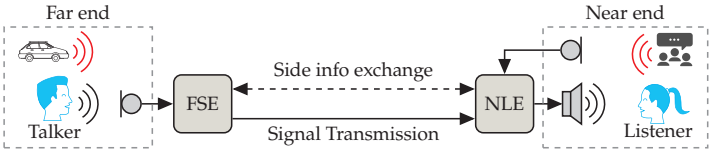
(a) FE **unaware** of NE and NE **unaware** of FE. The fully blind scenario.



(b) FE **unaware** of NE and NE **aware** of FE. Fe is blind towards the NE.



(c) FE **aware** of NE and NE **unaware** of FE. NE is blind towards the FE.



(d) FE **aware** of NE and NE **aware** of FE. Fully joint is possible.

Fig. 10: Illustration of how side information might be shared in different directions for the four different scenarios of awareness.

of each other jointly optimized processing is possible. That is, we can derived an optimal processing that compensates for all noise types and ensures all processing steps work together. From the above, we also note that informed processing is a subset of joint processing. That is, all joint processing scenarios are informed processing scenarios, but not all informed processing is joint processing depending on the level of information exchange.

Although we illustrate the side-information exchange as going all the way between the far- and near-end in Fig 10, it does not necessarily have to be implemented as such in practice. We could imagine that the processing takes place at some centralized processor in the speech communication chain with more convenient access to side information. Alternatively, by placing processing correctly and transferring the correct information and signals, joint processing

Fig. 11: Signal Model with joint far- and near-end speech and listening enhancement.

can be placed into just the FSE or NLE block. In this thesis, we do not consider how to optimally place the processing blocks in the communication system. However, we note that it is important to consider in practice as it has a great impact on the overall performance in end-to-end speech enhancement [8, 35]. In addition to not considering the optimal placement of processing, it is common in the joint far- and near-end enhancement literature to assume that the correct knowledge or side information is available at the right place at the right time. That is, we assume information is transferred without issues and any synchronization concerns can be handled. We also make these assumptions throughout our works and in the following. However, how to handle these concerns are important topics in practice and for future research, cf. Section 8.

We are mainly concerned with the differences between joint and blind processing and discuss these differences in next sections. However, in Section 4.4, we also consider how existing methods on joint far- and near-end enhancement can be categorized in relation to the informed processing scenario.

4.2 Joint Enhancement Problem Formulation

We now present the joint far- and near-end speech enhancement problem formulation in a similar manner to the FSE and NLE problems.

We consider the signal model shown in Fig. 11, which is similar to the signal model introduced Section 3.1, apart from the use of a new *joint far- and near-end processing* block. Here the noisy multi-microphone signal, $\boldsymbol{X}_{k,i}$, is directly mapped to a signal, $P_{k,i}$, that is an estimate of the target speech, $S_{k,i}$, and is played out in the noisy near-end environment, i.e.,

$$P_{k,i}^{joint} = f_{joint}\left(\boldsymbol{X}_{k,i}; \boldsymbol{\phi}_{k,i}, \boldsymbol{\mathcal{S}}_S, \boldsymbol{\mathcal{S}}_U, \boldsymbol{\mathcal{S}}_N\right), \tag{37}$$

where the joint mapping is parameterized by $\boldsymbol{\phi}$.

Often joint processing consists of two modules: one that performs FSE and one that performs NLE. Hence, the optimal mapping parameters, $\boldsymbol{\phi}^*$, can be split into the parameters for each part, i.e., $\boldsymbol{\phi}^* = \left[\boldsymbol{\phi}_{FSE}^*, \boldsymbol{\phi}_{NLE}^*\right]$. These are used to perform the various processing steps of the joint in the optimal order. Therefore, we might refer to an FSE and NLE part of the joint processing.

The task of the joint mapping, f_{joint}, is to simultaneously perform both the FSE and NLE tasks. That is, map $\boldsymbol{X}_{k,i}$ to a signal, $P_{k,i}$, that is simultaneously

both an optimal estimate of $S_{k,i}$ and such that final signal received by the listener, $Z_{k,i}$, has the best possible performance. Therefore, f_{joint} might also take as input relevant side-information of both the clean speech, S, the far-end noise, U, and the near-end noise N. Hence, the task is to solve the optimization problem,

$$\begin{aligned} \boldsymbol{\phi}^*_{k,i} = \underset{\boldsymbol{\phi}_{k,i}}{\arg\max} \quad & \mathcal{I}\left(f_{joint}(\boldsymbol{X}_{k,i};\boldsymbol{\phi}_{k,i},\mathcal{S}_S,\mathcal{S}_U,\mathcal{S}_N), S_{k,i}\right) \\ \text{subject to} \quad & \mathcal{D}_l\left(f_{joint}(\boldsymbol{X}_{k,i};\boldsymbol{\phi}_{k,i},\mathcal{S}_S,\mathcal{S}_U,\mathcal{S}_N), S_{k,i}\right) \leq 0, \quad l = 1,\ldots,m, \\ & h_l\left(\boldsymbol{\phi}_{k,i}\right) \leq 0, \qquad\qquad\qquad\qquad\quad l = 1,\ldots,p. \end{aligned} \tag{38}$$

We can consider the optimal parameters, $\boldsymbol{\phi}^*$, and hence optimal mapping, f^*_{joint}, as determined by a single function, ζ, of all relevant signals, S, U, N, by solving the optimization problem (38). That is,

$$\boldsymbol{\phi}^* = \zeta_{(38)}(S, U, N) \tag{39}$$

Thus, the optimal parameters and mapping are derived jointly because all processing effects and noise sources are taken into account simultaneously in the optimization.

4.2.1 Definition of Blind Enhancement

Conversely to the joint scenario above, in blind processing, cf. Fig. 8, the NLE output is the output of a composite function of the FSE and NLE [129, 130], i.e.,

$$P^{blind}_{k,i} = f_{blind}\left(\boldsymbol{X}_{k,i}\right) = \left(f_{NLE} \circ f_{FSE}\right)\left(\boldsymbol{X}_{k,i}\right) = f_{NLE}\left(f_{FSE}\left(\boldsymbol{X}_{k,i};\boldsymbol{w}_{k,i}\right)\boldsymbol{\theta}_{k,i}\right), \tag{40}$$

where the optimal FSE and NLE mappings are determined independently of each other. First, the optimal parameters for the FSE mapping, $\boldsymbol{w}^*$, are determined as a function of only the relevant far-end signals, i.e, the speech, S, and far-end noise, U, by solving the optimization problem (25). Secondly, the optimal NLE parameters, $\boldsymbol{\theta}^*$, are determined as a function of only the signal coming from the far-end, Y, and the near-end noise, N, by solving the optimization problem (36). That is,

$$\boldsymbol{w}^* = \zeta_{(25)}(S, U), \quad \text{and} \quad \boldsymbol{\theta}^* = \zeta_{(36)}(Y, N) \tag{41}$$

Hence, the FSE optimization is blind towards the processing and noise situation occurring at the near-end. Similarly, the NLE optimization is blind towards the processing and noise situation at the far-end. Thus, we emphasize that the difference between blind and joint enhancement lies in the *optimization*, as illustrated in Equations (39) and (41). That is, how the optimal parameters are derived is the deciding factor.

4.3 Joint versus Blind Processing

The prevailing approach to speech and listening enhancement is blind processing, where the FSE and NLE are optimized separately. It is also the most simple processing, since it puts no additional requirements on the link between the far- and near-end, apart from those that might exist in terms of signal transmission. However, we now consider some of the consequences that might occur because of blindly optimizing the FSE and NLE modules and how these motivate joint processing.

At the far-end, the FSE (cf. Section 3.3) is often tuned towards an aggressive noise reduction that removes as much far-end noise as possible at the cost of an increase in speech distortions [9, 10, 16]. However, in the end-to-end communication scenario, where noise is also present at the near-end, residual far-end noise at sufficiently low levels might be partially obscured by the near-end noise and there might not be a need for such an aggressive noise reduction [9, 10].

At the near-end, the NLE (cf. Section 3.4) typically operates under the assumption that the far-end signal is devoid of noise, which can potentially lead to erroneously interpreting noise as speech. Furthermore, as the NLE is often designed to maximize SI by amplifying the signal coming from the far-end to overpower the near-end noise, the loudspeaker might play back amplified noise, distorted speech and musical tones [9, 10].

Furthermore, the signal coming from the far-end sets an upper limit on the effectiveness of the NLE performance [7, 124], see also Paper D. Hence, if using an MVDR or SDW-MWF with low speech distortion, it is possible that not enough far-end noise is removed given that there will be further disturbances at the near-end. On the other hand, if the environmental noise at the near-end is minimal compared to the existing far-end noise in the signal, enhancing the near-end SI by increasing power yields little advantage. Instead, it is probably more advantageous to amplify channels with a high far-end SNR [124]. In addition, when the far- and near-end are blind to each other's presence, multiple instances of FSE and NLE may occur and lead to degradations in the enhancement performance [8].

Because joint processing can take all noise sources and processing steps into account simultaneously, the goal of joint enhancement is to better balance between noise reduction, speech distortions and compensation for the far- and near-end noise to achieve an overall better performance than in the blind setting. For example, we can optimize for both the far- and near-end noise simultaneously, or we can make sure the NLE part is aware of the amount of far-end noise that remains after the FSE and takes this into account in the optimization.

The idea behind joint processing, of taking multiple, if not all, aspects of the FSE and NLE problems into account simultaneously, necessitates the existence

of something to take into account. However, if there is no noise, or only noise with very low power, in either the far- or near-end it is no longer necessary to take the noise into account. This means that as the SNR at either the far- or near-end rises, the joint problem converges towards either of the single-ended FSE or NLE problems. Thus, the performance of blind and joint processing methods converge at the higher SNRs, as long as they are similar in nature in their respective FSE and NLE parts [6, 124, 131], cf. Papers C and D. Furthermore, if the FSE reduces the far-end noise without speech distortion using, e.g., an MVDR beamformer, then as the number of available microphones increase the MVDR can remove considerably more noise without introducing speech distortions [59]. Thus, if we have the ability to add more microphones, or in other ways improve the FSE in a distortionless manner, such that we can transfer an almost clean speech signal to the near-end, we might not need to use a joint processing setup and can perform classic NLE. Hence, from a noise reduction and listening enhancement perspective, joint processing is expected to be most effective, when noise is present in both the far- and near-end environment. If noise is not present or if we can sufficiently suppress noise in a mostly distortionless manner, it might not be worth it to use joint processing since joint processing puts additional requirements upon the processing chain to work in practice.

In this thesis, we focus only on joint far- and near-end enhancement from the perspective of compensating for environmental noise in relation to the FSE and NLE problems. However, as mentioned, the speech communication channel consists of several other factors, notably a need for speech coding and transmission over a often wireless channel. Hence, joint processing could naturally be extended to also include some of these aspects since joint processing already puts some requirements on the transmission of side-information, cf. future research in Section 8.

4.4 Review of Joint Enhancement Literature

We present an overview of existing work in the area of joint far- and near-end speech and listening enhancement. The topic of joint far- and near-end enhancement is relatively new and has not been widely studied prior to this thesis, with only a few prior works appearing in 2016 [124] and 2017 [6, 9, 132]. However, the topic has received new interest in recent years where more studies contemporary to this thesis have appeared [8, 131, 133].

In Table 1, we categorize existing studies according to the four different awareness scenarios from Fig. 10 and the methodology of each study. Although some studies formulate a joint enhancement optimization problem similar to (38), i.e., with the intention of optimization for conditions at both ends simultaneously, the final algorithm might not be fully joint. That is, the optimal solution to the optimization problem does not involve side-information

Paper	Problem Formulation	Final Algorithm	Methodology	Channels
[6, 124]	10d: Fully joint	10b: FE unaware of NE & NE aware of FE	Classic	Multi
[9]	10d: Fully joint	10d: Fully joint	Classic & heuristic	Single
[132]	10d: Fully joint	10b: FE unaware of NE & NE aware of FE	Classic & heuristic	Single
[131]	10d: Fully joint	10b: FE unaware of NE & NE aware of FE 10a: Blind FSE & NLE modules	DNN	Single
[133]	10d: Fully joint	10a: Blind FSE & NLE modules	DNN	Single
[8]	10b: FE unaware of NE & NE aware of FE	10a: Blind FSE modules	DNN	Single

Table 1: Categorization joint enhancement works in terms of the problem formulations and final algorithms according to the side-information awareness scenarios presented in Fig. 10, as well as methodology for deriving algorithms.

exchange in both directions, e.g., the FSE module might not need to be aware of the NLE module. Therefore, we categorize both the initial problem formulation and final algorithm of each study according to which awareness scenario is required for the full problem formulation, and which awareness scenario best illustrates what is taken into account in the processing steps of the final algorithm. The methodology of each study is categorized into three categories: (1) Classic signal processing based optimization, (2) heuristic methods, and (3) DNN based methods. Furthermore, we note if a study consider multi-channel or single-channel enhancement.

4.4.1 Classic Multi-Channel Enhancement

Of the existing studies only one considers multi-channel enhancement at the far-end. Namely, Khademi et al. [6], which is an extension of [124], where they perform joint multi-channel FSE and NLE based on a classic signal processing optimization problem formulation similar to (38). Here the goal is to maximize an approximation of mutual information under an equal power constraint. The optimization problem includes all far- and near-end noise sources. Additionally, [6] introduces the concept of production noise to mimic variations in speech patterns, which is also included in the optimization problem. Hence, the proposed optimization problem clearly falls into the category of a fully joint problem. The solution is an MVDR beamformer followed by an NLE subband gain. The far-end MVDR is not affected by the near-end noise or NLE gain. However, the NLE gain is a function of the remaining far-end noise, speech and near-end noise. Thus, the final algorithm can be implemented in a way similar to Fig. 10b, where the FSE module is fixed and depends only on speech and far-end noise, while the NLE module receives side-information from the far-end and changes its processing when the FSE processing changes. The results of [6] showed small performance gains in objective SI metrics over similar methods where the remaining far-end noise was not transferred to the NLE

module. That is, comparisons were made between transferring knowledge of both the far-end noise and speech signal, and only transferring knowledge of the speech signal to the proposed NLE module. Hence, no comparison was made against fully blind methods, where a noisy far-end signal is processed as if it is clean speech. Furthermore, all tested methods showed only small performance increases over the unprocessed noisy signals. Additionally, an informal SI listening test with only seven participants was used to indicate statistically significant differences between the proposed and reference methods for low SNRs. Furthermore, it is noted that the proposed production noise is a weighting parameter for the importance of a particular subband similar to the SII [6]. Finally, the proposed maximization of mutual information requires several approximations and assumptions regarding signals and their distributions, see also Paper A.

4.4.2 Classic Single-Channel Enhancement

The remaining existing joint enhancement studies consider single-channel FSE. We further group these according to if they use classic signal processing or DNN based methods. In this subsection, we first consider the classic signal processing based methods.

Niermann et al. [9] considers single channel joint FSE and NLE by jointly optimally controlling a FSE single channel Wiener filter with an SII based NLE subband gain. They do not directly maximize a cost function, but solve a system of equations that were heuristically determined to provide good SQ and SI. The proposed optimal joint control adopts to knowledge of both the (estimated) far-end noise, near-end noise and speech power. Hence, both the problem formulation and final algorithm fall in the category of fully joint algorithms. The study in [9] reported brief results stating the proposed joint approached was preferred in informal listening tests for SI and SQ in comparison with a fully blind Wiener filter and SII based subband gain.

Zorilă and Stylianou [132] also consider single-channel joint FSE and NLE. Here the NLE-only method SSDRC [82] is expanded to improve SI when the incoming far-end signal is noisy and also improve SQ in quiet conditions. This is done by introducing Multi-Band SSDRC (MBSSDRC) and combining it with classic single-channel FSE based on Minimum Mean-Squared-Error (MMSE) optimization [134]. Hence, the work of [132] combines the heuristic based SSDRC with classic FSE optimization. First FSE is performed while unaware of the down stream NLE module and near-end noise, then side-information is passed forward along with the enhanced signal from the FSE module to the NLE module where MBSSDRC takes place. Thus, we categorize the original problem formulation as fully joint and the final algorithm of [132] is categorized similar to Fig. 10b, where only the NLE module is aware of FSE module but not vice-versa. However, we note that both SSDRC and MBSSDRC are heuristic

methods that are independent of the near-end noise. That is, the processing is only depends on the received far-end signal irregardless of the amount of near-end noise, which is also part of the motivation for [132] to expand to the multi-band case. The results show, for quiet near-end conditions, that the proposed method had better SQ than standard blind FSE and NLE with SSDRC in subjective listening tests. For noisy near-ends, the proposed method improved subjective SI of the noisy signal but was still surpassed by blind noise reduction and SSDRC.

4.4.3 DNN based Single-Channel Enhancement

In this subsection we present the DNN based joint enhancement studies, which can also be considered as contemporary to this thesis, as they were published after the start of this PhD study.

Shifas et al. [133] presents a fully DNN based single-channel method for joint FSE and NLE. Here a DNN is trained via a teacher-student approach with the SSDRC [82] as the NLE teacher. That is, the DNN takes a noisy speech signal as input, and must then remove noise and process the signal such that it mimics the output of the SSDRC [82] NLE method. Hence, the proposed DNN, similar to SSDRC, is independent of the near-end noise. The results show that performance is improved compared to MBSSDRC [132], but without comparison against other joint or blind enhancement methods.

We categorize the original problem formulation of [133] as a fully joint problem, as the intention is to enhance performance in the presence of both far- and near-end noise. However, the categorization of the final processing is more ambiguous. On the first hand, during training, the different blocks inside the DNN are optimized simultaneously for the optimal end-to-end performance. That is, the later stages of the DNN are optimized given that a particular processing has already occurred in the early layers. However, on the other hand, because the proposed method consists of a single DNN that only takes the noisy far-end speech as input, and has no need for any additional side-information to be directly passed along to parts of the network, it could in principle be placed anywhere in the processing chain. Hence, as the presented DNN is a black box and it is not possible to tell if any side-information is passed forward inside the network, we resort to the simplest explanation or categorization, which is blind processing (Fig. 10a). This is also easier to implement than informed processing which requires transfer of additional side-information.

Li et al. [131] presents single-channel DNN based joint FSE and NLE, where the DNN consists of an FSE and NLE module. The FSE modules takes noisy far-end speech as input and presents an enhanced speech signal. The enhanced speech signal is then processed by the NLE module to produce a signal with high SI in a given near-end noise, which is also given as input to the NLE mod-

ule. Thus, contrary to [133], the processing adapts to the near-end noise. Two versions are presented: an informed version, where a noise token (knowledge of the far-end noise) is passed to the NLE module, and a version with fully blind modules, where signals are passed forward without side-information. There are no skip connections between the FSE and NLE modules. The DNNs are trained by maximizing multiple SQ and SI estimators when noise is present at both the far- and near-end. Thus, similar to above, we categorize the problem formulation as a fully joint problem because all noise sources and processing steps are taken into account in the training. However, the final trained DNN versions are clearly an informed version with noise token passed forward to the NLE module (Fig. 10b) and a fully blind version without side-information transfer. The results of [131], show that both proposed versions perform better blind concatenation of DNNs that were trained separately and also against blind concatenation of a single-channel Wiener filter and the SSDRC [82]. Furthermore, the informed version of the proposed method performed slightly better than the blind version. Performance gains were seen in both objective SQ and SI metrics and in preference tests. However, no performance comparisons were made against other joint approaches.

Finally, we consider the work of Tan and Wang [8]. This study does not consider joint FSE and NLE, but instead considers the consequences of blindly conducting single-channel FSE twice. Hence, it is very relevant for justification of using joint or more informed approaches. For example, if an NLE module assumes it receives a noisy far-end signal, as in [132], and then conducts some form of FSE or noise reduction prior to its NLE step. Particularly, [8] shows that the FSE performance degrades severely when blind FSE is performed twice. To remedy this, they propose a new strategy for downstream DNNs conditioned on the previous enhancement, by extending the training data with already enhanced noisy speech samples. The results show that the downstream DNN is more robust to processing artifacts caused by upstream FSE. The problem formulation of [8] falls in the category of informed processing (Fig. 10b) as the downstream DNN is trained with knowledge of processing artifacts. However, at implementation time the downstream DNN does not receive any side-information from the upstream, hence it can be implemented blindly.

The above review and Table 1, show that most existing works do not require a fully joint implementation of the final algorithm. Furthermore, the results show that considering joint information in the problem formulation increases the final performance compared to the blind approach. However, it is noteworthy that the final algorithm does not necessarily require a fully joint side-information exchange to realize this performance gain. This is beneficial, since, as mentioned earlier, fully joint algorithms put more requirements on the final implementation due to the necessary side-information exchange, than algorithms with only forward-flowing information. However, when the

final algorithms are blind or only have side-information passed forward, the FSE part is fixed and does not update its processing with changing near-end conditions and may result in suboptimal performance for the given situation. Furthermore, for the DNN based approaches, processing is optimized during training but fixed at inference time, which poses generalization issues when subject to unseen conditions. Thus, the preferable algorithm category is not evidently clear and involves a trade-off between the generalization to new scenarios against implementation complexity. In the next Section, we present a perspective on how to handle that processing requirements change depending on environmental noise situation.

5 The Minimum Processing Perspective

In this Section, we present and motivate the concept of *minimum processing*, first introduced by [16] for FSE, and how it relates to existing works on NLE and joint enhancement in relation to preserving both SI and SQ.

5.1 Motivation

The effects of both SI and SQ are important for the overall experience of the listener. However, the significance and relative importance of SI and SQ dynamically shifts in response to varying environmental noise conditions [11–14]. In environments with high noise levels, clear and intelligible speech is essential for effective communication, and high SI can significantly impact the listener's perceived SQ, thus highlighting the crucial contribution of SI to SQ in noisy circumstances [12, 13]. Conversely, in quieter settings or environments with minimal background noise, the SI naturally approaches 100 % and the importance of SQ becomes more pronounced as further SI enhancements are unattainable [11–16, 129, 130]. While SI remains important for conveying information clearly, SQ influences the overall listening experience to a greater extent. Hence, in quieter conditions, listeners are more attuned to subtle nuances in speech SQ, such as clarity, naturalness, and particular absence of distortions [11–14, 132].

Generally, the goal of NLE methods has been to always maximize SI as seen in Section 3.4. The processing necessary to increase SI potentially introduces speech distortions, although these distortions might be masked or a necessary price to pay in lower SNRs, in high near-end SNRs they can lead to diminishing SQ [13, 14]. Furthermore, in blind NLE processing, when the far-end signal is wrongly assumed to be clean, the distortions caused by processing a noisy signal as if it is clean also degrade SQ, especially in high near-end SNRs [132].

Similarly, the goal of most FSE methods is to maximizing noise reduction, as seen in Section 3.3. That is, the goal is to leave only the clean speech signal

behind by designing FSE methods according to the inherent undesirability of noise [16]. However, this may come at the cost of sometimes severe distortions [16, 23]. This is apparent especially for the popular MMSE based Wiener filter, and is why the SDW-MWF was introduced to control the trade-off between noise reduction and speech distortion [66]. Furthermore, considering all noise as undesirable may lead to loss of otherwise important contextual noise [16, 135]. For example, in the hearing aid scenario, always removing all noise might lead to users feeling isolated or, e.g., make it difficult to navigate in traffic [16]. In a similar manner, for communication between two physically separated environments, such as we consider in this thesis, it might be important for the listener to receive some important cues about the talker's environment to help facilitate a common ground [135].

A particular way to help better control the trade-off between SQ and SI and alleviate excessive processing is to apply a *minimum processing principle* [16], which we introduce in more detail in the following.

5.2 Minimum Processing Principle

The concept of minimum processing beamforming was recently proposed by [16] to mitigate the effects of excessive processing in multi-channel FSE. Here, the goal is to ensure that the beamformer output, $Y_{k,i}$ is minimally processed compared to a certain reference signal, $S^R_{k,i}$, as long as a given performance criterion is fulfilled [16]. The reference signal, $S^R_{k,i}$, is a signal with some desired properties that the beamformer output should mimic as long as the performance constraint is satisfied. Particularly, the performance constraint may be an SI criterion. This is formulated by [16] as an optimization problem that must be solved for each subband, j, i.e.,

$$\begin{aligned} \underset{\boldsymbol{w}_{k,i}, k\in\mathbb{B}_j}{\arg\min} \quad & \mathcal{D}\left(\boldsymbol{Y}_{j,i}, \boldsymbol{S}^R_{j,i}\right) \\ \text{subject to} \quad & \mathcal{I}\left(\boldsymbol{Y}_{j,i}, \boldsymbol{S}_{j,i}\right) \geq I'_j, \end{aligned} \tag{42}$$

where $\boldsymbol{Y}_{j,i}$ is a vector containing all $Y_{k,i}$ for which $k \in \mathbb{B}_j$, and similarly for $\boldsymbol{S}_{j,i}$ and $\boldsymbol{S}^R_{j,i}$. Here, $I'_j \triangleq \min(I_j, I^{max}_j)$, where I_j is a preselected minimum requirement on the performance, and I^{max}_j is the maximum possible performance achievable for the current far-end input SNR, when ignoring the processing penalty [16]. Thus, contrary to the standard beamformers in Section 3.3.2, the minimum processing principle utilizes the performance constraint of the general FSE optimization problem in (25) to ensure high SI while also preserving a desired performance characteristic. Therefore, we refer to the standard FSE, NLE and joint enhancement topics of Section 3 and 4 as *maximum processing*.

To validate the minimum processing principle, [16] showed results for two important cases: an ambient preserving mode and an aggressive mode. For

both cases the MSE was used as processing penalty, $\mathcal{D}$, and the SII audibility was used as performance measure, $\mathcal{I}$, where the subband SDR, ψ^F, was used instead of the SNR, ξ^F, (cf. Section 3.2). In ambient preserving mode, the reference signal, $S^R_{k,i}$, is chosen as the noisy signal at a reference microphone [16]. Hence, the beamformer aims to processes the noisy speech as little as possible, as long as the SI target is met. Thus, if there is little noise in the microphone signal and the unprocessed signal has sufficient SI, the beamformer does nothing. In the aggressive mode, the reference signal is chosen as the output of an aggressive SDW-MWF with $\mu \gg 1$ [16]. This results in a beamformer that removes all far-end noise unless the subsequent distortion of speech signal results in an SI below the required threshold.

To better understand, how the minimum processing performance changes with varying input SNR, we first define I_j^{min} as the minimum possible performance achievable for the current far-end input SNR when the processing is minimized while ignoring the performance constraint. Then, denoting the performance of the optimal beamformer solution (42) by I^*, we have that

$$I^* = \max(I'_j, I_j^{min}) = \max(\min(I_j, I_j^{max}), I_j^{min}). \tag{43}$$

Now, assuming $I_j^{min} < I_j^{max}$, we note, from the definition of I'_j, that the minimum processing beamformer does not guarantee to always have a performance higher than the desired target, I_j, but only guarantees a performance higher than the minimum of I_j and I_j^{max}. This is illustrated in Fig. 12, where if for a given far-end input SNR, ξ_j^{FE}, the maximum performance achievable, I_j^{max}, is lower than the target, I_j, then the performance of the optimal beamformer I^*, is equal to the maximum performance, i.e., $I^* = I_j^{max}$. It is only when the input SNR is high enough such that $I_j^{max} \geq I_j$, that the beamformer is guaranteed to always achieve a performance greater than or equal to the desired performance, I_j. Particularly, as the input SNR increases, then the minimum performance, I_j^{min}, might become greater than I_j, i.e., $I^* = I_j^{min} \geq I_j$, and the optimal feasible solution is that which minimizes the processing penalty.

5.3 Review of Related Existing Works

Zahedi et al. [16] introduce the minimum processing principle for FSE. However, there has been no direct formulation or usage of the minimum processing principle in either NLE or joint enhancement. The main reasoning behind minimum processing is that of improving SQ by limiting distortions when SI is high. Therefore, it is relevant to consider how other works have looked at balancing improving SI while limiting SQ degradations due to distortions in NLE and joint enhancement. Particularly, as some existing works can be said to fall into the category of minimum processing enhancement even though this was not the original goal of these works.

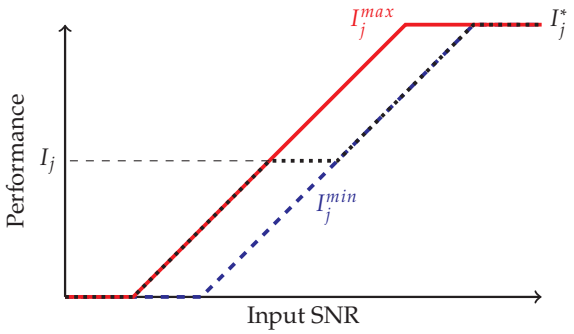


Fig. 12: Example illustration of the performance of the optimal minimum processing solution to (42) in an ideal setting. Inspired by results in [16, Fig. 4].

5.3.1 Minimum Processing in NLE

Although many existing NLE works are designed with the intention of maximizing SI [112–116], the processing does not always continue infinitely. Particularly, when the optimization metric has a build-in saturation where it clips. Thus, it can only be increased to a certain level and clipping occurs because processing is already maximized. For example, SII [39] (Section 3.2.2) based NLE algorithms, e.g. [80, 120], have this particular behavior. For NLE based on SII and similar metrics with clipping, the goal is still SI maximization, and the processing of the resulting algorithm only stops because of saturation build into the optimization measure. However, if the saturation level for the particular measure is not adjusted properly in the original design of the measure, this might still result in too much processing and a loss in speech quality.

5.3.2 Minimum Processing in Joint Enhancement

As we have a particular focus on joint far- and near-end enhancement in this thesis, we categorize the joint works introduced in Section 4.4 according to their processing goal and target metrics in Table 2. We note that the joint enhancement works of [6, 124] and [133] focus purely on maximizing SI. However, other joint enhancement take various steps to improve not only SI but also quality in the presence of near-end noise to various degrees.

The work of Niermann et al. [9] is similar to the minimum processing work we consider in this thesis, however [9] only considers single-channel FSE and does not directly solve a minimum processing optimization problem. As mentioned in Section 4.4, [9] considers joint control of a FSE single channel Wiener filter with a simplified SII based NLE subband gain, by solving a system of prioritized equations for each subband with the goal of minimizing speech distortions due to noise reduction while achieving a desired level of

SI. The constraints in prioritized order are: (i) not attenuation speech with the NLE gain, (ii) keeping the processed far-end noise below the near-end noise by a sufficient amount and (iii) speech overpowering the near-end noise by a sufficient amount. Hence, the goal is to always limits noise attenuation and speech distortions as much as possible, and then provide a sufficient SI [9]. Therefore, [9] provides a solution that is closely related to the minimum processing principle. However, the solution is not based on a particular sense of optimality and instead derived through heuristics and expert knowledge. In addition, [9] assumes the NLE gain does not result in any speech distortions, although if the NLE various greatly between subbands it can result in speech distortions. Furthermore, it is not considered if at certain points it is more beneficial to have more aggressive noise reduction at the cost of speech distortion and then limit the NLE gain instead. In addition, the SI constraint does not directly include the effect of the remaining far-end noise but only consider far-end noise power in relation to near-end noise power level. The SI constraint, and thereby achieved performance, is also not directly related to a desired level of SI or SNR target in an obvious way. Thus, the solution is not based on a particular sense of optimality, but derived through heuristics and expert knowledge.

Li et al. [131] utilize the power of large DNNs to optimize for a plethora of complex SI and SQ estimators simultaneously. This is possible with a Generative Adversarial Network (GAN) approach, where the generator produces enhanced speech and the discriminator mimics the objective metrics. Thus, by leveraging DNNs, [131] can choose to optimize for targets relating to both SI and SQ at the same time, with improvements achieved in both objective SQ and SI metrics and in subjective preference tests. However, as mentioned previously, it is not always that improvements in objective metrics with DNNs lead to improvements in subjective listening tests [60]. Furthermore, the black box nature of DNNs make it difficult to interpret and understand what processing steps lead to improvements.

Tan and Wang [8] use a DNN to maximize both SQ and SI when FSE is performed twice in a blind scenario, especially minimizing artifacts from multiple FSE instances. Particularly, they do so by minimizing the MSE between clean speech and speech estimated from either noisy signals or from signals that have already been enhanced once by various FSE methods. Hence, the goal is to both maximize SI and SQ, and minimize artifacts by minimizing a cost function. Thus, we categorize [8] as a maximum processing method, although the goal is also to minimize artifacts caused by excessive processing. Here [8] also has the strength of DNNs to achieve great objective SI and SQ performance, but also some of the same caveats as mentioned above.

The MBSSDRC of Zorilă and Stylianou [132] extends the otherwise NLE-only method of SSDRC to also focus on improving SQ in quiet near-end conditions when the received far-end signal is noisy. Particularly, by minimizing

Paper	Processing Goal	Target Metrics
[6, 124]	Maximize intelligibility	Mutual Information
[9]	Sufficient intelligibility with minimum artifacts	Joint control of SDW-SWF & Subband SII gain
[132]	Maximize intelligibility and minimize artifacts	Denoise: MSE Artifacts: SNR-adaptive gain thresholding
[131]	Maximize intelligibility and quality	Intelligibility: HASPI, ESTOI, SIIB Quality: PESQ, ViSQOL, HASQI, SI-SDR
[133]	Maximize intelligibility	Denoise and mimic SSDRC
[8]	Maximize intelligibility and quality	MSE

Table 2: Categorization joint enhancement works in terms of processing goal and target metrics.

artifacts from excessive NLE using gain thresholding depending on the SNR after MSE based noise reduction has been performed. The design of the MB-SSDRC consists of first classic maximum noise reduction by optimal MMSE then a heuristic expert knowledge approach to NLE with minimal artifacts. Thus, [132] includes both a maximum processing and minimum artifacts principle, although the minimal artifact NLE is not based on a particular sense of optimality.

6 Thoughts on Performance Comparisons

As mentioned in Section 3.2, it is important to evaluate the performance of speech and listening algorithms with both objective measures and subjective listening tests. In our work in Paper B, we conduct a listening test for SQ in the NLE case. Whereas in Paper D, we conduct listening tests for both SQ and SI in the joint far- and near-end enhancement case.

Speech and listening enhancement algorithms need to preserve or enhance both SI and SQ, since it is possible for speech to be highly intelligible but be of low quality [13, 23]. Therefore, it is important to measure other attributes than just SI [13]. However, SQ is highly subjective and difficult to evaluate reliably, whereas SI is an objective attribute and is more easily measured [23]. Furthermore, to the best of our knowledge, it is not very common to conduct SQ listening tests in situations where near-end noise is present, i.e., in NLE, or in the joint case where noise is present at both the far- and near-end, cf. Paper D. Therefore, in this Section we discuss evaluation of SQ in the presence of near-end noise, where best practices are unclear. To this end, we also discuss how the definition of SQ might differ between the perspectives of the far-end talker and near-end listener.

6.1 About the Definition of Speech Quality

The subjectivity of SQ largely stems from the fact that individual listeners have different standards for what constitutes "good" or "bad" quality [23]. SQ measures "how" speech is perceived and encompasses numerous dimensions and attributes, such as, e.g., naturalness and scratchiness [23]. Therefore, evaluations are often limited to a few dimensions, depending on the specific application [23, 40]. In this regard, how we define good SQ is of significant importance. In this discussion, we focus particularly on the perspective from which we define quality: the talker or the listener. SQ can have different meanings for the listener and the talker, and these meanings can vary depending on the communication scenario [13]. This is especially true when evaluating the performance of speech and listening enhancement algorithms.

Intrusive objective and subjective measures of SQ define basic audio quality as being the correct reproduction of the original reference signal [40, p. 370]. This does not not necessarily mean the correct waveform as in PESQ but at least some form of correct audio reproduction as in, e.g., [136]. Thus, all differences compared to the reference speech are considered as bad for SQ, as in the definition of the SDR. Hence, this definition of SQ can be related to the FSE goal of removing noise while preserving the talker's voice, or the speech coding goal of reproducing the original talker's voice with minimum artifacts. From the NLE perspective, this corresponds to the goal of increasing SI with minimum artifacts. Generally, we assume the preservation of the talker's voice can be related to the ability to recognize the talker, which for the talker is assumed to be a clear priority in non-critical communication scenarios, where SI is at a decent level.

The exact preservation of the talker's voice may not always be the main objective. Particularly, when considering NLE in noisy conditions, the main goal is often increasing SI no matter the costs to SQ, hence why the SNR is defined such that speech distortions are not considered as noise. This aligns with the listener's perspective, that in noisy situations it is more important to understand the speech than the talker's voice. Especially, if the listener does not know the talker. However, the listener may still prefer good SQ. The subjective SQ of the speech experienced by the listener can be quantified by a range of attributes describing speech and background noises without relying on the differences to the reference speech, c.f. [40, Sec. B.1]. However, if the listener knows the talker, the ability to recognize the talker and receive undistorted speech may become more important. These different perspectives on SQ depending on the talker and listener familiarity are summarized in Table 3.

Familiarity	Talker's perspective	Listener's perspective
Know each other	Desire no speech distortion and ability to be recognized.	Desire no speech distortion and ability to recognize talker.
Does not know each other	Desire no speech distortion and ability to be recognized.	Changes to voice irrelevant. Overall quality most important.

Table 3: The perspective of talker and listener on speech distortions and overall quality according to if they know each other or do not know each other.

6.2 Speech Quality Listening Tests with Near-end Noise

In both Paper B and D, we conducted listening tests for speech quality using the MUlti Stimulus with Hidden Reference and Anchor (MUSHRA) paradigm [137]. In a number of trials the participants were asked to evaluate audio quality on a scale from 0 to 100, segmented into five equal intervals denoted as *bad, poor, fair, good*, and *excellent*. Specifically, the participants were instructed to judge the *basic audio quality* in comparison to a known reference signal, without being provided any definition of audio quality. Each trial consisted of a reference signal and a number of signals to be rated including a hidden reference, a hidden anchor, an unprocessed noisy signal and a signal for each enhancement algorithm under test, cf. Papers B and D. The following discussion is based on observations from these tests and thoughts prior to these tests.

6.2.1 About the Anchor Signal

According to [137], the low anchor signal(s) should have impairments similar to those of the processing under test, while having limited content dependence. However, impairments due to the processing of most NLE and joint far- and near-end algorithms are content dependent, i.e., depend on both the speech and noise. Furthermore, the processing of different enhancement algorithms may also be vastly different [13]. Hence, it is difficult to create anchors using processing impairments that are fair to all algorithms. Since the environmental noise is independent of the speech content, it is, therefore, more suited to create anchors. Thus, anchors were created by adding significantly more noise than in the test-case SNRs. These very noisy signals are then not processed by any algorithm. Although this is not ideal, it was the most fair choice of anchor in the NLE and joint enhancement case.

6.2.2 About the Reference Signal

For subjective SQ evaluations in noisy near-end listening environments, the use of reference signals warrant some concern. Specifically, two concerns arise: firstly, if the test should include a known reference signal or not, and secondly

if a known reference signal should contain environmental noise. Reference signals play the particular role of being the signals to which all other signals are compared in a listening test [40]. Hence, the use of a reference signal plays a crucial role in connection with how we define (good) SQ.

Using a Reference Signal By using a reference signal, we give the listener the knowledge of what we determine as the best possible scenario, thus all differences between the original signal and the processed signal are considered bad. Hence, using a reference signal corresponds to a definition of SQ where we try to preserve the original signal as much as possible. The desire is to have the same SQ at both ends of the communication chain. That is, the talker sounds as much as possible as located in the same environment as the listener, and the talker's voice can be recognized, which can be of great importance or highly desired in some situations. When considering NLE, the environmental noise is added after the processing and might obscure some speech distortions [13]. Therefore, the use of reference signals can help measure if there are any detectable speech distortions because the participants are aware of what the undistorted speech sounds like and are instructed to focus on this as well.

However, not all changes to the speech signal are necessarily bad when listening in noise [13]. In NLE, the main objective is to increase the speech intelligibility which also affects the SQ [13]. Particularly, the NLE processing may introduce significant changes to the original speech that may still increase the quality of the speech above the unprocessed speech in the given noise situation, i.e., the speech could become more pleasant to listen to [13]. However, it is difficult to measure if participants prefer some processing when asked to compare with a reference, because they are forced to rank significant differences as bad, even if the more processed speech has a perceived higher quality. Particularly, we observed, that even when asked to judge "in relation to reference" some participants mentioned after the test, that they would judge a more processed version higher because it was more pleasant. Some also indicated that they would have judged a more processed version higher than they did, had they not been asked to compare to the reference. Therefore, it is important to properly instruct participants, such as to ensure they have the same behavior at all times.

Not using a Reference Signal As an alternative, by not using a reference signal we can consider a definition of SQ that does not put a particular focus on signal alterations. That is, we do not require the preservation of the original signal, but instead want to achieve the best possible SQ no matter what that is. This allows participants to grade signals independently of the original signal. Thus, the overall SQ is graded considering the fully combined experienced

of all processing artifacts and environmental noises. To still ensure a diverse grading scale usage, we may add high-quality anchors that are expected to score above the other signals, i.e., by still including a hidden reference signal, in addition to the low-quality anchors [40]. However, it is possible that the processed signals are graded higher than the high-quality anchor, as some participants may prefer some changes over the original signal, e.g., more bass.

Thus, when not using a reference signal participants grades may be more inconsistent, due to a varying definition of quality. Therefore, when not using a reference signal we need to be vigilant and enable checking of intra- and inter-rater reliability [23]. Furthermore, it was seen in [13], that even without being given a reference signal, when listening in quiet near-end conditions participants rated the clean unprocessed signal the highest. Hence, a certain standard of quality was learned by the participants [13]. However, the reasoning behind the participants scores of processed and unprocessed signals was also deemed to be unclear in certain situations [13].

Our work is focused on minimum processing and the preservation of the original clean speech as much as possible and limit speech distortions when SI is high. Therefore, we have used a reference signal, to enable better control of the grading scale and have a clear definition of high SQ in relation to the original speech signal. This corresponds to a listener centric view on SQ, where we assume some kind of prior knowledge of the talker's voice or at least a desire to recognize the talker. However, based on participant comments and the above discussion, it might be beneficial to consider alternatives to the MUSHRA paradigm for future SQ evaluation in near-end conditions. Promising alternatives are for example: using an overall MOS without any definition of quality as in [13, 132], using preference tests as in [9, 131], or using the SIG/BAK/OVRL paradigm [138] where a separate quality score is given to the signal, background noise and the overall signal.

Adding Noise to the Reference Signal When considering how to define the reference signal in near-end listening scenarios, we need to consider the noise situation, since the importance of SI and SQ differ depending on the noise conditions [13]. When designing NLE and joint enhancement algorithms these are designed for SI enhancement in noisy near-end environments, however when evaluating SQ it is often done in quiet near-end conditions when SI is maximized [13, 132]. Furthermore, when using a reference signal for SQ evaluation, it is common to consider the best possible scenario with no noise.

However, for the NLE case the noise is a significant part of the setup, and the intended use of the algorithms is in noisy settings. Hence, evaluating the algorithms outside these scenarios may be unfair to the algorithm. That is, any positive or negative effect of the NLE processing on the SQ should be considered across the overall acoustic scene, i.e., (processed) speech plus

noise. Therefore, it is interesting to consider if the reference should always be a perfect signal. However, if adding noise to the reference, it is possible we may obscure some parts of the clean speech signal. Thereby, any speech distortions that become audible in those regions after processing may not be detected by the participants.

Since most real use cases will involve different levels of background noise, listening situations without any noise are slightly unrealistic. Therefore, focusing solely on the ideal setting may limit the mapping of results to realistic situations. Furthermore, conducting listening tests closer to realistic scenarios improves reliability of results as indicators of performance, but evaluating performance in more specific scenarios also limit comparability of results between different listening tests.

The NLE work in Paper B is focused on minimum processing NLE in the presence of near-end noise. That is, the intended is preserve SQ when SI is high but noise is still present. Therefore, to evaluate SQ in relation to the original speech under the assumption that some noise is always present in the environment because it cannot be removed, we conducted a listening test where the reference signal was a noisy near-end signal with only low levels of noise instead of completely quiet conditions. We used such low levels of noise, that we expect similar results would have been achieved even without any noise added to the reference signal. However, by having noise on the reference we achieved a more realistic listening condition.

In Paper D, we consider joint far- and near-end enhancement. Here, there is a clear goal of transferring the far-end speech signal to the near-end listener with as little noise and distortion as possible. That is, the goal is to both remove far-end noise and also preserve SI and SQ in the noisy near-end environment. Hence, as in classic FSE [23], there is a clearer definition of the highest quality signal, i.e., the clean speech signal. Therefore, a clean speech signal was used as reference in Paper D. However, this also meant that no algorithm could ever achieve the highest quality level, because noise would always be present in the processed noisy signals, as the near-end noise could never be removed.

7 Scientific Contributions

The main body of this thesis consists of the collection of four accepted peer-reviewed papers in Part II, which documents the research contributions of our work. In this Section, we present the hypotheses and research questions, and the scientific contributions of the papers presented in Part II. Furthermore, we outline their relationship to the previously presented background topics.

The papers that constitute Part II are as follows:

[A] **A. J. Fuglsig**, J. Østergaard, J. Jensen, L. S. Bertelsen, P. Mariager and Z.-H. Tan, "Joint Far- and Near-End Speech Intelligibility Enhancement

based on the Approximated Speech Intelligibility Index," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[B] **A. J. Fuglsig**, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof and J. Østergaard, "Minimum Processing Near-end Listening Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[C] **A. J. Fuglsig**, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof and J. Østergaard, "Joint Minimum Processing Beamforming and Near-End Listening Enhancement," *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024.

[D] **A. J. Fuglsig**, Z.-H. Tan, L. S. Bertelsen, J. Jensen, J. C. Lindof and J. Østergaard, "Joint Far- and Near-end Speech and Listening Enhancement with Minimum Processing," *IEEE ACCESS*, 2024.

7.1 Hypotheses, Research Questions and Observations

The complete investigation spans the topics of NLE, joint far- and near-end speech and listening enhancement, and minimum processing based enhancement, following a logical progression revolving around the main hypotheses of this thesis:

(H.1) Joint enhancement provides better objective and subjective SI and SQ than blindly concatenated FSE and NLE.

(H.2) Minimum processing provides better SQ than maximum processing in quieter conditions while preserving high SI.

We address these hypotheses, by addressing the following individual research questions (RQ), observations (OB) and hypotheses (H) in each paper.

Paper A

(OB.A.1) Mutual information based joint enhancement [6] requires a complex signal model with production noise, several approximations and is closely related to ASII.

(RQ.A.1) What is the objective SI performance of optimal joint far- and near-end speech and listening based on ASII maximization?

Paper B

(OB.B.1) Minimum processing concept not formulated in NLE.

(H.B.1) Minimum processing NLE has higher SQ than maximum processing NLE with similar SI.

(RQ.B.1) What is the SI and SQ performance of optimal minimum processing NLE based on an ASII SI constraint and a MSE cost function?

Paper C

(OB.C.1) Minimum processing concept not formulated for joint enhancement.

(H.C.1) Joint minimum processing has a higher objective SI than blind minimum processing.

(H.C.2) Joint minimum processing has a higher objective SQ than blind minimum processing.

(RQ.C.1) What is the objective SI and SQ performance of optimal joint far- and near-end minimum processing based on an ASII SI constraint and a MSE cost function?

Paper D

(OB.D.1) Existing optimal joint far- and near-end minimum processing based on an ASII SI constraint and a MSE cost function only solved numerically.

(OB.D.2) Lacking existing performance comparison between: (1) joint and blind maximum processing methods, (2) joint and blind minimum processing methods, (3) blind minimum and maximum processing methods, (4) joint minimum and maximum processing methods.

(H.D.1) Joint processing has higher objective and subjective SI and SQ than blind processing for both the maximum and minimum processing case.

(H.D.2) Minimum processing has a higher objective and subjective SQ in higher SNRs than maximum processing for both the joint and blind processing case.

Table 4 provides a concentrated and clear overview of the contributions and answers to each of these hypotheses, research questions and observations, and is summarized below.

Fig. 13 illustrates the papers' relationship between each other and the previous covered topics of FSE, NLE, joint far- and near-end enhancement with both maximum and minimum processing. Similarly, Table 5 categories the presented papers' processing methodologies according to the joint enhancement scenarios and minimum and maximum processing goals similar to the existing works in Table 1 and 2 in Sections 5.3.2 and 4.4, respectively.

Paper	Observation (OB) / Hypothesis (H) / Research Question (RQ)	Contribution
[A]	OB: Mutual information based joint enhancement [6] requires a complex signal model with production noise, several approximations and is closely related to ASII.	Showed choices made in [6] regarding producing noise and approximation steps result in optimization target equivalent to ASII. Requires less complex signal model to optimize for ASII.
	RQ: What is the objective SI performance of optimal joint far- and near-end speech and listening based on ASII maximization?	Derived closed-form optimal joint far- and near-end enhancement based on ASII maximization. Results show ESTOI on par or better than [6].
[B]	OB: Minimum processing concept not formulated in NLE.	Extended minimum processing problem formulation to NLE.
	H: Minimum processing NLE has higher SQ than maximum processing NLE with similar SI.	Confirmed hypothesis with objective measures for SI and SQ and subjective listening test for SQ.
	RQ: What is the SI and SQ performance of optimal minimum processing NLE based on an ASII SI constraint and a MSE cost function?	Derived closed-form solution to minimum processing NLE optimization problem with ASII SI constraint and MSE cost function. Investigated performance with objective SI and SQ metrics and subjective SQ listening test.
[C]	OB: Minimum processing concept not formulated for joint enhancement.	Combine concepts of minimum processing and joint enhancement to formulate joint far- and near-end minimum processing problem.
	RQ: What is the objective SI and SQ performance of optimal joint far- and near-end minimum processing based on an ASII SI constraint and a MSE cost function?	Formulated joint far- and near-end optimization problem with MSE cost function, ASII SI constraint and an additional SQ constraint based on noise power. Experimentally investigated performance with numerical solution to optimization problem.
	H: Joint minimum processing has a higher objective SI than blind minimum processing.	Experimental study with ESTOI confirm that joint minimum processing achieves better SI than blind minimum processing.
	H: Joint minimum processing has a higher objective SQ than blind minimum processing.	Experimental study with PESQ show joint and blind minimum processing has similar performance. Blind concatenation preserves minium processing ability of individual methods.
[D]	OB: Existing optimal joint far- and near-end minimum processing based on an ASII SI constraint and a MSE cost function only solved numerically.	Derived closed-form solution to optimization problem proposed in Paper C. Extensive analysis of the solution and its behavior for varying SNR conditions.
	OB: Lacking existing performance comparison between: (1) joint and blind maximum processing methods, (2) joint and blind minimum processing methods, (3) blind minimum and maximum processing methods, (4) joint minimum and maximum processing methods.	Comprehensive investigation of both objective and subjective performance of blind and joint end-to-end enhancement methods in both the minimum and maximum processing case.
	H: Joint processing has higher objective and subjective SI and SQ than blind processing for both the maximum and minimum processing case.	Joint only better in objective SI in the minimum processing case, but not for maximum processing. No significant differences in subjective listening test. No differences in SQ.
	H: Minimum processing has a higher objective and subjective SQ in higher SNRs than maximum processing for both the joint and blind processing case.	Minimum processing achieves objective SI on par with maximum processing and higher objective SQ in higher SNRs. No significant differences in subjective listening test.

Table 4: Observations, hypotheses and research questions addressed in the papers in Part II and the respective contributions.

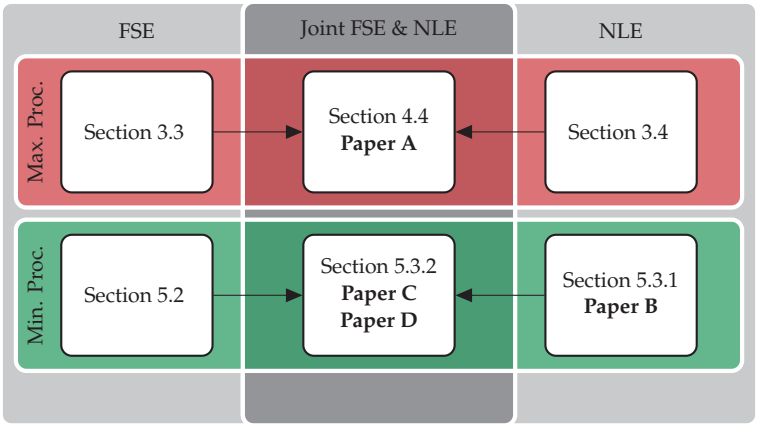


Fig. 13: Relationship between papers and enhancement topics covered in the introductory sections of the thesis.

Paper	Prob. Form.	Final Algorithm	Methodology	Channels	Proc. Goal	Target Metrics
[A]	10d: Fully joint	10b: FE unaware of NE & NE aware of FE	Classic	Multi	Max. SI	ASII
[B]	10a: Blind NLE	10a: Blind NLE	Classic	Single	Min. proc. with sufficient SI	SI: ASII Distortion: MSE
[C]	10d: Fully joint	10d: Fully joint	Classic & heuristic	Multi	Min. proc. with sufficient SI	SI: ASII Distortion: MSE
[D]	10d: Fully joint	10d: Fully joint	Classic & heuristic	Multi	Min. proc. with sufficient SI	SI: ASII Distortion: MSE

Table 5: Categorization of the papers in Part II in terms of the problem formulations, final algorithms according to the side-information awareness scenarios presented in Fig. 10, methodology, the processing goal and target metrics.

The investigation starts in the area of maximum processing joint far- and near-end enhancement, where little work existed prior to the start of this thesis. Noticing that existing state-of-the-art maximization of mutual information required several approximations and a complex signal model, we investigate the use of ASII as a suitable metric for multi-channel joint far- and near-end intelligibility enhancement on par with the state-of-the-art (Paper A). Observing that there were no direct formulation or usage of the minimum processing principle in NLE to improve SQ in high SNR conditions, we formulate and investigate the usage of minimum processing in NLE based on ASII (Paper B). Then, with the intent of leveraging the power of joint enhancement for improving SI and minimum processing for SQ, we combine the two main aspects of this thesis: joint far- and near-end enhancement, and minimum processing with ASII as SI metric (Paper C and D). Initially, we propose the joint far- and near-end minimum processing framework and conduct a proof of concept investigation with objective metrics and a numerical solution to the optimization problem (Paper C). We then derive a closed form solution and conduct a thorough and systematic investigation of the previously proposed joint far- and near-end minimum processing framework, and we focus on both the effects of minimum versus maximum processing and joint versus blind processing and their cross-combination (Paper D).

In the following section, we give a short and more detailed summary of each paper.

7.2 Contributions

[A] Joint Far- and Near-End Speech Intelligibility Enhancement based on the Approximated Speech Intelligibility Index

In this paper, we consider multi-channel joint far- and near-end SI maximization. Specifically, we extended the optimization of SI based on the ASII metric to the joint case, and derive a closed-form solution to the joint far- and near-end speech enhancement problem. Previous work jointly optimized far- and near-end SI based on mutual information [6]. However, this required a more elaborate signal model that includes natural speech variations, and the need for significant approximations and assumptions on the underlying signal distributions. We analyze the model choices and assumptions of the existing work, specifically how the relation between approximated mutual information and ASII motivate our model choices. The proposed solution and usage of ASII optimization requires a simpler well known signal model and is independent of the marginal signal distributions. Particular, we do not need to introduce or optimize for additional free parameters to model production and interpretation noise. We conduct an experimental comparison between the existing mutual information based optimization, with a newer production noise

model [139–141], and our proposed ASII optimization. The results show, that the proposed method, with a simpler signal model, achieves similar or slightly better ESTOI performance than the existing work.

[B] Minimum Processing Near-end Listening Enhancement

In this paper, we consider the area of NLE in combination with minimum processing. In particular, we take inspiration from the recently proposed concept of minimum processing beamforming [16], and propose a new *minimum processing NLE* rationale. In this context, the target speech undergoes minimal processing to minimize a processing penalty, as long as it meets a specific performance criterion, such as SI. The objective then shifts from solely enhancing SI to processing the target speech just sufficiently to maintain minimum SI under specific noise conditions, and simultaneously reducing speech distortions when noise levels are favorable. This is in contrast to existing NLE work, that focuses primarily on maximum processing which greatly increases SI in harsh noise environments, but this comes at the cost of degradations in SQ in favorable noise conditions. Thus, we prose a more general formulation of the NLE problem that focuses on controlling both SI and SQ in an adaptive manner depending on the noise conditions.

We present a case study where the processing penalty is the MSE between the NLE output signal and the clean speech, and the performance criterion is based on the ASII. We derive a closed-form solution to this optimization problem, which provides a computationally efficient gain rule that adapts to changing noise conditions, limiting distortions to the minimum necessary to achieve a desired SI. Experimental studies, that the proposed method achieves SQ comparable to or surpassing that of existing methods in both objective metrics and subjective listening evaluations, while also maintaining SI at levels equivalent to those of existing methods.

[C] Joint Minimum Processing Beamforming and Near-End Listening Enhancement

In this paper, we return to multi-channel joint far- and near-end speech and listening enhancement. Particularly, inspired by minimum processing beamforming [16] and minimum processing NLE [15] (Paper B), we propose a joint far- and near-end minimum processing framework to utilize both the advantageous of joint enhancement and minimum processing. Specifically, contrary to most existing joint works, the proposed framework is designed to enhance SI and preserve SQ by minimally altering the signal only as much as necessary to achieve a desired level of SI, while also limiting speech distortions in favorable noise conditions. Additionally, we expand upon the existing minimum processing FSE and NLE frameworks [15, 16] by jointly taken into account the

impacts of both FSE and NLE and the far- and near-end noise simultaneously.

We present an exemplary optimization problem, where the processing penalty is the MSE between the NLE output signal and the output of a reference beamformer subject to both an SI performance constraint based on ASII and an SQ performance constraint based on noise powers. We provide analytical solutions for specific boundary cases of interest and explore the overall performance through numerical optimization for the general case. Experimental comparison with blind concatenation of existing minimum processing FSE [16] and minimum processing NLE [15], show that the proposed joint approach improves performance compared to the concatenated approach in objective SI while preserving objective SQ. Furthermore, we see that concatenating existing minimum processing FSE and NLE methods preserves the abilities of the individual methods.

[D] Joint Far- and Near-end Speech and Listening Enhancement with Minimum Processing

In this paper, we present a comprehensive exploration and systematic analysis of the joint far-end and near-end minimum processing framework, which was initially introduced in [129] (Paper C). Our core contribution is the derivation of a closed-form analytical solution to the joint far- and near-end minimum processing optimization problem with MSE processing penalty, SI constraint based on the ASII, and a noise power constraint. We analyze the behavior of the optimal solution in great detail. Furthermore, we perform a systematic experimental evaluation using objective measures and listening tests for SI, listening effort, and SQ.

To evaluate the impact of both minimum processing in contrast to SI maximization, as well as joint processing compared to blind processing, we conduct comparisons with various existing methods: joint ASII maximization [142] (Paper A); blind concatenation of minimum processing FSE [16] and minimum processing NLE [15] (Paper B); and blind concatenation classic maximum processing, i.e., an MVDR beamformer [37] and NLE ASII maximization [46]. The results indicate that minimum processing maintains SI on par with maximum processing and preserves SQ in environments with higher SNRs, underscoring its effectiveness in end-to-end communication. While joint processing enhances objective SI in cases of minimum processing, it does not provide the same benefit when maximum processing is applied. However, the subjective listening tests did not show significant differences between the various methods, suggesting that optimizing far- and near-end aspects separately could be a viable and more straightforward alternative to joint optimization in certain speech and listening situations. The paper underscores the complexity of the relationship between SI, SQ and processing methods. The study provides detailed insights into the joint minimum processing optimization problem, and

emphasizes the need for future research to draw more definitive conclusions regarding joint far- and near-end optimization.

8 Future Research

This thesis has covered and introduced the concepts and combination of joint far- and near-end enhancement, and minimum processing. However, these are very elaborate topics and several aspects have not been covered in this work. In addition, our results and observations also warrant further investigations in multiple directions.

8.1 Extended Review of Joint Enhancement Techniques

The topic of joint far- and near-end enhancement is relatively new and has not received extensive prior study. Our research provides insights into this field and investigates the performance of joint and blind processing in relation to both maximum and minimum processing. However, our results also showed a lack of significant performance differences between the proposed joint and blind maximum processing methods contrary to some of the other existing works. The results also show that existing blind methods perform very well on their own. Therefore, our results also underscore the need for more comprehensive investigations with controlled implementations of joint and blind methods under identical conditions to gain a deeper understanding of the behavioral aspects of blind and joint processing. Consequently, a thorough review of existing joint enhancement works and a performance comparison under similar conditions and implementations is necessary.

8.2 More Advanced Enhancement Targets and Methods

So far the minimum processing frame work in both FSE [16], NLE (Paper B) and joint enhancement (Paper C and D), has been formulated as an optimization problem per subband instead of a broadband solution across the entire frequency domain. This is partly done for convenience, but is also necessary because the ASII and SII optimization targets rely on overlapping subbands leading to severely complicated optimization problems. However, the additional control of processing from a broadband solution might lead to an increase in performance. Especially, if considering other optimization targets.

The presented enhancement techniques in Papers A – D rely on optimization in relation to the ASII. However, the ASII is rather simple metric and does not correlate nearly as well with subjective SI as other more advanced measures such as mutual information, ESTOI, extended-SII etc. [43]. Furthermore, the DNN based joint enhancement techniques based on, e.g., ESTOI [131] have

shown great potential. Therefore, it is interesting to extend the presented joint minimum processing framework to more advanced SI and SQ predictors, with the possibility of providing even further enhancements that also lead to increases in subjective performance.

The proposed joint minimum processing method solution is derived for each individual subband. However, depending on the noise conditions, a feasible solution does not exists. Therefore, we propose different approaches to handle this in Papers C and D. However, it is important to gain a better understanding into this behavior, and future work involves investigating the feasibility conditions within the optimization problem. Additionally, we aim to derive optimal performance strategies even in cases where feasibility is challenging.

8.3 Beyond Intelligibility and Quality Enhancement

Although it is important to investigate potentials for increased SI and SQ with more advanced metrics, the results of Paper D and [6] indicate that joint or informed far- and near-end enhancement does not necessarily lead to significant performance gains compared to blind enhancement. Particularly, it can be seen that blind enhancement is able to achieve a high SI and SQ performance. Hence, because existing methods perform very well, there is less room for improvement in terms of SI and SQ. Furthermore, as mentioned in Section 4.3 joint processing has diminishing returns as environmental noise subsidies or the noise reduction abilities of FSE increases. Therefore, we suggests that investigations into performance of joint far- and near-end processing should extend beyond the traditional SI and SQ domains. Particularly, we are interested in the possible extensions and relations to speech coding and quantization, and evaluations of the increased processing complexity that might follow with joint enhancement methods.

8.4 Including ANC

In this thesis, the presented enhancement methods apply NLE prior to play back in the noisy near-end environment to improve SI and SQ. Hence, the near-end noise is countered only through alterations of the play-back signal. However, ANC methods are able to greatly increase SI and SQ by actively cancelling near-end noise via an anti-noise signal [105, 125–127]. Although ANC puts additional requirements on the placement of loudspeakers in relation to listener or necessitates the usage of headphones, ANC promises a great potential for advancing the proposed schemes. Particularly, in relation to minimum processing since ANC produces a less noisy near-end environment around the listener. Hence, the need for minimum processing and minimal speech distortions at high SI levels is even more important. Furthermore, from the joint

perspective the additional interaction between both FSE, NLE and ANC may result in additional conflicts in blind processing and require more joint control and optimization to ensure maximum benefit from all processing steps.

8.5 More Realistic Conditions

Our work investigated performance in ideal conditions, where we have assumed perfect knowledge of all relevant signal attributes (noise and speech powers). This was done to establish a baseline optimal performance. However, it is important to also investigate performance and behavior of joint and blind minimum processing in more realistic conditions, where speech and noise statistics are estimated from the recorded signals. In addition, an immediate extensions is including reverberation in the far-end and performing time-varying processing for a more realistic performance.

Furthermore, the presented processing techniques are linear. However, the non-linear nature of DNNs have shown good results in joint enhancement [131, 133], similarly possible non-linear processing steps may arise in practice particularly in regards to speech coding and quantization. Therefore, including more non-linear processing for possible performance gains is of particular interest, especially in relation to the inclusion of non-linear processing in the form of speech coding in the signal model and optimization.

8.6 Speech Coding and Information Transfer

In the context of joint or informed processing in practice, a crucial aspect is the necessary bidirectional transfer of side-information and synchronization between the far- and near-end or a central processor. However, our current assumptions – namely, no synchronization issues and no information transfer problems – does not hold true in general. Therefore, future investigations should explore the impact of information loss or mis-synchronization on the feasibility of joint processing. Additionally, identifying mitigation steps and determining the necessary amount of information for successful joint processing is essential. Notably, in cases of informed processing, we may find that less information suffices for achieving the desired outcomes as the enhancement steps are not updated fully jointly.

In speech communication between physically separated environments, the process of encoding and decoding speech plays a central role. However, this encoding introduces inherent quantization noise, commonly referred to as coding noise. Additionally, for joint enhancement to be effective, the side-information that must be transferred – along with the speech – also needs to undergo encoding. In this thesis, we have assumed the absence of coding noise. However, future investigations should delve into several critical aspects related to speech coding. These include: robustness to coding, exploring

how joint processing remains robust in the presence of coding artifacts. Bit rate considerations, investigating the potential increase in bit rates required to carry both enhanced signals and side-information and balancing the trade-off between data rate and enhancement performance is crucial. Mitigation strategies, identifying effective steps to minimize the impact of coding noise on joint enhancement or steps to limit the necessary bit rates. Particular, joint optimization of FSE, NLE and coding, where we optimize the coding process alongside the enhancement in end-to-end communication. Thus, improving overall performance in terms of both bit rate, SI and SQ, by incorporating even more processing steps and noises into the optimization. Especially, utilizing joint knowledge of the noise characteristics of the near-end environment to improve the rate-distortion tradeoff. Additionally, understanding how coding noise interacts with the enhancement process can inform better design choices.

Finally, in the sense of joint enhancement and information transfer, it is interesting to expand the optimization to new scenarios. Particularly, where enhancement must be done for multiple far-end environments with only one near-end, or where a single far-end broadcasts to multiple near-end environments as in online meetings. This poses new interesting scenarios, where processing must be balanced across multiple channels and information transferred in other directions, where perhaps the far-end is aware of the near-ends but not wise versa.

References

[1] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[2] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, Jul. 2015.

[3] É. Lombard, "Le signe de l'élévation de la voix," *Annales des Maladies de l'Oreille et du Larynx*, vol. 37, no. 2, pp. 101–119, 1911.

[4] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, Sep. 1988.

[5] P. Assmann and Q. Summerfield, "The Perception of Speech Under Adverse Conditions," in *Speech Processing in the Auditory System*. New York: Springer-Verlag, 2004, vol. 18, pp. 231–308.

[6] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Intelligibility Enhancement Based on Mutual Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1694–1708, Aug. 2017.

[7] T.-C. Zorilă, Y. Stylianou, S. Flanagan, and B. C. J. Moore, "Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss," *The Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 189–196, Jan. 2017, publisher: Acoustical Society of America.

[8] K. Tan and D. Wang, "Improving Robustness of Deep Learning Based Monaural Speech Enhancement Against Processing Artifacts," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 6914–6918.

[9] M. Niermann, P. Jax, and P. Vary, "Joint Near-End Listening Enhancement and far-end noise reduction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 4970–4974.

[10] M. Niermann, P. Vary, and P. Jax, "Near-End Listening Enhancement: The Impact of Far-End Noise Reduction," in *Proc. German Annu. Conf. Acoust.*, 2019, pp. 1371–1374.

[11] J. Rennies, A. Pusch, H. Schepker, and S. Doclo, "Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements," *The Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. EL315–EL321, Oct. 2018.

[12] R. Pricken, M. Wältermann, E. Parotat, M. Soloducha, and A. Raake, "Quality Aspects of Near-End Listening Enhancement Approaches in Telecommunication Applications," in *Proceedings of DAGA 2017*. Kiel: German Acoustical Society (DEGA), 2017, pp. 872–875.

[13] Y. Tang, C. Arnold, and T. J. Cox, "A Study on the Relationship between the Intelligibility and Quality of Algorithmically-Modified Speech for Normal Hearing Listeners," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, vol. 1, no. 1, p. 10, Jun. 2018.

[14] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech & Language*, vol. 28, no. 4, pp. 858–872, Jul. 2014.

[15] A. J. Fuglsig, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof, and J. Østergaard, "Minimum Processing Near-End Listening Enhance-

ment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2233–2245, 2023.

[16] A. Zahedi, M. S. Pedersen, J. Østergaard, T. U. Christiansen, L. Bramsløw, and J. Jensen, "Minimum Processing Beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2710–2724, 2021.

[17] P. B. Denes and E. N. Pinson, *The speech chain: the physics and biology of spoken language*, 2nd ed. New York, NY: Freeman, 1993.

[18] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[19] D. Michelsanti, "Audio-Visual Speech Enhancement Based on Deep Learning," PhD thesis, Aalborg University, 2020.

[20] A. Radford, M. Atkinson, D. Britain, H. Clahsen, and A. Spencer, *Linguistics: An Introduction*, 2nd ed., A. Radford, Ed. Cambridge ; New York: Cambridge University Press, 2009.

[21] J. C. L. Ingram and H. Chenery, *Neurolinguistics: An Introduction to Spoken Language Processing and its Disorders*. Cambridge, UNITED KINGDOM: Cambridge University Press, 2007.

[22] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.

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

[24] K. N. Stevens, *Acoustic Phonetics*, ser. Current Studies in Linguistics. Cambridge, Mass.: MIT Press, 2000, no. 30.

[25] C. J. Plack, *The Sense of Hearing*, 2nd ed. New York: Psychology Press, Taylor & Francis Group, 2014.

[26] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, Mass: MIT Press, 1997.

[27] F. Spoor, T. Garland, G. Krovitz, T. M. Ryan, M. T. Silcox, and A. Walker, "The primate semicircular canal system and locomotion," *Proceedings of the National Academy of Sciences*, vol. 104, no. 26, pp. 10 808–10 812, Jun. 2007.

[28] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Bingley: Emerald, 2012.

[29] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic Beamforming for Hearing Aid Applications," in *Handbook on Array Processing and Sensor Networks*. John Wiley & Sons, Ltd, 2010, pp. 269–302.

[30] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[31] R. Zamir, *Lattice Coding for Signals and Networks.* Cambridge University Press, 2014, oCLC: 889704920.

[32] R. Veldhuis and M. Breeuwer, *An introduction to source coding*. New York: Prentice Hall, 1993.

[33] T. Berger, *Rate distortion theory: a mathematical basis for data compression*, ser. Prentice-Hall series in information and system sciences. Englewood Cliffs, N.J: Prentice-Hall, 1971.

[34] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*, 3rd ed., ser. Always learning. Harlow: Pearson, 2014, oCLC: 935118828.

[35] M. A. Niermann, "Digital enhancement of speech perception in noisy environments," PhD thesis, RWTH Aachen University, Aachen, 2019.

[36] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge University Press, 2004.

[37] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[38] P. Hoang, "User-Symbiotic Speech Enhancement for Hearing Aids," Ph.d, Aalborg University, 2022.

[39] American National Standards Institute, *Methods for Calculation of the Speech Intelligibility Index*. New York, N.Y: Acoustical Society of America, 2017, vol. ANSI S.35-1997.

[40] S. Bech and N. Zacharov, *Perceptual Audio Evaluation-Theory, Method and Application: Bech/Perceptual Audio Evaluation-Theory, Method and Application*. Chichester, UK: John Wiley & Sons, Ltd, Apr. 2006.

[41] A. H. Andersen, "Speech Intelligibility Prediction for Hearing Aid Systems," PhD thesis, Aalborg University, 2017.

[42] M. Pedersen, "Data-Driven Speech Intelligibility Prediction," Ph.D. dissertation, Aalborg University, 2023, publisher: Aalborg Universitetsforlag.

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

[44] R. Xia, J. Li, M. Akagi, and Y. Yan, "Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, Mar. 2012, pp. 4465–4468.

[45] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[46] C. H. Taal, J. Jensen, and A. Leijon, "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[47] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

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

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

[50] J. H. L. Hansen and B. L. Pellom, "An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms," in *The International Conference on Speech and Language Processing*, 1998, pp. 2819–2822.

[51] ITU-T, "Recommendation P.862: Perceptual evaluation of speech quality (PESQ)," International Telecommunication Unition, Recommendation ITU-T P.862, Feb. 2001.

[52] M. Brandstein and D. Ward, Eds., *Microphone arrays: signal processing techniques and applications*, ser. Digital signal processing. New York: Springer, 2001.

[53] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement: with 18 tables*, ser. Signals and Communication Technology. Berlin: Springer, 2005.

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

[55] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, 1st ed., ser. Synthesis Lectures on Speech and Audio Processing. Springer Cham, 2013.

[56] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[57] Yiteng Huang, J. Benesty, and Jingdong Chen, "Analysis and Comparison of Multichannel Noise Reduction Methods in a Common Framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 957–968, Jul. 2008.

[58] K. Eneman, H. Luts, J. Wouters, M. Büchler, N. Dillier, W. Dreschler, M. Froehlich, G. Grimm, V. Hohmann, R. Houben, A. Leijon, A. Lombard, D. Mauler, M. Moonen, H. Puder, M. Schulte, A. Spriet, and M. Vormann, "Evaluation of signal enhancement algorithms for hearing instruments," in *2008 16th European Signal Processing Conference*, Aug. 2008, pp. 1–5.

[59] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New Insights Into the MVDR Beamformer in Room Acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[60] I. López-Espejo, A. Edraki, W.-Y. Chan, Z.-H. Tan, and J. Jensen, "On the deficiency of intelligibility metrics as proxies for subjective intelligibility," *Speech Communication*, vol. 150, pp. 9–22, May 2023.

[61] K. Tesch and T. Gerkmann, "Insights Into Deep Non-Linear Filters for Improved Multi-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023.

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

[63] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, Aug. 2007.

[64] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969, conference Name: Proceedings of the IEEE.

[65] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel Wiener filter for multiple sources scenarios," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, Nov. 2012, pp. 1–5.

[66] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech Distortion Weighted Multichannel Wiener Filtering Techniques for Noise Reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 199–228.

[67] J. C. R. Licklider and I. Pollack, "Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech," *The Journal of the Acoustical Society of America*, vol. 20, no. 1, pp. 42–51, Jan. 1948.

[68] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 277–282, Aug. 1976.

[69] J. D. Griffiths, "Optimum Linear Filter for Speech Transmission," *The Journal of the Acoustical Society of America*, vol. 43, no. 1, pp. 81–86, Jan. 1968.

[70] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Noise-dependent post-filtering," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2004, pp. I–457.

[71] S. Yoo, J. Boston, J. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and Ching-Chung Li, "Speech enhancement based on transient speech information," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, Oct. 2005, pp. 62–65.

[72] B. Sauert, G. Enzner, and P. Vary, "Near End Listening Enhancement With Strict Loudspeaker Output Power Constraining," in *Proc. of Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, Sep. 2006, p. 4.

[73] M. D. Skowronski and J. G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, vol. 48, no. 5, pp. 549–558, May 2006.

[74] B. Sauert, G. Enzner, and P. Vary, "Near End Listening Enhancement With Strict Loudspeaker Output Power Constraining," in *Proc. of Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, Sep. 2006, p. 4.

[75] S. D. Yoo, J. R. Boston, A. El-Jaroudi, C.-C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1138–1149, Aug. 2007.

[76] P. S. Chanda and S. Park, "Speech Intelligibility Enhancement using Tunable Equalization Filter," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, Apr. 2007, pp. IV–613–IV–616, iSSN: 2379-190X.

[77] C. Tantibundhit, J. R. Boston, C. C. Li, J. D. Durrant, S. Shaiman, K. Kovacyk, and A. El-Jaroudi, "New signal decomposition method based speech enhancement," *Signal Processing*, vol. 87, no. 11, pp. 2607–2628, Nov. 2007.

[78] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to Speech Intelligibility Index," in *2009 17th European Signal Processing Conference*, Aug. 2009, pp. 1844–1848.

[79] ——, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *2010 18th European Signal Processing Conference*, Aalborg, Denmark, Aug. 2010, pp. 1919–1923.

[80] ——, "Recursive Closed-Form Optimization of Spectral Audio Power Allocation for Near End Listening Enhancement," *ITG-Fachtagung Sprachkommunikation*, vol. Paper 8, p. 4, Oct. 2010.

[81] J. L. Hall and J. L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 280–285, Jan. 2010.

[82] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise Intelligibility Improvement Based on Power Recovery and Dynamic Range Compression," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Portland, USA, Aug. 2012, pp. 2075–2079.

[83] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4061–4064, iSSN: 2379-190X.

[84] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. ISCA Interspeech*, Portland, OR, 2012, pp. 955–958.

[85] P. N. Petkov, G. E. Henter, and W. B. Kleijn, "Maximizing Phoneme Recognition Accuracy for Enhanced Speech Intelligibility in Noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1035–1045, May 2013.

[86] M. Zhang, P. N. Petkov, and W. B. Kleijn, "Rephrasing-Based Speech Intelligibility Enhancement," in *Interspeech*, Lyon, France, 2013, p. 5.

[87] V. Aubanel and M. Cooke, "Information-Preserving Temporal Reallocation of Speech in the Presence of Fluctuating Maskers," in *Proc. of ISCA Interspeech*, Lyon, France, 2013, pp. 3592–3596.

[88] J. B. Crespo and R. C. Hendriks, "Speech reinforcement in noisy reverberant environments using a perceptual distortion measure," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, May 2014, pp. 910–914.

[89] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles," *Computer Speech & Language*, vol. 28, no. 2, pp. 629–647, Mar. 2014.

[90] J. B. Crespo and R. C. Hendriks, "Multizone Speech Reinforcement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 54–66, Jan. 2014.

[91] C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, "Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion," *Computer Speech & Language*, vol. 28, no. 2, pp. 665–686, Mar. 2014.

[92] H. Schepker, J. Rennies, and S. Doclo, "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 2692–2706, Nov. 2015.

[93] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal Near-End Speech Intelligibility Improvement Incorporating Additive Noise and Late Reverberation Under an Approximation of the Short-Time SII," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 851–862, May 2015.

[94] W. B. Kleijn and R. C. Hendriks, "A Simple Model of Speech Communication and its Application to Intelligibility Enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 303–307, Mar. 2015.

[95] Y. Tang and M. Cooke, "Learning static spectral weightings for speech intelligibility enhancement in noise," *Computer Speech & Language*, vol. 49, pp. 1–16, May 2018.

[96] S. Jebaruby, N. Singh, and M. Jeeva, "Weighted Energy Reallocation Approach for Near-end Speech Enhancement," in *2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, vol. 1, Mar. 2019, pp. 516–522.

[97] H.-Y. Dong and C.-M. Lee, "Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, p. 13, May 2018.

[98] A. Fallah and S. van de Par, "A Speech Preprocessing Method Based on Perceptually Optimized Envelope Processing to Increase Intelligibility in Reverberant Environments," *Applied Sciences*, vol. 11, no. 22, p. 10788, Jan. 2021.

[99] M. Niermann, P. Jax, and P. Vary, "Near-end listening enhancement by noise-inverse speech shaping," in *2016 24th European Signal Processing Conference (EUSIPCO)*, Aug. 2016, pp. 2390–2394.

[100] M. Niermann and P. Vary, "Listening Enhancement in Noisy Environments: Solutions in Time and Frequency Domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 699–709, 2021.

[101] E. Jokinen, U. Remes, P. Alku, E. Jokinen, U. Remes, and P. Alku, "Intelligibility Enhancement of Telephone Speech Using Gaussian Process Regression for Normal-to-Lombard Spectral Tilt Conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, pp. 1985–1996, Oct. 2017.

[102] G. Li, H. Ruimin, R. Zhang, and X. Wang, "A mapping model of spectral tilt in normal-to-Lombard speech conversion for intelligibility en-

hancement," *Multimedia Tools and Applications*, vol. 79, no. 27-28, pp. 19 471–19 491, Jul. 2020.

[103] R. Zhang, R. Hu, G. Li, and X. Wang, "Spectral Tilt Estimation for Speech Intelligibility Enhancement Using RNN Based on All-Pole Model," in *MultiMedia Modeling*, ser. Lecture Notes in Computer Science, I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, and S. Vrochidis, Eds. Springer International Publishing, 2018, pp. 144–156.

[104] H. Li and J. Yamagishi, "Multi-Metric Optimization Using Generative Adversarial Networks for Near-End Speech Intelligibility Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3000–3011, 2021.

[105] G. Li, R. Hu, X. Wang, and R. Zhang, "A near-end listening enhancement system by RNN-based noise cancellation and speech modification," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 483–15 505, Jun. 2019.

[106] H. Li, S.-W. Fu, Y. Tsao, and J. Yamagishi, "iMetricGAN: Intelligibility Enhancement for Speech-in-Noise Using Generative Adversarial Network-Based Metric Learning," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1336–1340.

[107] D. Li, C. Zhu, and L. Zhao, "D2StarGAN: A Near-Far End Noise Adaptive StarGAN for Speech Intelligibility Enhancement," *Electronics*, vol. 12, no. 17, Aug. 2023.

[108] K. Nathwani, F. Hafiz, A. Swain, and R. Biswas, "Speech Intelligibility Enhancement using an Optimal Formant Shifting Approach," in *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Sep. 2021, pp. 120–125.

[109] T. Ngo, R. Kubo, and M. Akagi, "Increasing speech intelligibility and naturalness in noise based on concepts of modulation spectrum and modulation transfer function," *Speech Communication*, vol. 135, pp. 11–24, Dec. 2021.

[110] J. Pak, I. Choi, Y. G. Jin, and J. W. Shin, "Multichannel speech reinforcement based on binaural unmasking," *Signal Processing*, vol. 139, pp. 165–172, Oct. 2017.

[111] C. Chermaz and S. King, "A Sound Engineering Approach to Near End Listening Enhancement," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1356–1360.

[112] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications

in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, May 2013.

[113] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[114] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. N. Petkov, B. Sauert, and P. Vary, "Optimizing Speech Intelligibility in a Noisy Environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, Mar. 2015.

[115] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Interspeech*, Lyon, France, 2013, pp. 3552–3556.

[116] J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-Enhancing Speech Modifications — The Hurricane Challenge 2.0," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1341–1345.

[117] C. Chermaz, C. Valentini-Botinhao, H. Schepker, and S. King, "Near End Listening Enhancement in Realistic Environments," *Proceedings of the ICA 2019 and EAA Euroregio : 23rd International Congress on Acoustics*, vol. integrating 4th EAA Euroregio 2019 : 9-13 September 2019, pp. 5731–5735, 2019.

[118] S. Gordon-Salant, "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing," *The Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1599–1607, Dec. 1986.

[119] K. D. Kryter, "Methods for the Calculation and Use of the Articulation Index," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1689–1697, Nov. 1962.

[120] B. Sauert, *Near-End Listening Enhancement: Theory and Application*, 1st ed., ser. Aachener Beiträge zu digitalen Nachrichtensystemen. Aachen: Wissenschaftsverlag Mainz, 2014, no. 36, oCLC: 880393716.

[121] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, Feb. 2006.

[122] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A Low-Complexity Spectro-Temporal Distortion Measure for Audio Processing Applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1553–1564, Jul. 2012.

[123] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Hoboken, N.J: Wiley-Interscience, 2006.

[124] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Jointly optimal near-end and far-end multi-microphone speech intelligibility enhancement based on mutual information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, Mar. 2016, pp. 654–658.

[125] S. Kuo and D. Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87, no. 6, pp. 943–973, Jun. 1999.

[126] N. V. George and G. Panda, "Advances in active noise control: A survey, with emphasis on recent nonlinear techniques," *Signal Processing*, vol. 93, no. 2, pp. 363–377, Feb. 2013.

[127] F. Cheng, X. Wang, L. Gang, W. Tu, and J. Wang, "Speech Intelligibility Enhancement in Strong Mechanical Noise Based on Neural Networks," in *Advances in Multimedia Information Processing – PCM 2017*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 702–712.

[128] M. Kolbæk, Z.-H. Tan, and J. Jensen, "On the Relationship Between Short-Time Objective Intelligibility and Short-Time Spectral-Amplitude Mean-Square Error for Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 283–295, Feb. 2019.

[129] A. J. Fuglsig, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof, and J. Østergaard, "Joint Minimum Processing Beamforming and Near-end Listening Enhancement," in *IEEE 2024 Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. Seoul: IEEE, May 2024, p. 5.

[130] A. J. Fuglsig, Z.-H. Tan, L. S. Bertelsen, J. Jensen, J. C. Lindof, and J. Østergaard, "Joint Far- and Near-end Speech and Listening Enhancement with Minimum Processing," 2024, pre-print submitted to IEEE ACCESS.

[131] H. Li, Y. Liu, and J. Yamagishi, "Joint Noise Reduction and Listening Enhancement for Full-End Speech Enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2023, pp. 1–5.

[132] T.-C. Zorilă and Y. Stylianou, "On the Quality and Intelligibility of Noisy Speech Processed for Near-End Listening Enhancement," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 2023–2027.

[133] M. P. Shifas, C. Zorilă, and Y. Stylianou, "End-to-End Neural Based Modification of Noisy Speech for Speech-in-Noise Intelligibility Improvement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 162–173, 2022.

[134] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985, conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

[135] B. Hect, J. Teevan, and Sellen, "The "Leaf Blower Problem" and the importance of common ground," Sep. 2021, Microsoft Research.

[136] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Speech quality assessment with WARP-Q: From similarity to subsequence dynamic time warp cost," *IET Signal Processing*, vol. 16, no. 9, pp. 1050–1070, 2022.

[137] ITU-R, "Recommendation BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Unition, Recommendation ITU-R BS.1534-3, Oct. 2015.

[138] ITU-T, "Recommendation P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," International Telecommunication Unition, Recommendation ITU-T P.835, Nov. 2003, issue: ITU-T P.862.2.

[139] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, Jan. 2018.

[140] ——, "An intelligibility metric based on a simple model of speech communication," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.

[141] ——, "On the information rate of speech communication," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 5625–5629.

[142] A. J. Fuglsig, J. Østergaard, J. Jensen, L. S. Bertelsen, P. Mariager, and Z.-H. Tan, "Joint Far- and Near-End Speech Intelligibility Enhancement Based on the Approximated Speech Intelligibility Index," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7752–7756.

Part II

Papers

Paper A

Joint Far- and Near-End Speech Intelligibility Enhancement based on the Approximated Speech Intelligibility Index

Andreas Jonas Fuglsig, Jan Østergaard, Jesper Jensen, Lars Søndergaard Bertelsen, Peter Mariager and Zheng-Hua Tan

The paper has been published in
2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7752–7756, 2022.

The layout has been revised.

Abstract

This paper considers speech enhancement of signals picked up in one noisy environment which must be presented to a listener in another noisy environment. Recently, it has been shown that an optimal solution to this problem requires the consideration of the noise sources in both environments jointly. However, the existing optimal mutual information based method requires a complicated system model that includes natural speech variations, and relies on approximations and assumptions of the underlying signal distributions. In this paper, we propose to use a simpler signal model and optimize speech intelligibility based on the Approximated Speech Intelligibility Index (ASII). We derive a closed-form solution to the joint far- and near-end speech enhancement problem that is independent of the marginal distribution of signal coefficients, and that achieves similar performance to existing work. In addition, we do not need to model or optimize for natural speech variations.

1 Introduction

Speech communication systems, such as mobile telephony, hearing aids and intercom systems, are required to work in numerous environments. As a consequence, the user environment is often noisy which can lead to intelligibility problems.

For speech communication systems we may consider two different environments, cf. Fig. A.1; the far-end environment (at the target talker) and the near-end environment (at the listener). Both the far- and near-end environment are often noisy, which leads to degradations in both speech quality and Speech Intelligibility (SI) for the listener. To remedy these effects, speech enhancement techniques may be applied at both the far- and near-end.

Depending on the availability, Far-end Speech Enhancement (FSE) algorithms may utilize either a single or multiple microphones [1–4]. In contrast to the far-end scenario, in Near-end Listening Enhancement (NLE) [5, 6] the interfering noise is mixed with the target speech after processing. Therefore, the

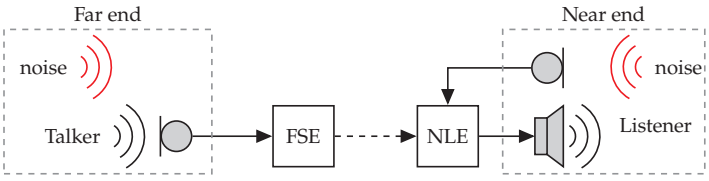


Fig. A.1: Classic speech communication system with Far-end Speech Enhancement (FSE) and Near-end Listening Enhancement (NLE).

noise cannot be reduced by the usual post-processing techniques of the former scenario. Instead, before playback in the noisy environment, NLE increases SI by adaptively pre-processing the FSE signal received from the far-end, while exploiting knowledge about the near-end noise.

Most work on NLE assumes the signal received from the far-end is noise-free [6]. However, in many communication scenarios both the target talker and listener may be in noisy environments. Even so, until recently the processing to mitigate the effects of disturbances in the far- and near-end environments have been considered separately. However, recently, in [7–9] it was shown that optimization of SI under joint consideration of the far- and near-end noise is superior to disjoint processing. The work of [7] considers jointly controlling a single-channel noise reduction filter along with a post-filter gain for NLE designed to increase SI. In [9] a new training strategy is proposed for deep learning based single-channel enhancement given that speech has already been processed at the far-end. For joint multi-microphone FSE and NLE [8, 10] proposes to optimize the Mutual Information (MI) [11] between the clean speech and the signal received by the listener. The results of [8] are the first to show both theoretically and experimentally that joint processing, using knowledge of processing and conditions at both ends, is superior to the classic disjoint processing.

MI as a SI optimization objective provides a target that unifies heuristic views on SI and mathematically founded SI measures [8, 12]. However, solving it in closed form requires simplifying assumptions. The resulting optimization objective is an SNR-type of measure that is approximately equal to the Approximated Speech Intelligibility Index (ASII) [8, 13]. Furthermore, the method of [8] depends on the choice of the correlation of the so-called production and interpretation noise with the clean speech. With the choice made in [8], the objective function of [8] reduces fully to the ASII, whereas for more recent choices in [14] it does not.

In this paper, we illustrate how, by using a simpler well established signal model and optimizing for the ASII directly, we can derive a closed-form solution to the joint far- and near-end speech enhancement problem that is independent of the marginal distribution of signal coefficients, and without the need for introducing additional parameters in terms of a production and interpretation noise model. The proposed approach can also be seen as an extension of the ASII optimization problem [13] to joint far- and near-end optimization. Furthermore, we analyze model choices and assumptions of [8], and how these relate the approximated MI of [8] to the ASII, thus motivating our model choices. Finally, we experimentally compare the performance of [8] using the production noise model that was later derived by some of the same authors in [14–16] to our proposed ASII based optimization. We see, that our proposed method achieves similar or slightly better intelligibility in terms of ESTOI [17] when the near-end SNR is low and the far-end SNR is intermediate

or high.

In summary, the contributions of this paper are: (i) A closed-form solution to ASII based joint far- and near-end speech enhancement, (ii) which is optimal independently of underlying marginal signal distributions, (iii) which does not introduce additional free parameters, e.g., in terms of production and interpretation noise, and (iv) which performs as good or slightly better than existing schemes.

2 Existing Work Based on Mutual Information

MI-based methods for joint far- and near-end SI enhancement [8, 10] improve SI by maximizing the MI, $I(S;Z)$, between the clean speech, S, and the signal received by the listener, Z.

2.1 Existing model assumptions

In [12], production and interpretation noise terms are introduced to model natural variations in speakers and listeners, respectively, and adopted into the signal model of [8]. The production noise, Q, is due to the convolution of the time-domain clean speech and the vocal tract, hence, theoretically, it should be a *multiplicative* noise in the frequency domain. However, in [8] in order to simplify mathematical expressions,

1. Multiplicative production noise is modelled as *additive*.

Thus, the single-microphone signal model of [8] in the absence of processing is, in the complex short-time DFT (STFT) domain,

$$Z_{k,i} = d_{k,i}S_{k,i} + d_{k,i}Q_{k,i} + U_{k,i} + N_{k,i} + W_{k,i}, \tag{A.1}$$

where k is the frequency-bin index and i the time-frame, W is the interpretation noise, U is the far-end environmental noise, N is the near-end environmental noise, and $d_{k,i}$ are the time-frequency coefficients of the room transfer function from target talker to the microphone. The work on MI [8] relies on several common signal model assumptions in the speech processing literature, e.g., that speech and noise STFT coefficients are statistically independent. However, to derive a speech enhancement procedure based on MI, [8] introduces additional assumptions and approximations;

2. The production noise, Q, and interpretation noise, W, are independent of the clean speech level and may be represented by a fixed gain (correlation), $\rho_{0,k}$, at each frequency band.

3. Critical band powers are assumed to be zero-mean independent Gaussian random variables.

In [8], the third assumption is needed since critical band powers are in-fact Chi-squared distributed and the MI between Chi-squared random variables is not expressible in closed form. However, we note that critical band powers are positive by definition, hence a zero-mean model is not necessarily appropriate.

2.2 Approximated Mutual Information vs ASII

The resulting approximated MI expression in [8] is

$$I(S;Z) \approx -\sum_j \frac{1}{2}\log\left(1-\rho_{0,j}^2\frac{\xi_j}{\xi_j+1}\right), \tag{A.2}$$

where ξ_j is the SNR in critical band j. By [8, sec. VIII.A] we may take a first order Taylor approximation of the MI in $\rho_{0,j}^2$ around zero, such that we can approximate the cost function as,

$$I(S;Z) \approx \sum_j I_j\frac{\xi_j}{\xi_j+1}, \tag{A.3}$$

for $\rho_{0,j}^2 \to 0^+$, where $I_j \triangleq -\frac{1}{2}\log(1-\rho_{0,j}^2)$. In the absence of a production noise model at the time, the values for $\rho_{0,j}$ where derived in [8] based on the band importance functions, γ_j, of the SII [18] such that $\rho_{0,j}^2 = 1-2^{-2\gamma_j}$. Inserting this, the cost function is

$$I(S;Z) \approx \sum_j -\frac{1}{2}\log(1-(1-2^{-2\gamma_j}))\frac{\xi_j}{\xi_j+1}. \tag{A.4}$$

We recognize, (A.4) resembles the ASII introduced in [13] as a cost function for SI enhancement. The ASII is defined as,

$$ASII \triangleq \sum_j \gamma_j f(\xi_j), \quad f(\xi_j) \triangleq \frac{\xi_j}{\xi_j+1}, \tag{A.5}$$

where the weights γ_j are the critical-band importance functions as defined in [18], and $f(\xi_j)$ is the audibility function per critical band. We notice from [18, Table 1] that band importance functions, γ_j, are in the interval of $[0.01, 0.06]$, resulting in $\rho_{0,j}^2 \in [0.0138, 0.0798]$. As shown in [10, Fig. 1] the approximation (A.3) holds for $\rho_{0,j}^2 \leq 0.4$. Hence, we can conclude the choice of $\rho_{0,j}^2 = 1-2^{-2\gamma_j}$ to be sufficiently close to zero for equality to hold in (A.4). Thus, the MI problem in [8] is equal to the ASII problem, when the parameter modelling production- and interpretation noise, $\rho_{0,j}$, is chosen according to the band importance functions of the SII.

3 Signal Model

In this section we introduce the proposed signal model, cf. Fig. A.2. The single-microphone signal model follows,

$$X_{k,i} = d_{k,i}S_{k,i} + U_{k,i},\ Y_{k,i} = vX_{k,i},\ Z_{k,i} = Y_{k,i} + N_{k,i}, \tag{A.6}$$

where $X_{k,i}$ is the recorded signal in STFT domain, i.e., the clean speech, $S_{k,i}$, recorded by the microphone contaminated by the far-end noise, $U_{k,i}$. To increase SI of the received message, the noisy microphone signal, $X_{k,i}$, is linearly processed prior to playout, producing the modified signal $Y_{k,i}$. The signal received by the listener, $Z_{k,i}$, is finally contaminated by the noise in the near-end environment, $N_{k,i}$. The speech and noise processes, S, U, and N, are assumed to be stationary sequences of complex random vectors consisting of the STFT coefficients. Both the far-end noise, U, and the near-end noise, N are assumed to be independent of each other and of the target speech, S. These assumptions are similar to [8]. However, compared to [8] we do not need to model a multiplicative production noise as additive, or to introduce an additional interpretation noise. Further, we do not need assumptions on the particular marginal distributions of the signals.

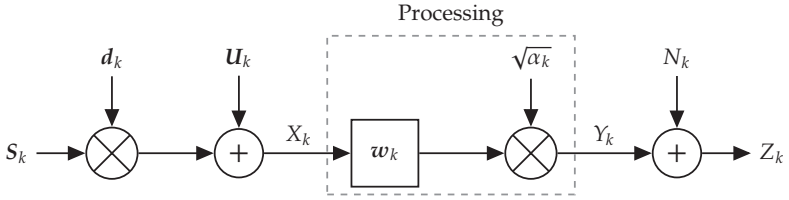


Fig. A.2: Our signal model of optimal joint SI enhancement.

3.1 Multi-Microphone Signal Model

Let us denote the acoustic transfer function from source to microphone m by $d_{k,i,m}$ with vector notation, $\boldsymbol{d}_{k,i} = [d_{k,i,1}, \ldots, d_{k,i,m}]^T$, and letting vector $\boldsymbol{U}_{k,i}$ denote far-end noise recorded by the microphones. Then the noisy microphone signals are given by

$$\boldsymbol{X}_{k,i} = \boldsymbol{d}_{k,i}S_{k,i} + \boldsymbol{U}_{k,i}. \tag{A.7}$$

Denoting the linear multi-microphone processor by $\boldsymbol{v}_{k,i}$, the processed microphone signal is,

$$Y_{k,i} = \boldsymbol{v}_{k,i}^H \boldsymbol{d}_{k,i}S_{k,i} + \boldsymbol{v}_{k,i}^H \boldsymbol{U}_{k,i} \tag{A.8}$$

where super-script H denotes conjugate transpose.

4 Optimal ASII Linear Processor

In this section we derive the optimal linear processor based on the ASII defined in (A.5). The derivation steps are similar to [8] and [13]. However, contrary to [8] we consider the ASII instead of MI. Further, we expand on [13] by including joint (multi-microphone) processing with far-end noise.

The energy of the clean speech signal within one critical band, j, and time-frame, i, is defined as

$$\mathcal{S}_{j,i}^2 \triangleq \sum_k |S_{k,i}|^2 |H_j(k)|^2, \tag{A.9}$$

where $H_j(k)$ is the STFT coefficients of the j'th critical band filter. Similarly, we define the critical band energy of the near-end noise, $\mathcal{N}_{j,i}^2$, and the processed far-end speech, $\widetilde{\mathcal{S}}_{j,i}^2$, and noise, $\widetilde{\mathcal{U}}_{j,i}^2$. Since we assume stationarity of the speech and noise, we can disregard the time-index, i, and let the average energy per DFT bin and critical band be based on a long-term average over several short-time frames,

$$\sigma_{S_k}^2 \triangleq \frac{1}{I} \sum_i |S_{k,i}|^2, \quad \sigma_{\mathcal{S}_j}^2 \triangleq \sum_k |H_j(k)|^2 \sigma_{S_k}^2, \tag{A.10}$$

where I is the total number of frames. Similar definitions hold for the noise terms U and N. The critical band filters, $H_j(k)$, are normalized such that the total energy is preserved in critical bands, i.e.,

$$\sum_j \sigma_{\mathcal{S}_j}^2 = \sum_k \sigma_{S_k}^2. \tag{A.11}$$

The critical band SNR at the near-end listener is then,

$$\xi_j = \frac{\sigma_{\widetilde{\mathcal{S}}_j}^2}{\sigma_{\widetilde{\mathcal{U}}_j}^2 + \sigma_{\mathcal{N}_j}^2}. \tag{A.12}$$

Inserting this into (A.5), we have that

$$f(\xi_j) = \frac{\sum_k |H_j(k)|^2 \boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k \sigma_{S_k}^2}{\sum_k |H_j(k)|^2 \left(\boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k \sigma_{S_k}^2 + \boldsymbol{v}_k^H \Sigma_{U_k} \boldsymbol{v}_k + \sigma_{N_k}^2 \right)} \tag{A.13}$$

$$\triangleq f_j\left(\{\boldsymbol{v}_k, \Theta_k\}\right), \tag{A.14}$$

where $\Theta_k = (\sigma_{S_k}^2, \Sigma_{U_k}, \sigma_{N_k}^2)$. In order to limit loudspeaker overload or unpleasant playback levels we invoke the following equal power constraint,

$$\sum_k \boldsymbol{v}_{k,i}^H \boldsymbol{d}_{k,i} \boldsymbol{d}_{k,i}^H \boldsymbol{v}_{k,i} \sigma_{S_{k,i}}^2 = \sum_k \sigma_{S_{k,i}}^2. \tag{A.15}$$

That is, for each time frame i the total power of the clean speech is unaltered by processing. The joint far and near-end SI enhancement problem with equal power constraint is then,

$$\begin{array}{ll} \sup\limits_{\{\boldsymbol{v}_k\}\in\mathbb{C}^M} & \sum_j \gamma_j f_j\left(\{\boldsymbol{v}_k, \Theta_k\}\right) \\ \text{subject to} & \sum_k \boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k \sigma_{S_k}^2 = \sum_k \sigma_{S_k}^2 . \end{array} \tag{A.16}$$

We now introduce the real and positive variable, α_k, to perform a variable transformation $\boldsymbol{v}_k = \alpha_k^{1/2}\boldsymbol{w}_k$ with the additional constraint $\boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k = \alpha_k$. This leads to the equivalent problem,

$$\begin{array}{ll} \sup\limits_{\{\boldsymbol{v}_k\}\in\mathbb{C}^M, \{\alpha_k\}\in\mathbb{R}_+} & \sum_j \gamma_j f_j\left(\{\alpha_k, \boldsymbol{v}_k, \Theta_k\}\right) \\ \text{subject to} & C_1 \ : \ \sum_k \alpha_k \sigma_{S_k}^2 = \sum_k \sigma_{S_k}^2, \\ & C_2 \ : \ \boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k = \alpha_k, \ \forall k. \end{array} \tag{A.17}$$

The objective function can be rewritten in terms of α_k and $\boldsymbol{w}_k$, i.e.,

$$f_j\left(\{\alpha_k, \boldsymbol{w}_k, , \Theta_k\}\right) = \frac{\sum_k |H_j(k)|^2 \alpha_k \sigma_{S_k}^2}{\sum_k |H_j(k)|^2 \left(\alpha_k \sigma_{S_k}^2 + \alpha_k \boldsymbol{w}_k^H \Sigma_{U_k} \boldsymbol{w}_k + \sigma_{N_k}^2\right)}.$$

We notice that $\boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k = \alpha_k \Leftrightarrow \boldsymbol{d}_k^H \boldsymbol{w}_k = 1$. Hence, writing the optimization problem in terms of $\boldsymbol{w}_k$ and α_k we have,

$$\begin{array}{ll} \sup\limits_{\{\boldsymbol{w}_k\}\in\mathbb{C}^M, \{\alpha_k\}\in\mathbb{R}_+} & \sum_j \gamma_j f_j\left(\{\alpha_k, \boldsymbol{w}_k, \Theta_k\}\right) \\ \text{subject to} & C_1 \ : \ \sum_k \alpha_k \sigma_{S_k}^2 = \sum_k \sigma_{S_k}^2 \\ & C_2 \ : \ \boldsymbol{d}_k^H \boldsymbol{w}_k = 1. \end{array} \tag{A.18}$$

We can separate (A.18) across the two variables [19, p. 133], i.e.,

$$\sup_{\{\alpha_k\}\in\mathbb{R}_+, \mathcal{C}_1} \ \sup_{\{\boldsymbol{w}_k\}\in\mathbb{C}^M, \mathcal{C}_2} \ \sum_j \gamma_j f_j\left(\{\alpha_k, \boldsymbol{w}_k, \Theta_k\}\right).$$

The inner optimization problem across $\{\boldsymbol{w}_k\}$ corresponds to the standard Minimum Variance Distortionless Response (MVDR) beamforming problem with the solution [2],

$$\boldsymbol{w}_k^* = \frac{\Sigma_{U_k}^{-1} \boldsymbol{d}_k}{\boldsymbol{d}_k^H \Sigma_{U_k}^{-1} \boldsymbol{d}_k}, \quad \forall k. \tag{A.19}$$

Inserting the MVDR solution into (A.18), the remaining problem is

$$\begin{array}{ll} \sup\limits_{\{\alpha_k\}\in\mathbb{R}_+} & \sum_j \gamma_j \left(\frac{\sum_k |H_j(k)|^2 \alpha_k \sigma_{S_k}^2}{\sum_k |H_j(k)|^2 \left(\alpha_k \sigma_{S_k}^2 + \alpha_k \sigma_{B_k}^2 + \sigma_{N_k}^2\right)} \right) \\ \text{subject to} & C_1 \ : \ \sum_k \alpha_k \sigma_{S_k}^2 = \sum_k \sigma_{S_k}^2, \end{array} \tag{A.20}$$

where $\sigma_{B_k}^2 \triangleq \boldsymbol{w}_k^{*H} \Sigma_{U_k} \boldsymbol{w}_k^*$ is the residual far-end noise after processing by the MVDR beamformer.

4.1 Critical-band near-end optimization

We derive, similarly to existing work on SI enhancement [6–8, 13], the optimal near-end processor based on the assumption that all frequency gains within a critical band j are the same, i.e.,

$$\alpha_k = \alpha_{k'}, \forall k, k' \in \mathcal{K}_j, \tag{A.21}$$

where $\mathcal{K}_j$ is the set of frequency bins belonging to the j'th critical band. The gains are then later on converted back to DFT domain.

Starting from the optimization problem (A.20) we have,

$$\begin{array}{ll} \sup\limits_{\{\alpha_j\}\in\mathbb{R}_+} & \sum_j \gamma_j \frac{\alpha_j \sigma^2_{\mathcal{S}_j}}{\alpha_j \sigma^2_{\mathcal{S}_j} + \alpha_j \sigma^2_{\mathcal{B}_j} + \sigma^2_{\mathcal{N}_j}} \\ \text{subject to} & C_1 \ : \ \sum_j \alpha_j \sigma^2_{\mathcal{S}_j} = \sum_j \sigma^2_{\mathcal{S}_j}. \end{array} \tag{A.22}$$

We notice each term in the sum is concave in α_j for $\alpha_j \geq 0$. Therefore, the weighted sum of these concave functions is also concave. We describe the problem by the Lagrangian cost-function [19],

$$\mathcal{L} = \sum_j \frac{\gamma_j \alpha_j \sigma^2_{\mathcal{S}_j}}{\alpha_j \sigma^2_{\mathcal{S}_j} + \alpha_j \sigma^2_{\mathcal{B}_j} + \sigma^2_{\mathcal{N}_j}} - \nu \left(\sum_j \alpha_j \sigma^2_{\mathcal{S}_j} - r \right) + \sum_j \lambda_j \alpha_j,$$

where $r = \sum_j \sigma^2_{\mathcal{S}_j}$, and ν and λ_j are the Lagrangian multipliers for the energy constraint and inequality constraint in (A.22). The KKT conditions [19] for the optimization problem are,

$$r = \sum_j \alpha_j \sigma^2_{\mathcal{S}_j}, \ 0 \leq \alpha_j, \ 0 \leq \lambda_j, \ 0 = \lambda_j \alpha_j, \forall j, \tag{A.23a}$$

$$0 = \gamma_j \frac{\sigma^2_{\mathcal{S}_j} \sigma^2_{\mathcal{N}_j}}{\left(\alpha_j \left(\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j} \right) + \sigma^2_{\mathcal{N}_j} \right)^2} - \nu \sigma^2_{\mathcal{S}_j} + \lambda_j, \forall j \tag{A.23b}$$

Isolating λ_j in (A.23b), then using the complimentary slackness condition to set $\lambda_j = 0$, we solve for the non-zero α_j,

$$\alpha_j = \max \left\{ \frac{\sqrt{\sigma^2_{\mathcal{N}_j} \gamma_j}}{\sqrt{\nu} \left(\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j} \right)} - \frac{\sigma^2_{\mathcal{N}_j}}{\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j}}, 0 \right\}, \ \forall j, \tag{A.24}$$

where ν is chosen such that the energy constraint in (A.23a) is satisfied,

$$\frac{1}{\sqrt{\nu}} = \left(r + \sum_{j \in \mathcal{J}} \frac{\sigma^2_{\mathcal{S}_j} \sigma^2_{\mathcal{N}_j}}{\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j}} \right) \Big/ \left(\sum_{j \in \mathcal{J}} \frac{\sigma^2_{\mathcal{S}_j} \sqrt{\sigma^2_{\mathcal{N}_j} \gamma_j}}{\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j}} \right), \tag{A.25}$$

here $\mathcal{J} = \{j \in \mathbb{N} : \alpha_j > 0\}$ denotes the set of frequency bins for which the optimal α_j are positive. We notice, that the set of indices $\mathcal{J}$ depends on the α_j [13]. Therefore, ν also depends on α_j. Hence, there is a recursive relationship between (A.24) and (A.25), which may be resolved by using, e.g., a bi-section method or evaluating (A.24) for a range of ν values, such that the energy constraint is satisfied [13]. Finally, to get the optimal frequency dependent gains, α_k^*, we weight the optimal, α_j^*, using the critical band filters, that is

$$\alpha_k^* = \sum_j |H_j(k)|^2 \alpha_j^*, \tag{A.26}$$

where the energy constraint is satisfied in both frequency and critical-band domain as per the normalization of the critical band filters in (A.11), i.e., $\sum_j \alpha_j^* \sigma_{\mathcal{S}_j}^2 = \sum_j \sigma_{\mathcal{S}_j}^2 = \sum_k \sigma_{S_k}^2 = \sum_k \alpha_k^* \sigma_{S_k}^2$.

The proposed processor is summarized in Fig. A.2. As in [8], the procedure consists of an MVDR beamformer, $\boldsymbol{w}_k$, followed by a frequency dependent gain, α_k. In contrast to the results of [8], the frequency bin gains, α_k, of our procedure are optimized specifically according to the ASII without approximations and assumptions on signal distributions and without introduction of additional free parameters to model natural speech variations.

5 Experimental Evaluation

We have seen (Section 2) that by using the production noise model choice in [8] the MI cost function reduces to the ASII. However, this choice of production noise is due to a lack of a more accurate model at the time. Recently, some of the authors have provided an estimated model for the production noise in [14–16], where $\rho_{0,j} = 0.75 \; \forall j$. We compare performance of our proposed method with [8] using the newer production noise model of [14]. We consider a Python simulation of a setup similar to that of [8, 20].

5.1 Experimental Setup

We consider a room with dimensions $3 \times 4 \times 3\,\mathrm{m}^3$, a single target speaker located at $[1.50, 3.00, 1]\,\mathrm{m}$, and an array with two microphones spaced 2 cm apart at $[1.50, 2.00, 1]\,\mathrm{m}$ and $[1.50, 2.02, 1]\,\mathrm{m}$. At the far-end there are three speech shaped noise sources located at $[0.50, 1.00, 1]\,\mathrm{m}$, $[0.75, 3.00, 1]\,\mathrm{m}$ and $[3.00, 1.60, 1]\,\mathrm{m}$, respectively. The near-end noise is pink. In addition to the far-end noise each microphone is subject to a 60 dB SNR white noise. The source signal is speech signals from five female and five male speakers from the TIMIT-database [21] sampled at 16 kHz. Signals were processed block-wise based on a DFT with 32 ms Hann windows with 50% overlap. Since we

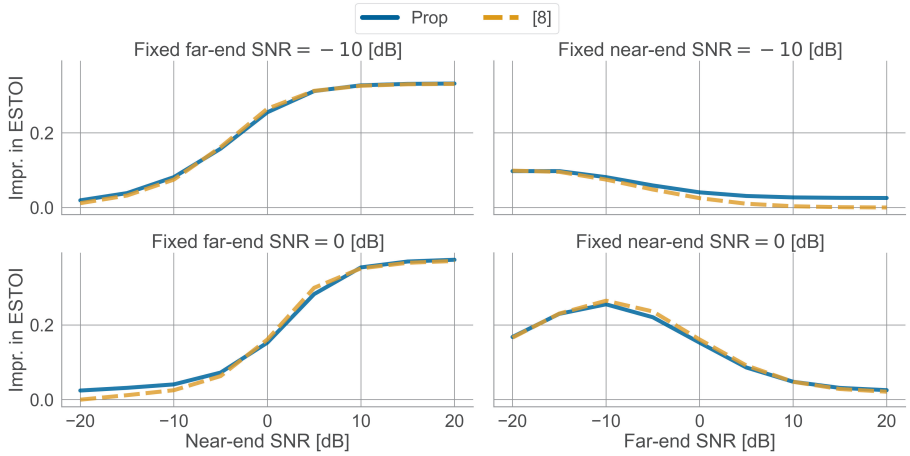


Fig. A.3: ESTOI performance of the proposed method and [8], for varying near-end SNR and a fixed far-end SNR (left column), and varying far-end SNR and a fixed near-end SNR (right column).

assume stationarity, the MVDR beamformer, $\boldsymbol{w}_k$ and post-filter gains, α_k, are derived based on the long-term average spectrums, leading to time-invariant processing. The long-term power of the clean speech, far-end noise, and the near-end noise are assumed to be known. Hence, we do not estimate any of these spectrums. Furthermore, the room transfer functions are assumed to be known, and generated without reverberation using [22].

5.2 Results

Fig. A.3 shows improvements over 'unprocessed' in ESTOI [17], of the proposed method along with the method of [8]. The results show that generally the two methods achieve a similar performance, which is expected due to the similarity shown in Section 2. However, the proposed method is slightly better when the near-end SNR is low and the far-end SNR is intermediate or high. Thus, the production noise model choice of [14] leads to a slightly worse speech enhancement than using the model based on the SII weights [18]. ASII plots show identical performance between the two methods and are thus not reported.

6 Conclusion

We have derived a closed-form optimal linear processor for joint far- and near-end speech intelligibility enhancement based on ASII. The optimal processor

consists of an MVDR beamformer followed by a frequency dependent post gain. The derived processor is based on a simple model without relying on assumptions and approximations of the underlying marginal signal distributions. Furthermore, we do not need to model or optimize for natural variations in speech. Finally, as a consequence, the proposed processor has comparable or slightly better ESTOI performance than existing work.

References

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

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

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

[4] K. Eneman, H. Luts, J. Wouters, M. Büchler, N. Dillier, W. Dreschler, M. Froehlich, G. Grimm, V. Hohmann, R. Houben, A. Leijon, A. Lombard, D. Mauler, M. Moonen, H. Puder, M. Schulte, A. Spriet, and M. Vormann, "Evaluation of signal enhancement algorithms for hearing instruments," in *2008 16th European Signal Processing Conference*, Aug. 2008, pp. 1–5.

[5] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[6] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. N. Petkov, B. Sauert, and P. Vary, "Optimizing Speech Intelligibility in a Noisy Environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, Mar. 2015.

[7] M. Niermann, P. Jax, and P. Vary, "Joint Near-End Listening Enhancement and far-end noise reduction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 4970–4974.

[8] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Intelligibility Enhancement Based on Mutual Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1694–1708, Aug. 2017.

[9] K. Tan and D. Wang, "Improving Robustness of Deep Learning Based Monaural Speech Enhancement Against Processing Artifacts," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 6914–6918.

[10] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Jointly optimal near-end and far-end multi-microphone speech intelligibility enhancement based on mutual information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, Mar. 2016, pp. 654–658.

[11] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Hoboken, N.J: Wiley-Interscience, 2006.

[12] W. B. Kleijn and R. C. Hendriks, "A Simple Model of Speech Communication and its Application to Intelligibility Enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 303–307, Mar. 2015.

[13] C. H. Taal, J. Jensen, and A. Leijon, "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[14] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, Jan. 2018.

[15] ——, "An intelligibility metric based on a simple model of speech communication," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.

[16] ——, "On the information rate of speech communication," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 5625–5629.

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

[18] American National Standards Institute, *Methods for Calculation of the Speech Intelligibility Index*. New York, N.Y: Acoustical Society of America, 2017, vol. ANSI S.35-1997.

[19] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge University Press, 2004.

[20] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Joint near-end and far-end intelligibility enhancement," https://cas.tudelft.nl/Repository/repitem.php?id=3&ti=2, Aug. 2017, accessed: 2021-02-15.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

[22] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.

Paper B

Minimum Processing Near-end Listening Enhancement

Andreas Jonas Fuglsig, Jesper Jensen, Zheng-Hua Tan,
Lars Søndergaard Bertelsen, Jens Christian Lindof and
Jan Østergaard

The paper has been published in the
IEEE/ACM Transactions on Audio, Speech, and Language Processing Vol. 31,
pp. 2233–2245, 2023.

The layout has been revised.

Abstract

The intelligibility and quality of speech from a mobile phone or public announcement system are often affected by background noise in the listening environment. By pre-processing the speech signal it is possible to improve the speech intelligibility and quality — this is known as near-end listening enhancement (NLE). Although, existing NLE techniques are able to greatly increase intelligibility in harsh noise environments, in favorable noise conditions the intelligibility of speech reaches a ceiling where it cannot be further enhanced. Actually, the focus of existing methods solely on improving the intelligibility causes unnecessary processing of the speech signal and leads to speech distortions and quality degradations. In this paper, we provide a new rationale for NLE, where the target speech is minimally processed in terms of a processing penalty, provided that a certain performance constraint, e.g., intelligibility, is satisfied. We present a closed-form solution for the case where the performance criterion is an intelligibility estimator based on the approximated speech intelligibility index and the processing penalty is the mean-square error between the processed and the clean speech. This produces an NLE method that adapts to changing noise conditions via a simple gain rule by limiting the processing to the minimum necessary to achieve a desired intelligibility, while at the same time focusing on quality in favorable noise situations by minimizing the amount of speech distortions. Through simulation studies, we show the proposed method attains speech quality on par or better than existing methods in both objective measurements and subjective listening tests, whilst still sustaining objective speech intelligibility performance on par with existing methods.

1 Introduction

Real-life speech communication, e.g., with mobile phones or public announcements, takes place in a large variety of often noisy places. Here, environmental noises such as cars, trains, construction work and other people talking may interfere with speech perception and degrade both the speech intelligibility (SI) and the speech quality (SQ). In near-end listening scenarios, the noise sources are physically present in the environment where the listener is located, cf. the right-hand part of Fig. B.1. Therefore, in the near-end scenario, we cannot extract the clean speech signal by removing noise from a noisy input signal as done in far-end speech enhancement, when there is noise in the left-hand side of Fig. B.1, e.g., [1, 2]. Instead, several other techniques exist for increasing the SI and SQ in noise by modifying the speech signal received from the far-end prior to playback in the noisy environment, and are known as Near-end Listening Enhancement (NLE).

Both SI and SQ are important factors for the listening experience, but the importance of SI and SQ changes depending on the acoustic situation [3–5]. Sometimes the requirements to SI and SQ may even be at a conflict and a

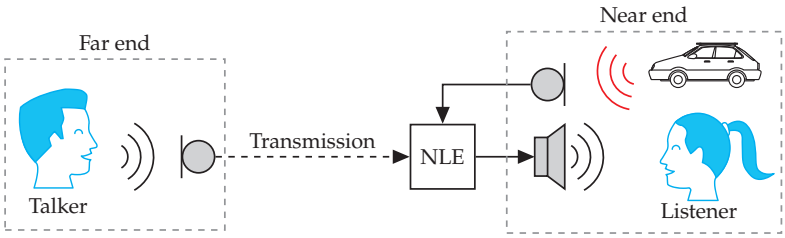


Fig. B.1: Basic principle of near-end listening enhancement (NLE), where the received far-end signal is processed prior to playout in a noisy environment.

trade-off must be made [5]. Increased intelligibility can lead to an increase in experienced SQ in noisy conditions [4, 5]. That is, when listening in noise, the SI is an important contributing factor to the experienced SQ [5]. Even though NLE processing may introduce distortion to the target speech signal, which could lead to decreases in SQ, these speech distortions may be masked by the harsher environmental noise in these noisy situations [5, 6]. However, speech becomes naturally more intelligible as noise conditions improve [3–5]. Hence, in favorable noise conditions, the intelligibility of unprocessed speech approaches 100 % so that it cannot be further enhanced [3]. In fact, further excessive or unnecessary processing of the speech signals may lead to speech distortions and quality degradations because the noise can no longer mask these processing distortions [5]. Therefore, it may be useful to adapt the NLE processing to the noise situations, such that it can be activated and deactivated in the best possible way [5].

Traditionally, the goal of NLE algorithms is solely to maximize the SI for the listener in the noisy environment by modifying the time-frequency characteristics of the input speech such that it is not masked by the environmental noise, cf. the reviews in [7–9] and the contributions to the Hurricane challenges [10, 11]. Existing NLE techniques can be categorized into two main classes. The first one is the heuristic or expert driven approaches, e.g., [12–24]. These consider various approaches regarding, e.g., formants, transients, Lombard speech, spectral shaping and dynamic range compression, or plain audibility. Although the class of heuristically based NLE techniques provide significant improvements in SI, they are often not derived according to specific objective optimality criteria. Instead, they are non-parametric and based on subjective expert experiences and knowledge. Hence, they cannot claim optimality and maybe are not adapting well to changing environments.

The second class of NLE algorithms is based on the main idea of manipulating the input speech such that a target intelligibility metric is maximized when the noise conditions are known. One of the most widely used optimization targets for SI enhancement is the Speech Intelligibility Index (SII) [25] or variations thereof, which have been used for NLE in numerous studies, e.g., [26–32].

The SII based approaches [26, 27] show good performance but fall behind the state-of-the-art heuristic [19] in subjective tests [10, 11], because the optimization targets do not correlate well with subjective intelligibility across varying noises and degradations [33]. Furthermore, some methods, e.g., [26, 27], require solving an optimization problem in real-time with varying execution time [23]. The SII based solutions also rely on simplifying assumptions about frequency gains being constant across frequency subbands [32]. Recently, deep neural network (DNN) based approaches [34, 35] have been able to optimize more advanced measures such as the extended short-time objective intelligibility (ESTOI) measure [36] that correlate more with subjective tests than the simpler metrics such as SII [33]. However, DNN methods come at the cost of significant memory usage, black-box solutions, and possible problems in generalizing to new acoustic scenarios.

We note for near-end listening with headphones, adaptive noise cancellation (ANC) techniques [37, 38] may be employed, which — instead of processing the target speech signal — aim at adaptively cancelling the noise by adding an anti-phase noise component to the speech signal before playout in the noisy environment [39, 40]. However, ANC with classic adaptive filtering has generally been insufficient for improving intelligibility outside headphone use until the use of DNNs [39]. Hence, NLE based on speech modification is still the predominant approach.

Common for the objective SI metrics such as SII and the Glimpse model [41] is that audibility is the decisive factor of intelligibility. Therefore, the principle for all NLE algorithms is to adjust the SNR for the perceptually important parts of the speech [39]. Hence, the simplest solution to the NLE problem is to increase the power of the clean speech signal until the noise is sufficiently masked, i.e., increasing the SNR (almost indefinitely). However, such an increase in speech power may lead to various problems, most significantly possible hearing damage to the near-end listener, but also problems with loudspeaker overload and unpleasant playback levels [28]. Therefore, many NLE approaches take a power constraint into account, such that an equal power constraint is maintained between the unprocessed and processed clean speech signal. However, this also considerably limits the potential for increasing the SI [24, 42].

Existing methods are designed to always achieve the desired output power level. Thus, the power is always increased (or decreased) to the maximum allowed level even though this may not be necessary and may lead to excessive processing and a decrease in SQ or SI [5, 8]. Thus, increasing the allowed output power too much, in order to overpower the noise, may cause problems for SQ. Hence, we need to control the amount of processing in a different way, such that we can achieve both good SI and SQ depending on the noise situation.

In this work, we take inspiration from the area of far-end noise reduction.

Particularly, the work of Zahedi et al. [43], where the concept of *minimum processing beamforming* is proposed. In this new rationale, the goal is to ensure a minimum level of SI performance, while guiding the noise reduction capabilities of the beamforming towards a particular performance; ambient-preserving, aggressive noise reduction or standard Wiener filtering.

In this paper, we propose the concept of *minimum processing near-end listening enhancement*. The proposed concept provides a more general formulation of the NLE problem that focuses on both SI and SQ in an adaptive manner. The concept is based on minimizing a speech processing penalty subject to a certain intelligibility performance constraint. The point is no longer only to maximize SI but instead process the target speech signal just enough to achieve a minimum desired intelligibility in the given noise conditions, while minimizing the amount of speech distortions in favorable noise situations. Thus, the concept is adapting to the noise conditions by increasing SI when needed and ensuring an automatic focus on SQ, when the SI is already sufficient by automatically reducing the amount of processing when the desired SI is achieved.

We present an exemplary case study where the processing penalty is the mean-square error (MSE) between the processed near-end signal and the clean speech, and the performance criterion used in the optimization is an intelligibility estimator based on the approximated SII (ASII) [27]. While it is common in SII based works to assume that gains in the same subband are the same, we do not assume this upfront in this work. Instead, the optimization problem in this work is specifically built as a function of the gains in STFT domain to formulate a more general optimization problem without this common assumption. The case study shows we are able to adapt to changing noise conditions via a simple gain rule that is computationally efficient and which does not require online optimization. Furthermore, it is a result of our work that for the given case study the gains within a subband are the same, thus the derived solution mathematically validates the simplifying assumption of constant gains made in existing SII based works.

Finally, the proposed NLE method achieves SQ performance on par or better than existing methods in both objective measurements and subjective listening tests, while maintaining objective SI performance on par with existing methods for a wide range of SNRs.

The rest of the paper is organized as follows. In Section 2 we introduce our signal model. Section 3 introduces the proposed minimum processing NLE and we make our case study. Section 4 introduces practical considerations and summarizes the algorithm. Section 5 presents the experimental setup and objective SI and SQ performance. Section 6 presents a subjective SQ listening test and discusses the results. We conclude the paper in Section 7.

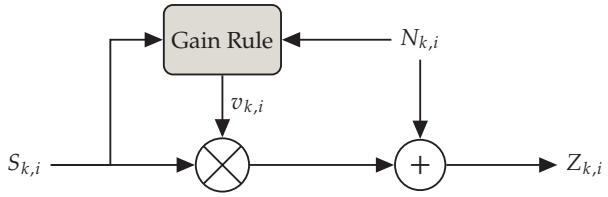


Fig. B.2: Block diagram representation of frequency domain NLE signal model.

2 Signal Model

In this work we consider a signal model where speech and noise are represented in the complex short-time discrete Fourier transform (STFT) domain, cf. Fig. B.2, Here $S_{k,i}$ is the clean speech signal received from the far-end in frequency-bin k and time-frame i. In most general settings the far-end signal is noisy, however, we assume an adequately clean version is available or can be achieved through proper noise reduction at the far-end. $N_{k,i}$, is the time-frequency representation of the additive near-end noise. As is common in the literature, we assume the speech, S, and noise, N, are uncorrelated of each other. To increase the SI (and SQ) of the signal received by the near-end listener, the speech signal, S, is linearly pre-processed prior to play out in the near-end environment. Thus, we assume that the signal presented to the near-end listener in STFT domain, $Z_{k,i}$, follows

$$Z_{k,i} = v_{k,i} S_{k,i} + N_{k,i}, \tag{B.1}$$

where $v_{k,i}$ denotes the linear pre-processing gains. The NLE gains constitute the core of NLE, and determining a proper NLE gain rule is the core task of NLE.

2.1 Subband Model

We focus on NLE based on (the perceptually driven) ASII for speech intelligibility prediction. ASII is part of a family of well-known speech intelligibility and quality predictors such as ESTOI [36], STOI [44], SII [25], extended SII (ESII) [45], the hearing-aid speech perception index (HASPI) [46], the hearing-aid speech quality index (HASQI) [47] and perceptual evaluation of speech quality (PESQ) [48], that all mimic aspects of human speech perception. That is, signals are analyzed in, e.g., octave bands, fractional octave bands or critical bands similar to the human ear. In general, these bands are referred to as subbands. Perceptually motivated subbands may be defined such that multiple frequency bins contribute to the same and/or multiple subbands, with an individual weight for each frequency bin-subband pair. We index subbands with index j and frequencies with index k.

To illustrate the use of subbands in this work, let $\omega_{j,k}$ denote the non-negative filter weights that implement the j'th subband filter, $\mathbb{B}_j$ denote the set of frequency bins that contribute to the j'th subband with $j \in \{1, \ldots, J\}$ and J the total number of subbands. Then, the clean speech spectrum level within one subband, j, and time-frame, i, is defined as

$$\sigma^2_{S_{j,i}} \triangleq \sum_{k \in \mathbb{B}_j} \omega_{j,k} \sigma^2_{S_{k,i}}, \tag{B.2}$$

where $\sigma^2_{S_{k,i}}$ is the clean speech spectrum level in time-frequency bin k, i. For the sake of brevity and ease of reading we assume any normalization of the subband filter weights, $\omega_{j,k}$, is already included in the weights. We elaborate further on the definition of $\omega_{j,k}$ and the relation between subbands and frequency bins in Appendix A.

Finally, we assume the speech and noise are processes of complex random vectors comprised of the STFT coefficients. In practice one could estimate the statistics of these processes online, and the mathematical framework can be applied to the time-varying case on a per time-frame basis. Therefore, for brevity of notation we disregard the time-index, i, and assume we are considering a certain time frame i, unless otherwise is stated.

3 Minimum Processing Near-End Listening Enhancement

3.1 Concept

For a particular subband j, we denote by $\boldsymbol{S}_j \in \mathbb{C}^{|\mathbb{B}_j|}$ the vector containing all S_k for $k \in \mathbb{B}_j$, where $|\mathbb{B}_j|$ denotes the number of frequency bins in the j'th subband. Similarly, we create the vector $\boldsymbol{Z}_j$ by stacking Z_k for $k \in \mathbb{B}_j$. Furthermore, let $\mathcal{D}_j(\cdot,\cdot)$ and $\mathcal{I}_j(\cdot,\cdot)$ be finite non-negative functionals, where $\mathcal{D}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right)$ measures the distortion (processing penalty) between the clean speech signal, S, and the signal presented to the near-end listener, Z, and $\mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right)$ is an intelligibility or performance estimator for the NLE in subband j. Then, the minimum processing near-end listening enhancer in subband j is defined as the solution to the following optimization problem:

$$\min_{\{v_k\} \in \mathbb{R}_+,\, k \in \mathbb{B}_j} \quad \mathcal{D}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right) \quad \text{s.t.} \quad \mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right) \geq I'_j. \tag{B.3}$$

The term I'_j is defined as

$$I'_j = \min\left(I_j, I_j^{\max}\right), \tag{B.4}$$

where I_j is a desired minimum requirement on the NLE intelligibility performance $\mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right)$ and $I_j^{\max}$ is the maximum achievable performance when disregarding the processing penalty $\mathcal{D}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right)$, i.e., when the performance $\mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right)$ is maximized in an unconstrained manner. We highlight the generality of the problem formulation in (B.3), in that $\mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right)$ is a general "performance criterion" which could reflect any type of performance aspect of interest to the application, e.g., SI, SQ or other measures of interest.

3.2 Case Study

In this paper, we study the case, where the processing penalty, $\mathcal{D}_j$, is the MSE criterion, and the performance criterion, $\mathcal{I}_j$, is an intelligibility estimator based on the ASII [27]. We solve the problem analytically for any given minimum performance constraint and subband definition. Furthermore, we show how we can select I_j to achieve a desired total minimum performance across all subbands, A^*, i.e, we guarantee that the optimum solution satisfies

$$\sum_j \mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right) \geq A^*. \tag{B.5}$$

3.2.1 Processing Penalty

To ensure the processing does not distort speech excessively and the playback volume is not increased infinitely, we introduce a processing penalty [43]. We consider a MSE processing penalty, where the MSE is evaluated in subband domain instead of DFT domain to ensure compatibility with the minimum processing NLE formulation in (B.3). That is, the j'th subband MSE processing penalty is

$$\mathcal{D}_j(\boldsymbol{S}_j, \boldsymbol{Z}_j) = \sum_{k \in \mathbb{B}_j} \omega_{j,k} \left(1 - v_k\right)^2 \sigma_{S_k}^2. \tag{B.6}$$

See Appendix B for the derivation.

3.2.2 Performance Criterion

We consider, as an example of the proposed frame work, a performance criterion based on the ASII [27]. The original work on minimum processing in [43] considered a performance criterion based on the SII [25]. Both SII and ASII consider intelligibility to be a weighted sum of intermediate measures of the audibility of speech in a subband as a function of the subband SNR. In SII, the band audibility is determined by a function where the long-term SNR is log-transformed and clipped between -15 dB and 15 dB and normalized to range between zero and one. In ASII, the audibility is determined by a non-linear

approximation of the log-transform and clipping via a sigmoidal function of the subband SNR. The advantage of using the ASII is that its nonlinear approximation of the SII provides a nice mathematical tractability circumventing the need for dealing with clipping of SNRs. Based on the ASII, have the following performance criterion for the j'th subband,

$$\sum_{k\in\mathbb{B}_j} \omega_{j,k} v_k^2 \sigma_{S_k}^2 \geq \sigma_{\mathcal{N}_j}^2 I_j^{\xi}, \tag{B.7}$$

where $I_j^{\xi} \triangleq \frac{I_j}{1-I_j}$, and I_j is the desired subband audibility performance. The details are shown in Appendix C.

3.3 Optimization Problem and Solution

Joining the results of (B.6) and (B.7), the minimum processing NLE problem (B.3), which we consider, and its solution is given in the following theorem.

Theorem 1 *The minimum processing NLE problem* (B.3) *with MSE processing penalty* (B.6) *and ASII performance criterion* (B.7) *is*

$$\begin{array}{ll} \min\limits_{\{v_k\}\in\mathbb{R}_+, k\in\mathbb{B}_j} & \sum_{k\in\mathbb{B}_j} \omega_{j,k} \left(1-v_k\right)^2 \sigma_{S_k}^2, \\ \textit{subject to} & \sum_{k\in\mathbb{B}_j} \omega_{j,k} v_k^2 \sigma_{S_k}^2 \geq \sigma_{\mathcal{N}_j}^2 I_j^{\xi}. \end{array} \tag{B.8}$$

The optimal minimum processing NLE *gains are*

$$v_{k,j}^{MP} = \begin{cases} 1 & \textit{if } \sigma_{S_j}^2 \geq \sigma_{\mathcal{N}_j}^2 I_j^{\xi} \\ \sqrt{\frac{\sigma_{\mathcal{N}_j}^2 I_j^{\xi}}{\sigma_{S_j}^2}} & \textit{otherwise} \end{cases}, \quad \forall k \in \mathbb{B}_j. \tag{B.9}$$

In Appendix D we derive the *minimum processing NLE*, i.e., the optimal gains in (B.8).

We immediately have the following result regarding the optimal minimum processing NLE subband gains (B.9).

Corollary 1 *The optimum gains for the minimum processing NLE* (B.8) *are equal for all frequencies within a subband j.*

This is an important result because most work on subband SNR based enhancement make this assumption for convenience in the joint optimization across all subbands [9, 27, 30, 32, 49]. However, as we show, it can be deduced from (B.9) that the optimal solution does not depend upon k, and it is optimal to have the same gain across the entire subband when optimizing for each subband individually.

For the optimal gain (B.9), the first case takes effect when the unprocessed speech signal already satisfies the performance constraint, i.e., the speech has an intelligibility at or above the desired level. Hence, there is no reason to process the speech signal and the optimal solution is to do nothing, i.e., $v_{k,j}^{\mathrm{MP}} = 1$ for $k \in \mathbb{B}_j$, and thus maximize SQ by minimizing the MSE.

In the second case, the unprocessed speech is below the desired audibility level or equivalently the desired SNR. The optimal gain then increases the speech power such that the subband SNR is exactly at the desired level. That is, the speech power in the j'th subband is increased just enough to satisfy the subband audibility constraint. This illustrates how the optimal solution provides the minimum required processing necessary to achieve the desired intelligibility. Thus, the two cases of (B.9) illustrate how our approach has the advantage of achieving a desired intelligibility level, while automatically adjusting the processing to minimize processing artifacts.

Squaring the gains of (B.9), $(v_{k,j}^{\mathrm{MP}})^2$, we can see that the optimal processed speech power (B.32) is increased proportionally to the near-end noise power. This is also the naturally expected solution to ASII and SII based NLE approaches given sufficient ability to increase speech power [24, 26, 27]. Thus, we provide a mathematical justification for the heuristic approach of increasing speech power proportionally to the noise in [24].

Depending on the choice of subband definition, the subbands may overlap such that multiple frequencies contribute to multiple subbands indexed by $\mathbb{F}_k$. Therefore, the optimum gain, $v_{j,k}^*$ may also contribute to multiple subbands as indicated by the dependence on both j and k. Using the weights of these contributions, $\omega_{j,k}$, we can weigh the filter gains through the subband filters. Thus, we have the following corollary.

Corollary 2 *The NLE processor in frequency bin k is given as*

$$v_k^{MP} = \sqrt{\sum_{j \in \mathbb{F}_k} \omega_{j,k} \left(v_{k,j}^{MP} \right)^2}. \tag{B.10}$$

From this corollary, we see for overlapping subbands, that gains that were previously equal do not remain the same after being projected back to STFT domain.

4 Practical Considerations

4.1 Preventing excessive sound levels

The proposed optimal minimum processing increases speech power to the necessary level to achieve the desired intelligibility. However, this may lead

to speech levels that are unpleasant for the user. In extreme cases, if not controlled, such high signal levels could even cause hearing damage. Usually in most NLE work, this is prevented by a power equality constraint [24, 27]. To prevent excessive output levels, in our work, we put an upper limit on the maximum processed speech power within each subband as described in [31, Sec. 2.2.5]. The idea is to limit the gain applied to each subband, j, such that the resulting subband power of the enhanced speech signal does not increase beyond a maximum subband power, $P_S^{\max}$. That is,

$$\overline{v}_{k,j}^{\mathrm{MP}} = \min\left\{v_{k,j}^{\mathrm{MP}}, v_{j,\max}\right\} \quad \forall k \in \mathbb{B}_j, \tag{B.11}$$

with maximum subband gain

$$v_{j,\max} = \sqrt{\frac{P_S^{\max}}{\sigma_{S_j}^2}}. \tag{B.12}$$

This is done prior to the subband filtering of the gains in (B.10). Thus, $\overline{v}_{k,j}^{\mathrm{MP}}$ is inserted on the right-hand side of (B.10).

4.2 Choosing the intelligibility limit per band

The proposed minimum processing NLE as well as the work of [43] considers a performance constraint, I_j, for each particular subband $j \in \{1, \ldots, J\}$. Thus, the processed speech achieves a desired performance in each particular subband. This target must be decided upon for each of the J subbands before processing. In this section, we show how, when using the ASII criterion, we can select and achieve a single total target intelligibility, A^*, instead of having to select a desired performance constraint for each of the J subbands. This single total target intelligibility is then converted to a per band intelligibility target.

Let I_j be a given minimum requirement on the performance in a particular subband, j, i.e., we require

$$f(\xi_j) \geq I_j. \tag{B.13}$$

Then the processed ASII score satisfies

$$\mathrm{ASII} = \sum_j \gamma_j f(\xi_j) \geq \sum_j \gamma_j I_j. \tag{B.14}$$

That is, given that the minimum performance requirement is met for each band, the processed ASII is greater than or equal to a weighted sum of the performance criterions, where the weights are the band importance functions.

The band importance functions, γ_j, in the SII [25] describe how some subbands are more important for intelligibility than others. Therefore, it is natural to require higher audibility limits for the more important subbands

and lower audibility limits for the less important bands. We can achieve this by using the band importance functions.

Let A^* be the required minimum total ASII performance, and let the subband audibility limits be given as

$$I_j = \frac{A^* \gamma_j}{\sum_i \gamma_i^2}. \tag{B.15}$$

Then by inserting in (B.14) the resulting ASII score satisfies

$$\text{ASII} \geq \sum_j \gamma_j \frac{A^* \gamma_j}{\sum_i \gamma_i^2} = \frac{A^*}{\sum_i \gamma_i^2} \sum_j \gamma_j^2 = A^*, \tag{B.16}$$

that is, the total processed ASII score is lower bounded by the required performance. This means we can guarantee a total ASII greater than or equal to a target ASII, A^*.

As a special case, we see that removing the weights and selecting a fixed audibility limit across subbands $I_j = A^*$, $\forall j$ also achieves ASII $\geq A^*$. In our simulations we use the weighted audibility limits, as these have shown a slightly better performance than having the same fixed limit for each subband.

4.3 Estimating statistics

When processing time-varying signals, such as speech and non-stationary noise, the statistics of the signals change over time. Therefore, the statistics must be estimated and updated in time using, e.g., recursive averaging. However, if the statistics are updated too fast or abruptly, the optimal gains in (B.9) can change suddenly between time-frames, which causes audible distortions to the target speech signal. To circumvent this, it is common in the literature, e.g., [24, 43], to use slowly time-varying processing, where the recursive averaging is across several frames (seconds). Therefore, in this work, we let the average energy per DFT bin and critical band be based on a long-term average over several short-time frames,

$$\sigma_{S_k}^2 \triangleq \frac{1}{I} \sum_i |S_{k,i}|^2, \tag{B.17}$$

$$\sigma_{S_j}^2 \triangleq \sum_k \omega_{j,k} \sigma_{S_k}^2, \tag{B.18}$$

where I is the total number of frames. Similar expressions hold for the near-end noise signal $N_{k,i}$. Thus, the estimated statistics in this work do not change over time, and the implemented processing is time-invariant. For time-varying estimates, (B.17) should be updated to a recursive or moving average.

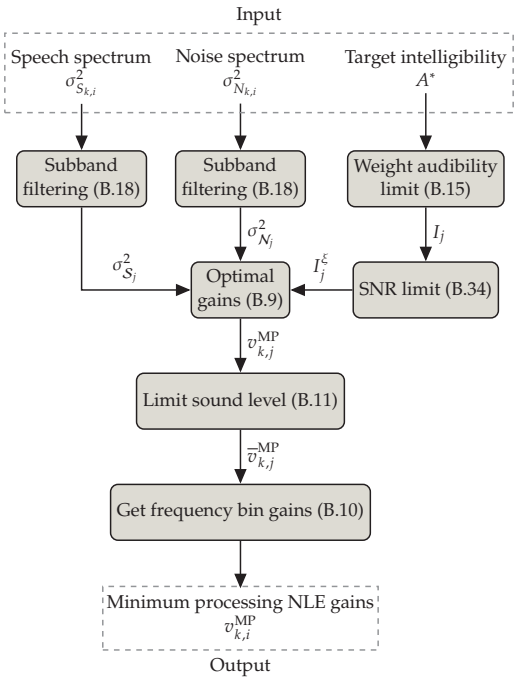


Fig. B.3: Flowchart of the proposed minimum processing NLE gain rule.

4.4 Algorithm summary

The proposed minimum processing NLE algorithm with ASII intelligibility criterion is summarized with the flowchart in Fig. B.3.

5 Objective Performance Evaluation

In this section, we evaluate the performance of the proposed minimum processing NLE method (B.9) in the NLE setup shown in Fig. B.2, where the target speech signal is processed prior to play out in the noisy environment, and the gain rule follows the proposed algorithm summarized in Fig. B.3. We evaluate objective SI and SQ performance, and compare performance with existing related state-of-the-art NLE methods. For intelligibility scoring, we use the chosen performance criterion ASII [27] and the more general intelligibility predictor ESTOI [36]. The goal of minimum processing is to reproduce the clean unprocessed input speech signal as well as possible. Therefore, quality is estimated using PESQ [50] and Segmental SNR (Seg-SNR) [51].

5.1 Experimental Setup

The speech material used for the evaluation are English sentences each a few seconds in duration from the test set in the TIMIT-database [52] sampled at 16 kHz. In each trial a speaker is selected randomly without replacement across trials. Afterwards, a random sentence is selected amongst those available for the selected speaker. Finally, the speech excerpt is padded with 0.5 s of silence at the beginning and 0.125 s of silence at the end.

For the noise signals we consider white noise, synthetic speech shaped noise, as well as noise recorded inside a car traveling at 130 km/h. For the recorded noise, a random excerpt of the appropriate length is cut out of the noise recording for each trial.

Speech and noise signals were transformed to time-frequency domain using an STFT with 32 ms Hann windows with 50% overlap at a sampling frequency of 16 kHz. For the subbands, we consider a total of $J = 30$ overlapping auditory filters with linearly spaced center frequencies on the equivalent rectangular bandwidth scale from 150 Hz to 8000 Hz [53], see Appendix A for further details on the subband definition.

The target intelligibility is set to $A^* = 0.7$, unless otherwise is stated. We choose this target, because an ASII value of 0.7 corresponds to almost full intelligibility [33], and since any higher value requires significantly more processing (see also Section 5.3). The band intelligibility limits, I_j are weighted according to (B.15) for all trials.

For the maximum subband gain, $v_{j,\max}$, in (B.12), the maximum allowed power $P_S^{\max}$ is chosen such that

$$10 \log \left(\frac{P_S^{\max}}{P_0} \right) = 100 \, \mathrm{dB_{SPL}}. \tag{B.19}$$

Here P_0 represents the digital reference power corresponding to a reference sound pressure [31]. We make the convention that a signal power of P_0 corresponds to a signal with an $RMS = 1$, thus

$$P_0 = 10^{\frac{-100 \, \mathrm{dB_{SPL}}}{10}}. \tag{B.20}$$

Finally, we conduct a total of 10 trials and evaluate the performances in each trial. The average performance across the trials is then taken as the final score for each combination of noise, SNR and gain rule.

5.2 Reference methods

We consider three reference methods that are similar to our work: (1) The original ASII optimization of [27] because we also optimize for ASII, (2) the well known SII optimization of [26] that clips SNRs above and below a certain

level and (3) the very recent NoiseProp algorithm of [24] because it has a simple gain-rule very similar to ours. The method of [24] has an additional tuning parameter, $\rho_{NV} \in [0, 1]$. It is mentioned in [24], that informal listening has shown $\rho_{NV} \in [0.5, 0.8]$ is a good choice. Therefore, we select the value $\rho_{NV} = 0.7$ for all experiments.

All reference methods are implemented following the procedure described in the previous subsection. The only difference is the applied gain rule. That is, for ease of comparison all methods are implemented as time-invariant even though the original work may have considered slowly time-varying processing, e.g., [24, 26].

All the reference methods are based on a target speech power equality constraint. Hence, they are designed to keep the total power across frequency at a certain level, P_{ref}. However, the proposed method is able to increase the power as much as needed, achieving a total processed power level denoted by P_{MP}. Therefore, for a fair comparison, the reference methods are also allowed to increase the power to that level. That is, the reference methods are implemented such that $P_{\text{ref}} = P_{\text{MP}}$.

5.3 Minimum Processing Effect

Initially, we illustrate the effects of the proposed minimum processing solution in terms of processing penalty versus intelligibility performance and resulting increases in speech power. The effects are considered using a white noise source to limit the importance of the spectral distribution of noise.

Fig. B.4 shows that for increasing values of the target intelligibility, A^*, the achieved ASII (top) along with the corresponding MSE processing penalty (middle) and increase in speech power (bottom). The lower limit case of target AII $A^* = 0$ corresponds to the case of no processing. The ASII is upper bounded by $A^* \leq 1$ and is only achievable in the limit case of infinite SNR, therefore we only show the trend up to $A^* = 0.9$. The top panel in Fig. B.4 shows, that when the target intelligibility $A^* > 0$, we are able to improve the performance over the unprocessed case ($A^* = 0$) for a large range of SNRs. From the middle and bottoms panels in Fig. B.4, we see that the greater the target intelligibility and the lower the input SNR, the increases in intelligibility performance come at the cost of a greater need for processing and increased speech power. Finally, as the near-end SNR improves, the target intelligibility is more easily achieved and the processing required to achieve a given target decreases, as can be seen by a transition from a high level of processing to no processing.

5.4 Objective Performance

Fig. B.5 shows the ASII, ESTOI, PESQ and Seg-SNR performance as function of the near-end SNR for the proposed minimum processing NLE and the refer-

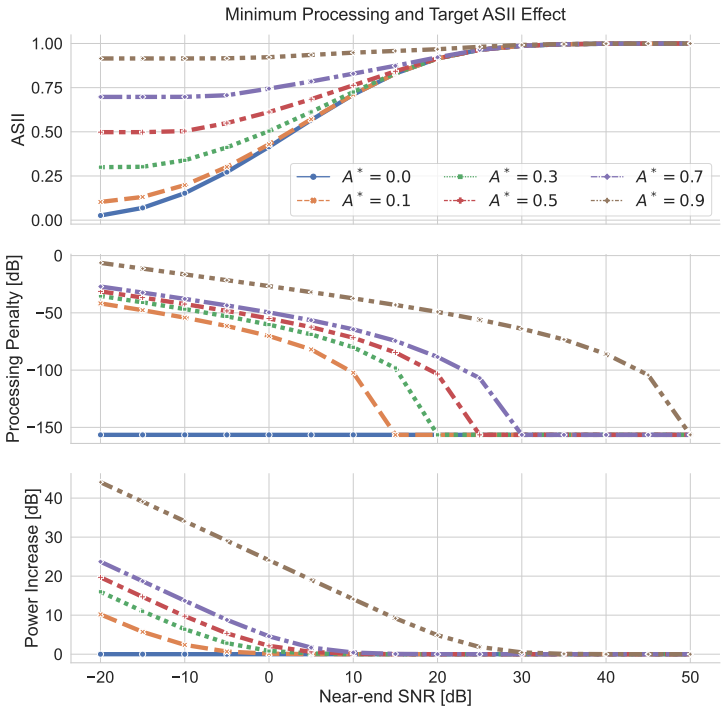


Fig. B.4: The achieved ASII (top), and the corresponding MSE processing penalty (middle) and power increase (bottom) versus input near-end SNR for varying choices of target intelligibility. The legend in the top panel applies to all plots.

ence methods in both car noise and speech shaped noise. Further experiments using cafeteria and babble noise sources show performance similar to experiments using speech shaped noise, however these are omitted here do to space limitations.

5.4.1 Estimated intelligibility performance

From the SI metrics ASII and ESTOI, we see that the proposed method improves SI over the unprocessed speech to the desired level at the lower SNRs. The proposed method even achieves the best ESTOI performance in car noise compared to the reference methods. In ASII, the reference methods are able to gain a slightly better SI than the proposed. This is expected, as they try to maximize the SI given the allowed power constraint and not just reach a sufficient level of SI. As the SNR increases and the speech becomes naturally intelligible, the proposed method follows the unprocessed performance above the target intelligibility. Thus, the proposed method always achieves the desired SI performance or better. Similar performance can be seen with SII maximization by

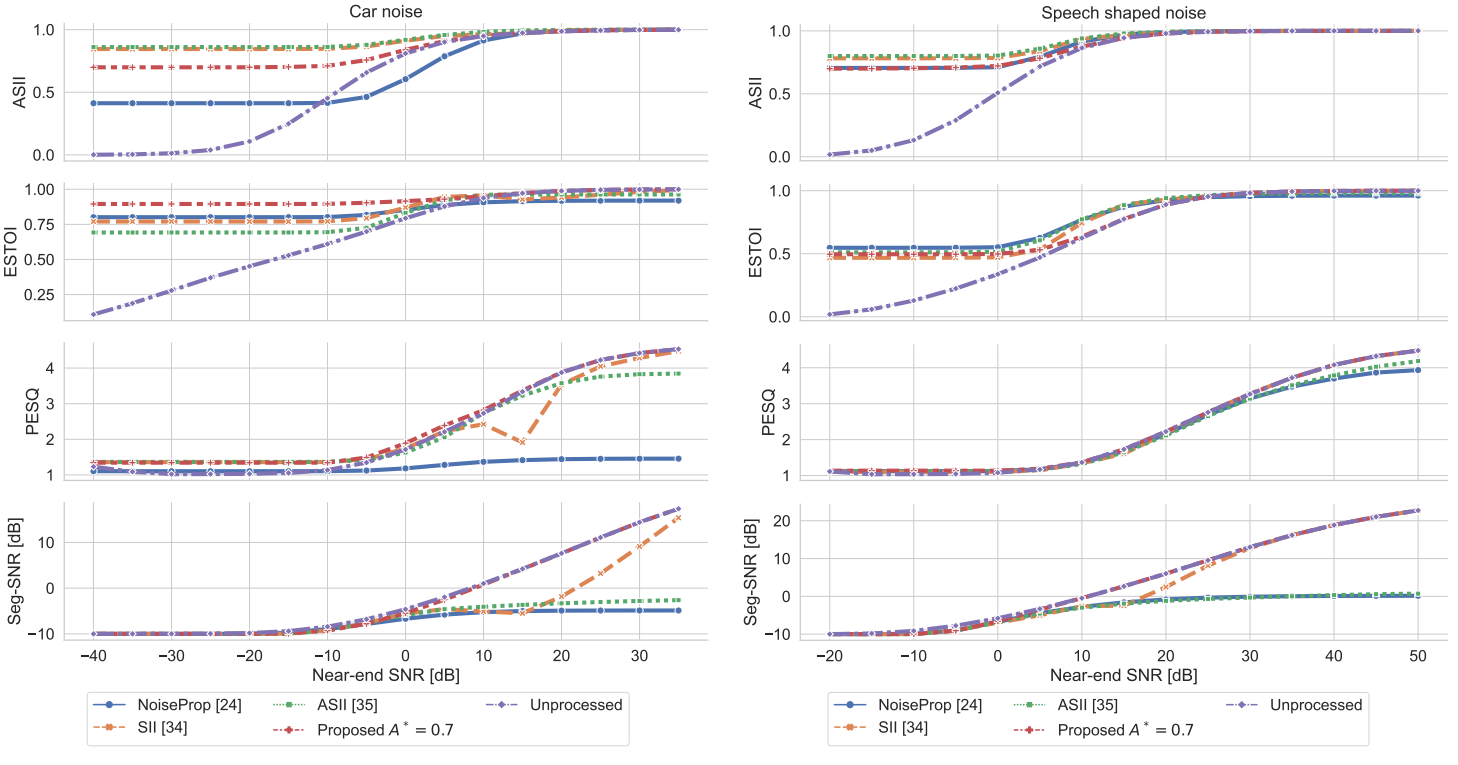


Fig. B.5: ASII, ESTOI, PESQ and Seg-SNR scores in (a) car noise and (b) speech shaped noise. Legend apples to all plots.

Sauert [26], however this does not happen until the SII clips the subband SNRs above 15 dB and the SII is fully maximized. Additionally, as the unprocessed SI reaches its ceiling the reference methods are no longer able to further enhance SI by maximization.

5.4.2 Estimated speech quality performance

Generally, from the PESQ and Seg-SNR plots, we see that all Seg-SNR and PESQ scores are at the lowest levels for very low SNRs, and it is difficult to tell any difference in quality performances. This is expected, since the near-end SNR is low and the noise is the dominant part of the signals. As the input SNR increases so does the PESQ and Seg-SNR scores, where the unprocessed clean speech is considered to be the maximum level of SQ. This is due to PESQ and Seg-SNR measuring quality by how similar the signal presented to the listener, Z, is to the clean reference, S. Hence, for high SNR this is the maximum achievable level when the near-end noise is insignificant and $Z \approx S$.

For increasing SNRs in both car noise and speech shaped noise, we see that the proposed method follows the maximum SQ performance of the unprocessed signal, when the intelligibility is at the desired level and the processing switches off. Similarly, the method of Sauert [26] switches off automatically at maximum SII, but does so at a higher level of intelligibility, thereby incurring more speech distortions than the proposed method. On the other hand, the reference methods of Taal [27] and Niermann [24] never switch off and continue processing at the higher SNRs causing large speech distortions.

Comparing between car noise and speech shaped noise we see that the PESQ and Seg-SNR show less speech distortion in speech shaped noise. Further, since car noise is a less severe noise, speech distortions occur at lower SNRs and are more clearly heard in car noise. Thus, excessive processing is more detrimental to SQ in better noise conditions.

5.5 Gain dynamics

To show the advantages of minimum processing in comparison with existing methods, we illustrate how the proposed method and [24] affect the speech spectra in different noise conditions. We consider the behavior in car noise for an input SNR of -30 dB and 10 dB. In these SNRs both the proposed method and [24] achieve very high SI, however they have a vastly different SQ at 10 dB as seen in Fig. B.5a.

Fig. B.6 shows an example of the clean speech and noise spectrum (top row), as well as the derived optimal gains for both methods (bottom row) for both -30 dB SNR (left column) and 10 dB SNR (right column). Looking at the left-hand column with low SNR, we see that the proposed method provide optimal gains that raise the speech power sufficiently above the near-end noise, and the

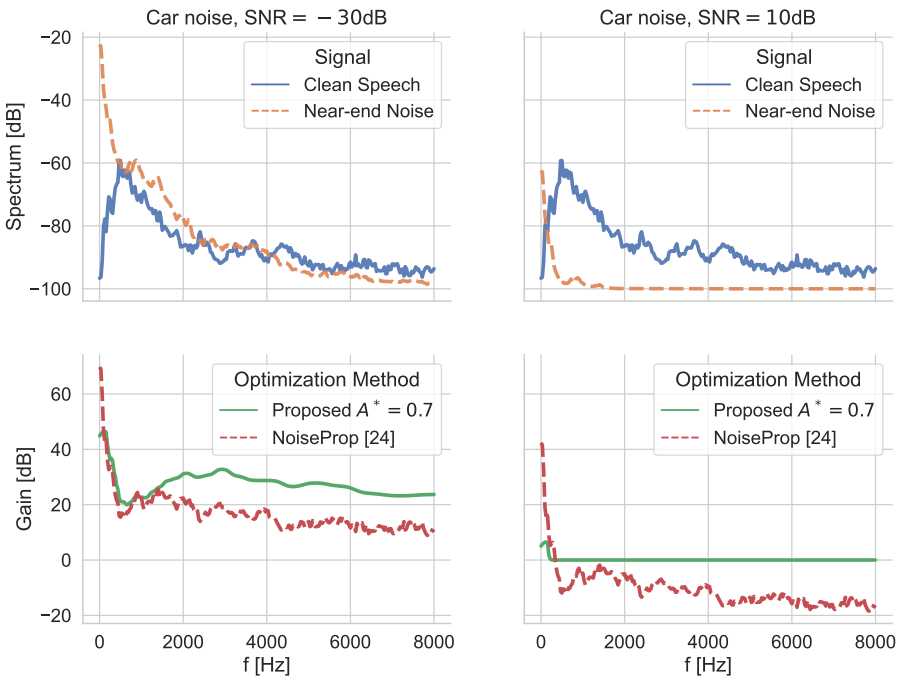


Fig. B.6: Clean speech and near-end noise spectra (top row), and optimal gains (bottom row) for car noise in −30 dB SNR (left column) and 10 dB SNR (right column).

reference method [24] optimal gains are proportional to the near-end noise. Now considering the difference between the two SNRs (columns), we see how the proposed method automatically minimizes the amount of processing as the SNR increases and the SI is at the ceiling level. This results in high SI at low SNRs, and both high SI and SQ at high SNRs in Fig. B.5a. On the other hand, the reference [24] method, which always tries to maximize SI, continues to process the speech signal proportionally to the near-end noise even at the high SNR when SI is at the ceiling level. This continued processing introduces audible speech distortions, which result in low SQ scores at the higher SNRs as seen in Fig. B.5a.

6 Subjective Speech Quality Test

In this section, we evaluate the performance of the proposed minimum processing NLE method (B.9) by a subjective listening test for SQ. Speech quality and general audio quality is highly subjective and may not always be represented well by objective measures. For example, grading SQ in the presence of noise using PESQ is difficult due to a high sensitivity to the environmental

noise compared to the speech distortion [43]. Therefore, to further investigate the performance of the proposed procedure we perform a listening test to asses the SQ in the near-end listening scenario, where we cannot process the noise but only the clean speech component. We consider the proposed minimum processing in comparison with the highly similar and state-of-the-art NoiseProp method of [24]. We investigate performance in both car noise and speech shaped noise.

6.1 Listening Test Setup

The listening test was conducted by 21 (4 female, 17 male) volunteering untrained listeners. The participants had an age span of 22 to 59 years with an average age of 39.6 years. All participants had self-reported normal hearing. The average test time including training was 38 minutes.

The listening test was conducted in a silent room using a Lenovo T570 laptop equipped with a USB sound card (DragonFly Black) and a pair of closed headphones (Beyerdynamic DT-770 Pro 32 ohm) for reporting and audio playback. The user interface was based on the Web Audio Evaluation Tool [54]. All audio stimuli were normalized to the same perceived loudness of -30 LUFS following [55] and the EBU R128 [56] recommendation for loudness normalization as implemented within the Web Audio Evaluation Tool [57]. Participants were allowed to adjust the general volume to a comfortable level during the training session of the test.

6.2 Procedure

We conducted a listening test following the MUlti Stimulus with Hidden Reference and Anchor (MUSHRA) [58] paradigm, where the audio quality is assessed on a scale from 0 to 100, divided into five equal intervals labelled as *bad, poor, fair, good* and *excellent*. The participants were instructed to grade the *basic audio quality* compared to a known reference signal. No other definition was provided of audio quality.

Each test participant was presented with 2 sequences of 8 trials, one sequence for each of the car and speech shaped noise. Each trial consisted of a reference signal (unprocessed signal in high SNR) and five other signals to be rated: 1 hidden reference, 2 systems under test (proposed method and NoiseProp [24]), 1 unprocessed signal, and 1 hidden anchor (unprocessed signal at lower SNR). The SNRs considered for the reference, anchor and test cases vary depending on the noise type, such that the perceptual quality was not affected too severely by a loss in intelligibility. For car noise, the reference and anchor signal SNRs were set to 15 dB and -30 dB respectively. For speech shaped noise, the reference and anchor signal SNRs were set to 25 dB and -5 dB respectively. For both noise types, four trials were used to evaluate the systems

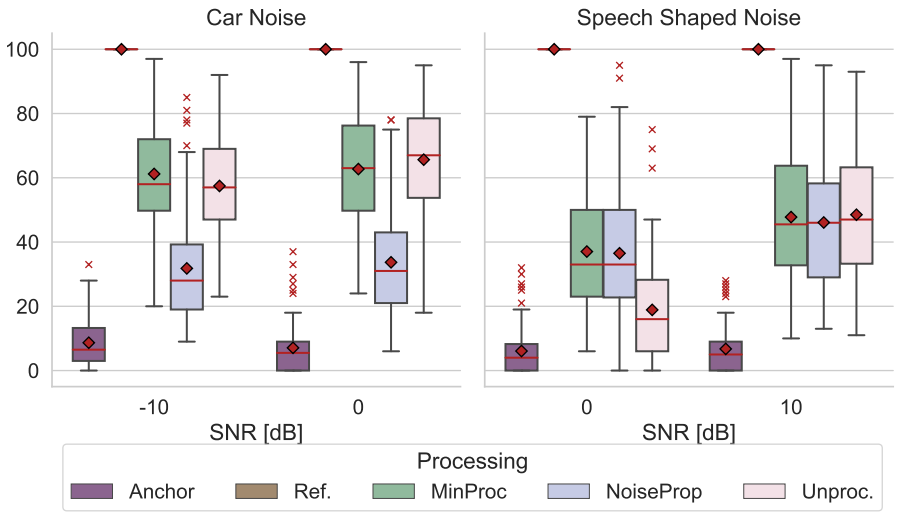


Fig. B.7: Boxplot illustration of the MUSHRA listening test results for car noise (left) and speech shaped noise (right). Medians and means are indicated by red horizontal lines and diamonds, respectively. Outliers (according to the 1.5 interquantile range rule) are indicated by red crosses.

at an input SNR of 0 dB. The last 4 trials were used to evaluate the systems at −10 dB and 10 dB for car noise and speech shaped noise, respectively.

Before the actual experiment, each participant went through a short training session, where they were able to get used to the task at hand, the different audio stimuli and the user interface. The sentences in the training were different from those in the actual experiment and did not contribute to the final test score.

6.3 Listening Test Results and Discussion

The average scores across trials and participants for each noise, processing and SNR condition are shown in Fig. B.7 with boxplots.

The speech shaped noise results at an SNR of 0 dB in Fig. B.7, show that processing with the proposed method and NoiseProp [24] to increase intelligibility has a positive effect on subjective quality compared to the unprocessed performance, when the noise situation is more severe. As the SNR increases to 10 dB, we cannot detect any effect on the subjective SQ compared to the unprocessed performance. This is in contrast to the PESQ results in Fig. B.5b, where no difference was seen.

The listening test results for car noise at both −10 dB and 0 dB SNR show that NoiseProp [24] deteriorates SQ, while the proposed minimum processing does not. Both methods have comparable SI performance at these SNRs with the proposed method slightly outperforming the NoiseProp, as shown by

ESTOI in Fig. B.5a. The SQ results are also confirmed with PESQ at 0 dB SNR in Fig. B.5a, where there is a score difference of almost 2 points. At input SNR of −10 dB PESQ only shows a minor difference between the two methods. This illustrates, when the unprocessed SI is almost maximized due to favorable noise conditions, the excessive processing of NoiseProp, as seen Fig. B.6, can harm subjective SQ, even though the SI may be increased slightly.

These observations confirm the results of the objective SQ metrics as well as those in the literature [5], that SI enhancement may lead to both decreases and increases in SQ. Particularly, decreases in SQ are observed when the unprocessed intelligibility is already high and the processing becomes excessive. On the other hand, in harsh noise conditions, the environmental noise masks most speech distortions. This is also in line with the results of [5], where SI is shown to be the dominant factor for SQ in severe noise.

Existing NLE methods focus on maximizing SI. Therefore, given a certain allowed increase in speech power, the reference methods are sometimes able to gain a slightly higher SI than the proposed method as the ASII and ESTOI scores show in Fig B.5. However, the natural ceiling effect of SI occurs prior to the maximum value of the metric [33]. Therefore, the objective SQ and listening test results also show the achieved SI score may be unnecessarily high and lead to excessive speech distortion. Instead, by using a single intelligibility target, A^*, our proposed method and can be better linked to the onset of the ceiling effect for various noise types. Thus, by varying the target intelligibility, A^*, we see how the proposed method allows controlling the tradeoff between SI and SQ.

7 Conclusion

We proposed a (novel) near-end listening enhancement concept, where the output signal is optimized to have the minimum amount of processing artifacts (for example, distance from the clean speech signal), with the constraint that a certain performance criterion (for example, estimated intelligibility) is satisfied. The proposed concept is adaptive to environmental noise conditions by focusing on estimated intelligibility for high noise levels, and, as a consequences of the minimum processing paradigm, quality for low noise levels. The trade-off between intelligibility and quality can be controlled via the performance criterion. To demonstrate an instance of the proposed framework, we make a thorough investigation of an example case, where the performance criterion is based on the approximated speech intelligibility index and the processing metric is the mean squared error. We show that, this provides a simple gain rule based NLE, that raises the speech spectrum above the noise to exactly the minium required level for the desired intelligibility. The proposed method, is able to distribute the processing across subbands to maximize or

reach a target intelligibility by varying the desired intelligibility target between the subbands. If the noise conditions are favorable, the processing is automatically reduced leading to fewer processing artifacts and hence improved speech quality. Thereby, by allowing an increase in speech power, the proposed NLE technique is able to adaptively achieve any desired intelligibility level with a minimum of processing artifacts. Experimental studies verify the advantages of the concept in terms of both intelligibility and quality.

Future work will include expanding the proposed concept to jointly consider near-end listening enhancement together with far-end noise reduction for an increased intelligibility and quality performance. Additionally, it is interesting combining the proposed method with acoustic echo cancellation techniques when working with joint far-end and near-end optimization.

A Subband Weights

Assigning frequency bins to subbands and determining the weight of their contributions can be done in different ways, cf. [43, App. A]. We consider a gammatone filter bank model [53] that we normalize to preserve power between subband and frequency domain.

We focus on subbands in terms of overlapping auditory filters. Let h_j be the impulse response of the j'th subband auditory filter. Then, the energy of the clean speech signal, $\mathcal{S}_{j,i}$, is given as the convolution with between s and h_j, which in in time-subband domain is given as

$$\mathcal{S}_{j,i}^2 \triangleq \sum_{k\in\mathbb{B}_j} |S_{k,i}|^2 |H_j(k)|^2, \tag{B.21}$$

where $H_j(k)$ represents the DFT of h_j in frequency-bin k. We consider squared magnitude responses, $|H_j(k)|^2$ given by the gammatone filter bank derived in [53]. Furthermore, we want to ensure the total power is the same in both the frequency and subband domain, i.e.,

$$\sum_j \sigma_{\mathcal{S}_{j,i}}^2 = \sum_k \sigma_{S_{k,i}}^2. \tag{B.22}$$

Now inserting the filtering operation we have that

$$\sum_j \sigma_{\mathcal{S}_{j,i}}^2 = \sum_j \sum_k |H_j(k)|^2 \sigma_{S_{k,i}}^2 \tag{B.23}$$

$$= \sum_k \sum_j |H_j(k)|^2 \sigma_{S_{k,i}}^2. \tag{B.24}$$

From (B.24) we see that (B.22) is satisfied if $\sum_j |H_j(k)|^2 = 1$. We can achieve this by normalizing $H_j(k)$, i.e.,

$$H'_j(k) = \frac{H_j(k)}{\sqrt{\sum_l^J |H_l(k)|^2}}, \quad \forall j. \tag{B.25}$$

Hence, we let the subband filter weights, $\omega_{j,k}$, be the normalized squared magnitude response of h_j, i.e., $\omega_{j,k} = |H'_j(k)|^2$.

B MSE Processing Penalty

For the j'th subband we write the MSE processing penalty as

$$\mathcal{D}_j(\boldsymbol{S}_j, \boldsymbol{Z}_j) = \sum_{k \in \mathbb{B}_j} \omega_{j,k} \, \mathrm{E}\left[|S_k - Z_k|^2\right] \tag{B.26}$$

$$= \sum_{k \in \mathbb{B}_j} \omega_{j,k} \left((1 - v_k)^2 \, \sigma^2_{S_k} + \sigma^2_{N_k} \right), \tag{B.27}$$

where we have assumed the STFT coefficients are uncorrelated across frequency. Thereby, the MSE in a subband is the 'weighted average' of the MSE in the DFT domain where the weights are the subband filter weights. Since the near-end noise power, $\sigma^2_{N_k}$, is unaffected by the processing, v_k, we can disregard it from the processing penalty, and we have

$$\mathcal{D}_j(\boldsymbol{S}_j, \boldsymbol{Z}_j) = \sum_{k \in \mathbb{B}_j} \omega_{j,k} \, (1 - v_k)^2 \, \sigma^2_{S_k}. \tag{B.28}$$

C ASII performance criterion

The ASII [27] is defined as

$$ASII \triangleq \sum_j \gamma_j f(\xi_j), \tag{B.29}$$

where the weights γ_j are pre-defined constants (more specifically the band importance functions as defined in [25]),

$$f(\xi_j) \triangleq \frac{\xi_j}{\xi_j + 1}, \tag{B.30}$$

is the sigmoidal audibility function per subband, and ξ_j is the SNR per subband, i.e.,

$$\xi_j \triangleq \frac{\sigma^2_{\widetilde{S}_j}}{\sigma^2_{N_j}}, \tag{B.31}$$

where

$$\sigma^2_{\widetilde{S}_j} = \sum_{k\in\mathbb{B}_j} \omega_{j,k} v_k^2 \sigma^2_{S_k}, \tag{B.32}$$

is the processed speech power. Hence, the *processed* subband SNR is the ratio of the *processed* speech power to the near-end noise power at the listener.

Let I_j be a given minimum requirement on the audibility performance in a particular subband, j, i.e., we require

$$f(\xi_j) \geq I_j. \tag{B.33}$$

It is beneficial to interpret the subband audibility limits, I_j, in terms of the subband SNR.Using (B.30), it follows that (B.33) can be written as

$$\xi_j \geq \frac{I_j}{1-I_j} \triangleq I_j^{\xi} \tag{B.34}$$

where $I_j^{\xi} \triangleq \frac{I_j}{1-I_j}$, may be considered a desired minimum processed SNR. Therefore, instead of focusing on choosing an intelligibility limit, I_j, we can equivalently choose an SNR limit I_j^{ξ}, or interpret the desired audibility limits as a limit on the subband SNR. Hence, the performance criterion in (B.3) is chosen as the subband SNR, i.e.,

$$\mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right) = \xi_j = \frac{\sigma^2_{\widetilde{S}_j}}{\sigma^2_{\mathcal{N}_j}}. \tag{B.35}$$

Writing out the performance criterion we have that

$$\mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right) \geq I_j^{\xi} \tag{B.36}$$

$$\sum_{k\in\mathbb{B}_j} \omega_{j,k} v_k^2 \sigma^2_{S_k} \geq \sigma^2_{\mathcal{N}_j} I_j^{\xi}. \tag{B.37}$$

D Proof of Theorem 1

We first see that the optimization problem (B.3) can be written as (B.8) by inserting (B.6) and (B.7) in (B.3).

To prove that the solution to (B.8) is given by the weights in (B.9), we start by making an observation on the relationship between the MSE cost function and the performance constraint in (B.8). Clearly, the unconstrained solution to the MSE minimization problem is $v_k = 1$ for $k \in \mathbb{B}_j$, i.e., no processing obviously minimizes the MSE.

The interesting scenario is then that of $\sum_{k\in\mathbb{B}_j} \omega_{j,k}\sigma^2_{S_k} < \sigma^2_{\mathcal{N}_j} I_j^{\xi}$. Since we have no energy constraint and we must increase the speech energy to not violate

the performance constraint we can lower bound the gains as $v_k \geq 1$ for all $k \in \mathbb{B}_j$, i.e., we only increase the speech power and have no reason to decrease it. Hence, the optimization problem becomes

$$\begin{aligned} \min_{\{v_k\} \in \mathbb{R}_+, k \in \mathbb{B}_j} \quad & \textstyle\sum_{k \in \mathbb{B}_j} \omega_{j,k} \left(1 - v_k\right)^2 \sigma_{S_k}^2, \\ \text{subject to} \quad & \textstyle\sum_{k \in \mathbb{B}_j} \omega_{j,k} v_k^2 \sigma_{S_k}^2 \geq \sigma_{\mathcal{N}_j}^2 I_j^{\xi}, \\ & v_k \geq 1. \end{aligned} \tag{B.38}$$

Formulating the Lagrangian we have

$$\begin{aligned} \mathcal{L}(\{v_k\}, \lambda_j, \{\mu_k\}) = & \sum_{k \in \mathbb{B}_j} \omega_{j,k} \left(1 - v_k\right)^2 \sigma_{S_k}^2 + \mu_k (1 - v_k) \\ & + \lambda_j \left(\sigma_{\mathcal{N}_j}^2 I_j^{\xi} - \sum_{k \in \mathbb{B}_j} \omega_{j,k} v_k^2 \sigma_{S_k}^2 \right) \end{aligned}$$

Taking the derivative with respect to v_k for a particular $k \in \mathbb{B}_j$, we have

$$\frac{\partial}{\partial v_k} \mathcal{L} = -2\omega_{j,k} \left(1 - v_k\right) \sigma_{S_k}^2 - 2\lambda_j \omega_{j,k} v_k \sigma_{S_k}^2 - \mu_k. \tag{B.39}$$

Since we optimize over a non-convex set the optimization problem is non-convex. Thus, the Karush-Kuhn-Tucker conditions for this problem are not sufficient for a global optimum. However, the they are still necessary conditions, thus we use these to determine a solution to the problem.

$$v_k \geq 1, \quad \mu_k \geq 0, \quad \mu_k (1 - v_k) = 0, \quad \lambda_j \geq 0 \tag{B.40a}$$

$$\sigma_{\mathcal{N}_j}^2 I_j^{\xi} - \sum_{k \in \mathbb{B}_j} \omega_{j,k} v_k^2 \sigma_{S_k}^2 \leq 0 \tag{B.40b}$$

$$\lambda_j \left[\sigma_{\mathcal{N}_j}^2 I_j^{\xi} - \sum_{k \in \mathbb{B}_j} \omega_{j,k} v_k^2 \sigma_{S_k}^2 \right] = 0 \tag{B.40c}$$

$$-2\omega_{j,k} \left(1 - v_k\right) \sigma_{S_k}^2 - 2\lambda_j \omega_{j,k} v_k \sigma_{S_k}^2 - \mu_k = 0 \tag{B.40d}$$

Isolating μ_k in the last equation we have,

$$-2\omega_{j,k} \left(1 - v_k\right) \sigma_{S_k}^2 - 2\lambda_j \omega_{j,k} v_k \sigma_{S_k}^2 \geq 0. \tag{B.41}$$

Now solving for v_k,

$$v_k \geq \frac{1}{1 - \lambda_j} \tag{B.42}$$

Combining this with $v_k \geq 1$, we have

$$v_k = \max\left(1, \frac{1}{1-\lambda_j}\right). \tag{B.43}$$

We observe that $v_k = v_{k'}$ for all $k, k' \in \mathbb{B}_j$. That is, the gains are equal for all frequencies within subband j.

Now to determine λ_j, we first notice that if $\lambda_j = 0$ then $v_k = 1\ \forall k \in \mathbb{B}_j$. Then the performance constraint can only be satisfied if $\sigma^2_{\mathcal{S}_j} \geq \sigma^2_{\mathcal{N}_j} I_j^{\xi}$, i.e., if we are in the scenario that does not require processing. Therefore, we must have $\lambda_j > 0$ if $\sigma^2_{\mathcal{S}_j} < \sigma^2_{\mathcal{N}_j} I_j^{\xi}$. Furthermore, from (B.43) we see that the non-trivial solution requires $\lambda_j < 1$. Therefore, (B.40c) is only satisfied for $\lambda_j \in (0, 1)$ if the optimal λ_j satisfy

$$\sigma^2_{\mathcal{N}_j} I_j^{\xi} - \sum_{k \in \mathbb{B}_j} \omega_{j,k} \frac{1}{\left(1-\lambda_j\right)^2} \sigma^2_{\mathcal{S}_k} = 0 \tag{B.44}$$

$$\lambda_j = 1 - \sqrt{\frac{\sigma^2_{\mathcal{S}_j}}{\sigma^2_{\mathcal{N}_j} I_j^{\xi}}}. \tag{B.45}$$

Inserting this into (B.43) the optimal gain is

$$v_k^* = \max\left(1, \sqrt{\frac{\sigma^2_{\mathcal{N}_j} I_j^{\xi}}{\sigma^2_{\mathcal{S}_j}}}\right) = \sqrt{\frac{\sigma^2_{\mathcal{N}_j} I_j^{\xi}}{\sigma^2_{\mathcal{S}_j}}}, \tag{B.46}$$

where the second equality follows since we consider the case of $\sigma^2_{\mathcal{N}_j} I_j^{\xi} > \sigma^2_{\mathcal{S}_j}$. Finally, by inserting the optimum v_k^*, λ_j^* and μ_k^* in (B.40d), it can be seen that this is a stationary point of the Lagrangian and thus dual feasible. Hence, the solution is

$$v_k^* = \begin{cases} 1 & \text{if } \sigma^2_{\mathcal{S}_j} \geq \sigma^2_{\mathcal{N}_j} I_j^{\xi} \\ \sqrt{\frac{\sigma^2_{\mathcal{N}_j} I_j^{\xi}}{\sigma^2_{\mathcal{S}_j}}} & \text{otherwise} \end{cases}, \quad \forall k \in \mathbb{B}_j. \tag{B.47}$$

This completes the proof.

References

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

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[3] J. Rennies, A. Pusch, H. Schepker, and S. Doclo, "Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements," *The Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. EL315–EL321, Oct. 2018.

[4] R. Pricken, M. Wältermann, E. Parotat, M. Soloducha, and A. Raake, "Quality Aspects of Near-End Listening Enhancement Approaches in Telecommunication Applications," in *Proceedings of DAGA 2017*. Kiel: German Acoustical Society (DEGA), 2017, pp. 872–875.

[5] Y. Tang, C. Arnold, and T. J. Cox, "A Study on the Relationship between the Intelligibility and Quality of Algorithmically-Modified Speech for Normal Hearing Listeners," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, vol. 1, no. 1, p. 10, Jun. 2018.

[6] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech & Language*, vol. 28, no. 4, pp. 858–872, Jul. 2014.

[7] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, May 2013.

[8] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[9] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. N. Petkov, B. Sauert, and P. Vary, "Optimizing Speech Intelligibility in a Noisy Environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, Mar. 2015.

[10] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Interspeech*, Lyon, France, 2013, pp. 3552–3556.

[11] J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-Enhancing Speech Modifications — The Hurricane Challenge 2.0," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1341–1345.

[12] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 277–282, Aug. 1976.

[13] J. L. Hall and J. L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 280–285, Jan. 2010.

[14] K. Nathwani, F. Hafiz, A. Swain, and R. Biswas, "Speech Intelligibility Enhancement using an Optimal Formant Shifting Approach," in *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Sep. 2021, pp. 120–125.

[15] C. Tantibundhit, J. R. Boston, C. C. Li, J. D. Durrant, S. Shaiman, K. Kovacyk, and A. El-Jaroudi, "New signal decomposition method based speech enhancement," *Signal Processing*, vol. 87, no. 11, pp. 2607–2628, Nov. 2007.

[16] E. Jokinen, U. Remes, P. Alku, E. Jokinen, U. Remes, and P. Alku, "Intelligibility Enhancement of Telephone Speech Using Gaussian Process Regression for Normal-to-Lombard Spectral Tilt Conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, pp. 1985–1996, Oct. 2017.

[17] R. Zhang, R. Hu, G. Li, and X. Wang, "Spectral Tilt Estimation for Speech Intelligibility Enhancement Using RNN Based on All-Pole Model," in *MultiMedia Modeling*, ser. Lecture Notes in Computer Science, I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, and S. Vrochidis, Eds. Springer International Publishing, 2018, pp. 144–156.

[18] G. Li, H. Ruimin, R. Zhang, and X. Wang, "A mapping model of spectral tilt in normal-to-Lombard speech conversion for intelligibility enhancement," *Multimedia Tools and Applications*, vol. 79, no. 27-28, pp. 19 471–19 491, Jul. 2020.

[19] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise Intelligibility Improvement Based on Power Recovery and Dynamic Range Compression," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Portland, USA, Aug. 2012, pp. 2075–2079.

[20] C. Chermaz and S. King, "A Sound Engineering Approach to Near End Listening Enhancement," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1356–1360.

[21] B. Sauert and P. Vary, "Near End Listening Enhancement: Speech Intelligibility Improvement in Noisy Environments," in *2006 IEEE International*

Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1, May 2006, pp. 493–496.

[22] B. Sauert, G. Enzner, and P. Vary, "Near End Listening Enhancement With Strict Loudspeaker Output Power Constraining," in *Proc. of Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, Sep. 2006, p. 4.

[23] M. Niermann, P. Jax, and P. Vary, "Near-end listening enhancement by noise-inverse speech shaping," in *2016 24th European Signal Processing Conference (EUSIPCO)*, Aug. 2016, pp. 2390–2394.

[24] M. Niermann and P. Vary, "Listening Enhancement in Noisy Environments: Solutions in Time and Frequency Domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 699–709, 2021.

[25] American National Standards Institute, *Methods for Calculation of the Speech Intelligibility Index*. New York, N.Y: Acoustical Society of America, 2017, vol. ANSI S.35-1997.

[26] B. Sauert and P. Vary, "Recursive Closed-Form Optimization of Spectral Audio Power Allocation for Near End Listening Enhancement," *ITG-Fachtagung Sprachkommunikation*, vol. Paper 8, p. 4, Oct. 2010.

[27] C. H. Taal, J. Jensen, and A. Leijon, "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[28] H. Schepker, J. Rennies, and S. Doclo, "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 2692–2706, Nov. 2015.

[29] R. C. Hendriks, J. B. Crespo, J. Jensen, and C. H. Taal, "Optimal Near-End Speech Intelligibility Improvement Incorporating Additive Noise and Late Reverberation Under an Approximation of the Short-Time SII," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 851–862, May 2015.

[30] M. Niermann, P. Jax, and P. Vary, "Joint Near-End Listening Enhancement and far-end noise reduction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 4970–4974.

[31] B. Sauert, *Near-End Listening Enhancement: Theory and Application*, 1st ed., ser. Aachener Beiträge zu digitalen Nachrichtensystemen. Aachen: Wissenschaftsverlag Mainz, 2014, no. 36, oCLC: 880393716.

[32] A. J. Fuglsig, J. Østergaard, J. Jensen, L. S. Bertelsen, P. Mariager, and Z.-H. Tan, "Joint Far- and Near-End Speech Intelligibility Enhancement Based on the Approximated Speech Intelligibility Index," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7752–7756.

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

[34] H. Li, S.-W. Fu, Y. Tsao, and J. Yamagishi, "iMetricGAN: Intelligibility Enhancement for Speech-in-Noise Using Generative Adversarial Network-Based Metric Learning," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1336–1340.

[35] H. Li and J. Yamagishi, "Multi-Metric Optimization Using Generative Adversarial Networks for Near-End Speech Intelligibility Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3000–3011, 2021.

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

[37] S. Kuo and D. Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87, no. 6, pp. 943–973, Jun. 1999.

[38] N. V. George and G. Panda, "Advances in active noise control: A survey, with emphasis on recent nonlinear techniques," *Signal Processing*, vol. 93, no. 2, pp. 363–377, Feb. 2013.

[39] G. Li, R. Hu, X. Wang, and R. Zhang, "A near-end listening enhancement system by RNN-based noise cancellation and speech modification," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 483–15 505, Jun. 2019.

[40] F. Cheng, X. Wang, L. Gang, W. Tu, and J. Wang, "Speech Intelligibility Enhancement in Strong Mechanical Noise Based on Neural Networks," in *Advances in Multimedia Information Processing – PCM 2017*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 702–712.

[41] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, Feb. 2006.

[42] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to Speech Intelligibility Index," in *2009 17th European Signal Processing Conference*, Aug. 2009, pp. 1844–1848.

[43] A. Zahedi, M. S. Pedersen, J. Østergaard, T. U. Christiansen, L. Bramsløw, and J. Jensen, "Minimum Processing Beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2710–2724, 2021.

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

[45] 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," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

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

[47] ——, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *J. Audio Eng. Soc*, vol. 62, no. 3, pp. 99–117, 2014.

[48] ITU-T, "Recommendation P.862: Perceptual evaluation of speech quality (PESQ)," International Telecommunication Unition, Recommendation ITU-T P.862, Feb. 2001.

[49] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Intelligibility Enhancement Based on Mutual Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1694–1708, Aug. 2017.

[50] ITU-T, "Recommendation P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunication Unition, Recommendation ITU-T P.862.2, Nov. 2007.

[51] J. H. L. Hansen and B. L. Pellom, "An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms," in *The International Conference on Speech and Language Processing*, 1998, pp. 2819–2822.

[52] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

[53] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A Perceptual Model for Sinusoidal Audio Coding Based on Spectral Integration," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1292–1304, Jun. 2005.

[54] N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, "Web Audio Evaluation Tool: A browser-based listening test environment," Maynooth, Ireland, Jul. 2015, pp. 147–152.

[55] ITU-R, "Recommendation BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level," International Telecommunication Unition, Recommendation ITU-R BS.1770-4, Oct. 2015.

[56] EBU, "EBU recommendation R128 - Loudness Normalisation and Permitted Maximum Level of Audio Signals," European Broadcasting Union, Recommendation R-128, 2014.

[57] B. D. Man, "Web Audio Evaluation Tool," May 2022, version committed on 26-12-2021.

[58] ITU-R, "Recommendation BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Unition, Recommendation ITU-R BS.1534-3, Oct. 2015.

Paper C

Joint Minimum Processing Beamforming and Near-End Listening Enhancement

Andreas Jonas Fuglsig, Jesper Jensen, Zheng-Hua Tan, Lars Søndergaard Bertelsen, Jens Christian Lindof and Jan Østergaard

The paper has been published in the
2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pp. 485–489, 2024.

The layout has been revised.

Abstract

We consider speech enhancement for signals picked up in one noisy environment that must be rendered to a listener in another noisy environment. For both far-end noise reduction and near-end listening enhancement, it has been shown that excessive focus on noise suppression or intelligibility maximization may lead to excessive speech distortions and quality degradations in favorable noise conditions, where intelligibility is already at ceiling level. Recently [1, 2] propose to remedy this with a minimum processing framework that either reduces noise or enhances listening a minimum amount given that a certain intelligibility criterion is still satisfied. Additionally, it has been shown that joint consideration of both environments improves speech enhancement performance. In this paper, we formulate a joint far- and near-end minimum processing framework, that improves intelligibility while limiting speech distortions in favorable noise conditions. We provide closed-form solutions to specific boundary scenarios and investigate performance for the general case using numerical optimization. We also show concatenating existing minimum processing far- and near-end enhancement methods preserves the effects of the initial methods. Results show that the joint optimization can further improve performance compared to the concatenated approach.

1 Introduction

Speech communication systems, including, e.g., mobile phones, hearing aids, and intercom systems, need to work in a variety of often noisy situations which can degrade intelligibility and quality.

In speech communication systems, we may consider two distinct environments, cf. Fig. C.1: The far-end (the target talker location) and the near-end (the listener's location). Both environments may be susceptible to noise affecting the Speech Quality (SQ) and Intelligibility (SI) for the listener. To counter this, speech enhancement can be applied at both ends. Far-end Speech Enhancement (FSE) may employ single or multiple microphone noise reduction methods [1, 3–5]. Near-end Listening Enhancement (NLE) [2, 6, 7] leverage knowledge of the near-end noise to pre-process the received FSE signal for an optimal presentation with enhanced SI in the near-end background noise. We note that headphone listening can utilize adaptive noise control (ANC) methods [8]. However, ANC with classic adaptive filtering falls short outside headphone use [9]. Thus, ANC is beyond our scope as we address speech presentation in an open environment.

NLE algorithms have conventionally aimed to solely enhance SI, which may be beneficial at low SNRs but might diminish SQ at high SNRs due to excessive processing [7, 10–12]. Furthermore, many FSE methods are designed with a rationale targeting eliminating all background noise to retain only clean speech,

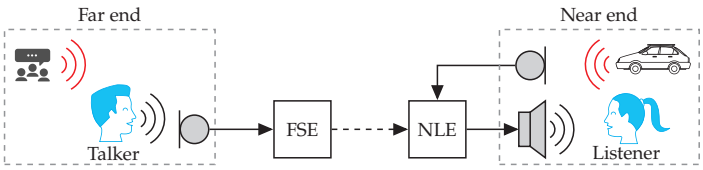


Fig. C.1: Basic communication system with Far-end Speech Enhancement (FSE) and Near-end Listening Enhancement(NLE).

causing potential excessive speech distortion or loss of contextual noise [1]. Therefore, the works of [1] and [2] apply a *minimum processing principle* to FSE and NLE, respectively, where the noisy signal [1] or the signal received from the far-end [2] is modified as little as possible while obtaining a desired level of SI. However, so far, the minimum processing principle has not been applied to situations where noise is present in both far-end and near-end environments simultaneously. In fact, until recently, addressing disturbances in both the far-end and near-end settings was approached as separate tasks [4–6]. However, recent work in [13–18] have shown that optimizing SI by jointly addressing the noise in both environments is more effective than handling them as separate disjoint problems.

In this paper, we formulate a joint far- and near-end minimum processing framework, which contrary to existing joint works only modifies the signal the minimum amount required to achieve a desired level of SI, and preserves SQ in favorable noise condition. Furthermore, it expands upon the existing minimum processing frameworks [1, 2] by jointly considering the effects of FSE and NLE for both far- and near-end noise simultaneously. Following [1, 2] we minimize a mean-square error (MSE) processing penalty subject to an estimated SI constraint in terms of the Approximated Speech Intelligibility Index (ASII) [19]. We derive closed-form solutions for interesting special cases of the problem, and solve the general case using numerical optimization. We perform an experimental evaluation where we compare the proposed approach to a concatenation of minimum processing FSE [1] and minimum processing NLE [2]. The results show, that the proposed method is able to greatly improve SI up until a desired level in noisy conditions while also limiting speech distortions in favorable noise conditions. In addition, we see that concatenation preserves the minimum processing abilities of the individual methods while being able to improve both SI and SQ in various noise conditions. Finally, we show the joint approach is able to further improve performance compared to the concatenation.

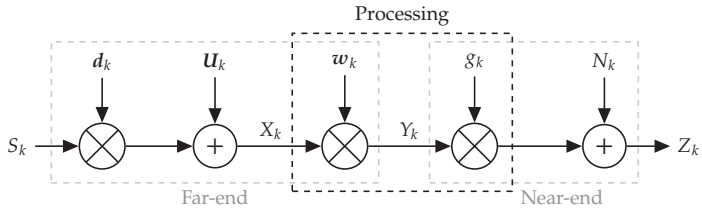


Fig. C.2: Signal model.

2 Signal Model

We consider a time-frequency domain representation of speech and noise signals with frequency index k. Since the statistics of the signals can be estimated online, and the mathematical framework can be applied on a per time-frame basis, we disregard the time index and assume we are considering a particular time frame, unless otherwise is stated. The signal model in frequency bin k, cf. Fig. C.2, is given by

$$\boldsymbol{X}_k = \boldsymbol{d}_k S_k + \boldsymbol{U}_k, \quad Y_k = \boldsymbol{w}_k^H \boldsymbol{X}_k, \quad Z_k = g_k Y_k + N_k, \tag{C.1}$$

where $\boldsymbol{X}_k \in \mathbb{C}^M$ is the noisy multi-microphone signal, S_k the clean speech signal recorded at the source location, $\boldsymbol{d}_k \in \mathbb{C}^M$ are acoustic transfer functions from the source to the microphones, and $\boldsymbol{U}_k \in \mathbb{C}^M$ is additive far-end noise, and M is the number of microphones. To increase SI and SQ, the noisy signal, $\boldsymbol{X}_k$, is linearly and spatially enhanced via a FSE noise reducing beamformer, $\boldsymbol{w}_k \in \mathbb{C}^M$, producing the modified signal Y_k. To further increase SI and SQ a NLE gain, $g_k \in \mathbb{R}_+$, is applied prior to playout. Finally, the signal, Z_k, received at the near-end, is contaminated by ambient noise, N_k, in the environment. We assume the speech and noise processes are uncorrelated and zero-mean random processes, which are independent across frequency [4]. We then have the speech distortion weighted covariance matrix $C_{X_k}^{(\mu)}$ of $\boldsymbol{X}_k$ is [1],

$$C_{X_k}^{(\mu)} \triangleq C_{S_k} + \mu C_{U_k} = \sigma_{S_k}^2 \boldsymbol{d}_k \boldsymbol{d}_k^H + \mu C_{U_k}, \tag{C.2}$$

where $\sigma_{S_k}^2$ is the clean speech power spectrum level in time-frequency bin k and C_{U_k} is the far-end noise covariance matrix of $\boldsymbol{U}_k$ and where $\mu \in \mathbb{R}_+$ is the speech distortion weight [3, 20].

In this paper, we process signals in perceptually relevant critical bands [21], with an individual non-negative filter weight, $\omega_{j,k}$, for each frequency bin-subband pair, where subbands are indexed by j and frequencies with index k. We let $\mathbb{B}_j$ denote the set of frequencies, k, that contribute to the j'th subband.

3 Minimum Processing Concept

To increase SI and SQ, the aim is to jointly determine a FSE beamformer, $\boldsymbol{w}_k$, for far-end noise reduction and a NLE gain, g_k, for pre-processing the signal before playout in near-end background noise.

Assume, as in [1], we are given a target reference signal, S_k^R, which may be the output of a beamformer with some desired properties, e.g., low speech distortion. Then for a particular subband, j, stack all S_k^R, S_k and Z_k for $k \in \mathbb{B}_j$ into vectors $\boldsymbol{S}_j^R$, $\boldsymbol{S}_j$ and $\boldsymbol{Z}_j$ [1, 2]. Additionally, let $\mathcal{D}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Z}_j\right)$ be a non-negative distortion measure (processing penalty) between the target reference signal, S_k^R, and the signal presented to the near-end listener, Z_k, and let $\mathcal{I}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Z}_j\right)$ be a finite non-negative SI estimator of NLE-processed speech, $\boldsymbol{Z}_j$ in subband j. Then, the joint far- and near-end minimum processing beamformer, $\boldsymbol{w}_k^{MP}$, and NLE gain, g_k^{MP}, in subband j are defined as the solution to the following optimization problem:

$$\underset{\{\boldsymbol{w}_k\},\{g_k\},k\in\mathbb{B}_j}{\arg\min} \quad \mathcal{D}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Z}_j\right) \quad \text{s.t.} \quad \mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right) \geq I_j'. \tag{C.3}$$

Here we consider the combined effects of all noise sources with far-end noise reduction and near-end listening enhancement simultaneously. This is contrary to [1] that only considers far-end noise reduction, and in a similar manner [2] that is only concerned with near-end listening enhancement under the assumptions of a clean far-end. Thus, instead of taking a classic blind concatenated approach, where we solve the two versions of (C.3) proposed in [1] and [2] in succession, while they are unaware of each other and the processing they apply. In our proposed joint approach, we solve (C.3) directly, such that all noise sources and processing steps of $\boldsymbol{w}_k$ and g_k are jointly taken into account at the same time.

4 Joint Minimum Processing

To avoid comb filtering effects and inspired by the results of [1] and [13], we propose the following parameterized multichannel noise reduction vector (beamformer), that is fixed across an entire subband,

$$\boldsymbol{w}_{j,k} \triangleq \alpha_j \boldsymbol{w}_k^{\mu_R} + (1-\alpha_j)\boldsymbol{w}_k^{\mu_0}, \tag{C.4}$$

as a solution to (C.3). Here the parameter $\alpha_j \in [0,1]$, and $\boldsymbol{w}_k^{\mu_R}$ and $\boldsymbol{w}_k^{\mu_0}$ are speech distortion weighted Multichannel Wiener Filters (MWFs) [1, 3]

$$\boldsymbol{w}_k^{\mu} \triangleq \left(C_{X_k}^{(\mu)}\right)^{-1} \sigma_{S_k}^2 \boldsymbol{d}_k, \tag{C.5}$$

with pre-selected speech distortion weights, μ_R and μ_0, such that the reference beamformer $\boldsymbol{w}_k^{\mu_R}$ has low speech distortion and $\boldsymbol{w}_k^{\mu_0}$ has high noise reduction [1].

Similarly to avoid comb filtering by the NLE gains, g_k, we assume they are fixed across an entire subband, i.e.,

$$g_k = g_i, \quad \forall k, i \in \mathbb{B}_j. \tag{C.6}$$

This is also in line with results of existing NLE literature [6, 13–15].

4.1 Processing Penalty

For the processing penalty, $\mathcal{D}_j(\cdot)$ we consider an MSE criterion [1, 2]. Since we want to have low speech distortion, we consider the reference signal, S_k^R, to be the output of the reference MWF, $\boldsymbol{w}_k^{\mu_R}$, which was chosen above to have the property of low distortion. Therefore, the minimum processing solution to (C.3), i.e., $\boldsymbol{w}_{j,k}$ and g_k, should minimize the distance to $\boldsymbol{w}_k^{\mu_R}$. That is, the processing penalty must punish excessive difference to the reference signal caused by both the beamforming and NLE post gain. We note, that an obvious way to increase the near-end output SNR is to increase g_k to infinity. However, this would lead to excessive speech distortions, infinite playback volume, and most importantly increase the difference to the reference signal leading to a violation of the minimum processing concept. Hence, we propose the following processing penalty,

$$\mathcal{D}_j(\boldsymbol{S}_j^R, \boldsymbol{Z}_j) = \left(1 - \alpha_j\right)^2 + \left(1 - g_j\right)^2. \tag{C.7}$$

Here the first term is the processing penalty incurred by the beamformer and pushes $\boldsymbol{w}_{j,k}$ close to $\boldsymbol{w}_k^{\mu_R}$. The second term is the penalty incurred by the NLE gain and pushes $g_k \boldsymbol{w}_{j,k}$ close to $\boldsymbol{w}_k^{\mu_R}$ and limits any speech distortions and excessive playback volume caused by the NLE gain.

4.2 Performance Criteria

We consider two different performance criteria; an intelligibility performance criterion and a new noise power criterion.

4.2.1 Intelligibility criterion

We consider a performance criterion based on the ASII [19] as in [2] whereas [1] uses SII. Letting I_j be a given minimum requirement on the ASII subband SI performance [19], the SI constraint in terms of the subband SNR, ξ_j, is [2,

App. C]

$$\xi_j \triangleq \frac{g_j^2 \delta_{S_j}(\alpha_j)}{g_j^2 \delta_{U_j}(\alpha_j) + \sigma^2_{\mathcal{N}_j}}, \quad \xi_j \geq \frac{I_j}{1 - I_j} \triangleq I_j^{\xi}, \tag{C.8}$$

where we consider $I_j^{\xi} \triangleq \frac{I_j}{1-I_j}$ as a target SNR, and $\delta_{S_j}(\alpha_j)$ and $\delta_{U_j}(\alpha_j)$ denote the processed speech and far-end noise power within one subband, j, for a given α_j, respectively. By evaluating $\boldsymbol{w}_{j,k}^H C_{U_k} \boldsymbol{w}_{j,k}$ and filtering into subbands we have

$$\delta_{U_j}(\alpha_j) = \alpha_j^2 \delta_{U_j}^{\mu_R} + (1-\alpha_j)^2 \delta_{U_j}^{\mu_0} + \alpha_j(1-\alpha_j)\delta_{U_j}^{cross}, \tag{C.9}$$

$$\delta_{U_j}^{\mu} \triangleq \sum_{k \in \mathbb{B}_j} \omega_{j,k} \left(\boldsymbol{w}_k^{\mu}\right)^H C_{U_k} \boldsymbol{w}_k^{\mu}, \tag{C.10}$$

$$\delta_{U_j}^{cross} \triangleq \sum_{k \in \mathbb{B}_j} \omega_{j,k} 2\Re \left\{ \left(\boldsymbol{w}_k^{\mu_0}\right)^H C_{U_k} \boldsymbol{w}_k^{\mu_R} \right\}. \tag{C.11}$$

A similar definition applies to the processed speech subband power, $\delta_{S_j}(\alpha_j)$. Since we modify the speech to increase SI, the subband SNR, ξ_j, is defined as the ratio of the *processed* speech subband power to the total *processed* noise power at the near-end listener [2, 19, 21]. This is different to [1], where SNR is clean speech power relative to the MSE between S and Y, i.e, all processing to the original speech is considered as a noise term and does not include near-end noise, N. In this work, the far-end SNR is defined as $\delta_{S_j}(\alpha_j)/\delta_{U_j}(\alpha_j)$. Now by defining the polynomial $p_{FSE}(\alpha_j)$, representing FSE SNR performance, as

$$p_{FSE}(\alpha_j) \triangleq \delta_{S_j}(\alpha_j) - \delta_{U_j}(\alpha_j) I_j^{\xi} \tag{C.12}$$

$$= \alpha_j^2 D_j^{\mu_R} + (1-\alpha_j)^2 D_j^{\mu_0} + \alpha_j(1-\alpha_j) D_j^{cross} \tag{C.13}$$

where $D_j^{\mu} \triangleq \delta_{S_j}^{\mu} - \delta_{U_j}^{\mu} I_j^{\xi}$, and $D_j^{cross} \triangleq \delta_{S_j}^{cross} - \delta_{U_j}^{cross} I_j^{\xi}$. We can then write the constraint as

$$\mathcal{I}_j = g_j^2 p_{FSE}(\alpha_j) \geq \sigma^2_{\mathcal{N}_j} I_j^{\xi}. \tag{C.14}$$

4.2.2 Noise power criterion

Since we consider far-end and near-end noise jointly, we have more knowledge about the processing and noise situation than in [1] and [2]. Therefore, looking at (C.8), we see that to increase the SNR and satisfy the audibility constraint, the processed far-end noise might need to overpower the near-end noise. However, depending on the noise powers this increase in SI may come at an undesired loss in SQ due to increased total noise levels. Therefore, to

limit distortions caused by excessive noise levels, in the new joint approach we impose a constraint on the processed far-end noise power,

$$10\log_{10}\left(g_j^2\delta_{U_j}(\alpha_j)\right) \leq 10\log_{10}\sigma^2_{\mathcal{N}_j} + \Delta_{U_j}, \tag{C.15}$$

where the parameter Δ_{U_j} controls how many dB the processed far-end noise can deviate from the near-end noise in subband, j.

4.3 Optimization Problem and Boundary Solutions

From the above derivations we have that the joint far- and near-end minimum processing speech enhancement problem (C.3) with the MSE processing penalty (C.7), ASII performance constraint (C.14) and noise power constraint (C.15) is

$$\begin{aligned} &\underset{\alpha_j, g_j \in \mathbb{R}_+}{\arg\min} \quad \left(1-\alpha_j\right)^2 + \left(1-g_j\right)^2 && (P_0)\\ \text{s.t. } &C_1 : g_j^2 p_{FSE}(\alpha_j) \geq \sigma^2_{\mathcal{N}_j} I_j^{\xi}, && C_3 : 0 \leq \alpha_j \leq 1,\\ &C_2 : g_j^2\delta_{U_j}(\alpha_j) \leq \sigma^2_{\mathcal{N}_j} 10^{\Delta_{U_j}/10}, && C_4 : 1 \leq g_j. \end{aligned}$$

We can solve this optimization problem using a grid search algorithm. Given the optimal solution (α_j^*, g_j^{MP}), the optimum minimum processing beamformer is then given as

$$\boldsymbol{w}_{j,k}^{MP} = \alpha_j^* \boldsymbol{w}_k^{\mu_R} + (1-\alpha_j^*)\boldsymbol{w}_k^{\mu_0}. \tag{C.16}$$

From (C.9) and (C.13), we see that both the processed far-end noise power, δ_{U_j}, and processed far-end SNR performance, p_{FSE}, include terms from each beamformer and a crossover term, and that the parameter α provides a trade off between the SNR/processing possible by the two candidate beamformers. Furthermore, we see that for $\alpha_j = 0$ or $\alpha_j = 1$, then $\boldsymbol{w}_{j,k} = \boldsymbol{w}_k^{\mu_R}$ or $\boldsymbol{w}_{j,k} = \boldsymbol{w}_k^{\mu_0}$, respectively, and thus the crossover terms vanish as well as the term accounting for the other beamformer. From this and inspection of the constraints we have the following lemma showing conditions for feasible boundary solutions. Proof omitted due to space limitations.

Lemma 1 *The beamformer combination weight, $\alpha_j^* = 1$ is a solution to (P_0) under one of the two following conditions: (i) If $D_j^{\mu_R} \geq \sigma^2_{\mathcal{N}_j} I_j^{\xi}$ and $\delta_{U_j}^{\mu_R} \leq \sigma^2_{\mathcal{N}_j} 10^{\Delta_{U_j}/10}$, with optimal NLE gain $g_j^* = 1$. (ii) If $D_j^{\mu_R} \in (0, \sigma^2_{\mathcal{N}_j} I_j^{\xi})$ and $g_j^2\delta_{U_j}^{\mu_R} \leq \sigma^2_{\mathcal{N}_j} 10^{\Delta_{U_j}/10}$, where the NLE gain is $g_j^2 = \sigma^2_{\mathcal{N}_j} I_j^{\xi}/D_j^{\mu_R}$.*

The beamformer combination weight, $\alpha_j^* = 0$ is a solution to (P_0) under one of the two following conditions: (i) If $D_j^{\mu_0} \geq \sigma^2_{\mathcal{N}_j} I_j^{\xi}$ and $\delta_{U_j}^{\mu_0} \leq \sigma^2_{\mathcal{N}_j} 10^{\Delta_{U_j}/10}$, with NLE gain

$g_j^* = 1$. *(ii) If* $D_j^{\mu_0} \in (0, \sigma_{\mathcal{N}_j}^2 I_j^{\xi})$ *and* $g_j^2 \delta_{\mathcal{U}_j}^{\mu_0} \leq \sigma_{\mathcal{N}_j}^2 10^{\Delta_{U_j}/10}$, *where the NLE gain is* $g_j^2 = \sigma_{\mathcal{N}_j}^2 I_j^{\xi} / D_j^{\mu_0}$.

Depending on the subband definition, multiple frequencies may contribute to multiple subbands indexed by $\mathbb{F}_k$. Therefore, the optimum beamformer $\boldsymbol{w}_{j,k}^{MP}$ and NLE gain g_j^{MP} may also contribute to multiple subbands. Letting $\eta_{j,k}$ denote the weight that accounts for the impact of this contribution, the optimal frequency dependent beamformer and NLE gain, respectively, are

$$\boldsymbol{w}_k^{MP} = \sum_{j \in \mathbb{F}_k} \eta_{j,k} \boldsymbol{w}_{j,k}^{MP} \quad \text{and} \quad g_k^{MP} = \sum_{j \in \mathbb{F}_k} \eta_{j,k} g_j^{MP}. \tag{C.17}$$

Depending on the subband noise powers, the constraints of (P_0) may be infeasible. For example, C_1 is infeasible if the far-end noise cannot be sufficiently reduced to produce a feasible far-end SNR. Similarly, C_2 is infeasible if the remaining far-end noise power is too high compared to the near-end noise power. We propose three ways to handle the infeasible situations.

If C_1 is infeasible: First find an α_j^* that maximizes the far-end SNR, $\delta_{S_j}(\alpha)/\delta_{U_j}(\alpha)$. Then to increase SI as function of g_j for a fixed α_j^*, select the NLE gain, g_j^*, such that the near-end noise does not decrease the SNR coming from the far-end more than $\Delta_{N_j} > 0$ dB. The gain, g_j^* is then clipped according to C_2 and C_4.

If C_2 is infeasible: Let $g_j^* = 1$ and find an α_j^* so as the near-end SNR, ξ_j, is close to I_j^{ξ} to approach satisfying both C_1 and C_2.

If the intersected constraints $C_1 \cap C_2$ are infeasible: Find an α_j^* such that the processed near-end SNR is close to I_j^{ξ} while adhering to C_2. Then select the NLE gain, g_j^*, such that the processed near-end SNR is maximized while minimizing g_j and satisfying C_2.

5 Experimental Evaluation

We compare performance between the proposed joint minimum processing method and the concatenation of the FSE [1] and NLE [2]. We investigate two scenarios: (i) When the target talker is in a babble noise setting, e.g., office environment, and the listener is driving in a car, and (ii) the reverse scenario where the talker is in car noise and the listener is in babble noise. The FSE beamformer of [1] is also parameterized according to two μ-MWF beamformers, and these beamformers are selected to be the same as in the proposed method, where we have $\mu_R = 0$, $\mu_0 = 5$. The per-band audibility target input, I_j, is weighted from a total SII target, $A^* = 0.7$, using the band importance functions of the SII [21], cf. [2, Sec. IV.B], in both the proposed and reference method. Similarly, for the proposed method, the parameters, Δ_{U_j}

and Δ_{N_j} are weighted for each subband from a single value of Δ_U, and Δ_N, respectively. Through informal listening tests we have selected $\Delta_U = 12\,\text{dB}$, and $\Delta_N = 10\,\text{dB}$ for scenario (i), and $\Delta_U = 0\,\text{dB}$, and $\Delta_N = 10\,\text{dB}$ for scenario (ii).

5.1 Experimental Setup

The far-end room dimensions are $3 \times 4 \times 3\,\text{m}^3$, with the target talker located at $[1.50, 3.00, 1]\,\text{m}$, and three noise sources located at $[0.50, 1.00, 1]\,\text{m}$, $[0.75, 3.00, 1]\,\text{m}$ and $[3.00, 1.60, 1]\,\text{m}$. The far-end has two microphones at $[1.50, 2.00, 1]\,\text{m}$ and $[1.50, 2.02, 1]\,\text{m}$. Each microphone is also subject to a 60 dB SNR white noise. The time-frequency representations of the speech and noise signals are based on a DFT with 32 ms windows with 50% overlap. We consider a total of $J = 30$ critical bands with center frequencies linearly spaced on the equivalent rectangular bandwidth scale from 150 Hz to 8000 Hz derived according to [22]. For simplicity, signals are processed in a time-invariant manner and power spectrums are evaluated as the long-term average across time-frames. The long-term power spectrums of the speech and noise are assumed to be known along with the room transfer functions, that are generated without reverberation using [23]. The speech material is sentences from the TIMIT [24] test set sampled at 16 kHz. Performance is evaluated across a total of 10 trials, where for each trial a speaker is selected randomly without replacement, and a random sentence is selected for the given speaker. We then average the performance across the trials for each combination of noise, SNR and enhancement method.

5.2 Results

Estimated SI and SQ performance is measured with ESTOI [25] and PESQ [26]. Table C.1 shows scores for the proposed and concatenated method alongside the unprocessed performance, with the best ESTOI and PESQ scores highlighted for each SNR and noise pair.

The results indicate that the proposed joint method and the concatenation method generally exhibit similar performance, as expected due to their overall similarity. However, in severe noise with low SNRs, where the unprocessed performance is very low, the proposed joint method overall outperforms the blind concatenation in ESTOI. As the SNRs increase, the unprocessed SQ and SI score rise naturally. Here, when the noise situation is more favorable, both methods are able to utilize their minimum processing designs and limit distortions to better preserve the natural SQ and SI, as seen by how the ESTOI performance is close to the high unprocessed scores, while the PESQ scores still improve or stay close to the unprocessed scores. Since both the proposed joint method and the individual steps in the concatenation of [1] and [2] are

Noise		SNR		ESTOI			PESQ		
FE	NE	FE	NE	Prop.	Blind	Unp.	Prop.	Blind	Unp.
B	C	0 dB	−25 dB	**0.482**	0.419	0.380	**1.052**	1.030	1.028
B	C	0 dB	0 dB	0.569	0.512	**0.629**	1.292	1.282	**1.320**
B	C	0 dB	15 dB	0.618	0.604	**0.680**	1.400	**1.465**	1.389
B	C	−10 dB	−30 dB	**0.262**	0.231	0.214	**1.044**	1.030	1.040
B	C	0 dB	−30 dB	**0.478**	0.416	0.314	**1.050**	1.030	1.024
B	C	10 dB	−30 dB	**0.529**	0.506	0.343	**1.034**	1.031	1.024
C	B	−10 dB	−20 dB	**0.533**	0.500	0.036	1.286	**1.317**	1.078
C	B	−10 dB	0 dB	**0.542**	0.511	0.460	1.284	**1.318**	1.147
C	B	−10 dB	20 dB	0.684	0.679	**0.748**	1.832	**2.012**	1.542
C	B	−20 dB	−10 dB	**0.470**	0.439	0.170	1.254	**1.291**	1.109
C	B	0 dB	−10 dB	**0.544**	0.535	0.194	**1.326**	1.325	1.128
C	B	20 dB	−10 dB	**0.544**	**0.544**	0.195	**1.326**	**1.326**	1.128

Table C.1: ESTOI and PESQ scores for the proposed joint method and the concatenation of [1] and [2] for various SNRs along with the unprocessed performance. Here B is babble noise and C is car noise.

designed with minimum processing in mind, we did not expect a big difference in SQ performance at high SNRs.

We also see, that the concatenation of [1] and [2] preserves the effects of the individual methods, i.e., the signal is only processed the minimum required amount to obtain a desired SI at the far- and near-end respectively, and preserves SQ in favorable noise conditions.

For far-end car noise we observe, that the blind method is able to increase PESQ slightly more than the proposed method, we expect this is caused by the slight variations between the constraints in the two methods. Hence, further benefits might be gained from adjusting the proposed method accordingly. However, because the concatenation is blind the FSE beamformer [1] may not remove a sufficient amount of noise for the NLE in [2] to be able to achieve the desired SI. Similarly, because the NLE in [2] is blind to noise coming from the far-end it might not provide a sufficiently high gain as it mistakes noise for speech. On the other hand, the proposed joint method can achieve a higher SI performance because it has access to all noise and processing information simultaneously.

6 Conclusion

We formulated a joint far- and near-end minimum processing framework, where the beamformed and near-end listening enhanced output signal is optimized to have the minimum amount of processing artifacts with the constraint that an intelligibility performance criterion is satisfied. The proposed method adapts to environmental noise conditions and focuses on improving intelligibility in very noisy conditions, and, by the minimum processing concept, quality in favorable noise conditions. We show closed-form solutions to interesting special cases of the optimization problem. Additionally, we show that speech enhancement using a blind concatenation of the existing far- and near-end minimum processing frameworks [1] and [2] preserves the minimum processing abilities of the individual methods, and that the concatenation is also able to improve both intelligibility and quality in various noise conditions. Results also show that the proposed joint method outperforms the simple blind concatenation in terms of intelligibility enhancement because the proposed method considers all noise sources and processing steps simultaneously.

References

[1] A. Zahedi, M. S. Pedersen, J. Østergaard, T. U. Christiansen, L. Bramsløw, and J. Jensen, "Minimum Processing Beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2710–2724, 2021.

[2] A. J. Fuglsig, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof, and J. Østergaard, "Minimum Processing Near-End Listening Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2233–2245, 2023.

[3] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic Beamforming for Hearing Aid Applications," in *Handbook on Array Processing and Sensor Networks*. John Wiley & Sons, Ltd, 2010, pp. 269–302.

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

[5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[6] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. N. Petkov, B. Sauert, and P. Vary, "Optimizing Speech Intelligibility in a Noisy Environment: A

unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, Mar. 2015.

[7] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[8] N. V. George and G. Panda, "Advances in active noise control: A survey, with emphasis on recent nonlinear techniques," *Signal Processing*, vol. 93, no. 2, pp. 363–377, Feb. 2013.

[9] G. Li, R. Hu, X. Wang, and R. Zhang, "A near-end listening enhancement system by RNN-based noise cancellation and speech modification," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 483–15 505, Jun. 2019.

[10] J. Rennies, A. Pusch, H. Schepker, and S. Doclo, "Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements," *The Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. EL315–EL321, Oct. 2018.

[11] R. Pricken, M. Wältermann, E. Parotat, M. Soloducha, and A. Raake, "Quality Aspects of Near-End Listening Enhancement Approaches in Telecommunication Applications," in *Proceedings of DAGA 2017*. Kiel: German Acoustical Society (DEGA), 2017, pp. 872–875.

[12] Y. Tang, C. Arnold, and T. J. Cox, "A Study on the Relationship between the Intelligibility and Quality of Algorithmically-Modified Speech for Normal Hearing Listeners," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, vol. 1, no. 1, p. 10, Jun. 2018.

[13] M. Niermann, P. Jax, and P. Vary, "Joint Near-End Listening Enhancement and far-end noise reduction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 4970–4974.

[14] A. J. Fuglsig, J. Østergaard, J. Jensen, L. S. Bertelsen, P. Mariager, and Z.-H. Tan, "Joint Far- and Near-End Speech Intelligibility Enhancement Based on the Approximated Speech Intelligibility Index," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7752–7756.

[15] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Intelligibility Enhancement Based on Mutual Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1694–1708, Aug. 2017.

[16] H. Li, Y. Liu, and J. Yamagishi, "Joint Noise Reduction and Listening Enhancement for Full-End Speech Enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2023, pp. 1–5.

[17] M. P. Shifas, C. Zorilă, and Y. Stylianou, "End-to-End Neural Based Modification of Noisy Speech for Speech-in-Noise Intelligibility Improvement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 162–173, 2022.

[18] T.-C. Zorilă and Y. Stylianou, "On the Quality and Intelligibility of Noisy Speech Processed for Near-End Listening Enhancement," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 2023–2027.

[19] C. H. Taal, J. Jensen, and A. Leijon, "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[20] M. Brandstein and D. Ward, Eds., *Microphone arrays: signal processing techniques and applications*, ser. Digital signal processing. New York: Springer, 2001.

[21] American National Standards Institute, *Methods for Calculation of the Speech Intelligibility Index*. New York, N.Y: Acoustical Society of America, 2017, vol. ANSI S.35-1997.

[22] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A Perceptual Model for Sinusoidal Audio Coding Based on Spectral Integration," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1292–1304, Jun. 2005.

[23] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

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

[26] ITU-T, "Recommendation P.862: Perceptual evaluation of speech quality (PESQ)," International Telecommunication Unition, Recommendation ITU-T P.862, Feb. 2001.

Paper D

Joint Far- and Near-end Speech and Listening Enhancement with Minimum Processing

Andreas Jonas Fuglsig, Zheng-Hua Tan,
Lars Søndergaard Bertelsen, Jesper Jensen, Jens Christian Lindof
and Jan Østergaard

The paper has been published in *IEEE ACCESS*, 2024.

The layout has been revised.

Abstract

This paper considers speech and listening enhancement for signals captured in one noisy environment that must be played back to a listener in another noisy environment. In both far-end speech enhancement and near-end listening enhancement, overly prioritizing noise suppression or maximizing intelligibility can result in undue speech distortions and reduced quality, especially when intelligibility is already high in favorable noise conditions. To address this, the use of a minimum processing framework has been proposed with the aim of reducing noise or enhancing listening to a minimum degree while ensuring that a specified intelligibility level is maintained. Furthermore, results have shown that jointly considering both environments improves performance compared to blindly concatenating far- and near-end methods. In blind processing, near-end listening enhancement typically assumes that the far-end signal is devoid of noise, potentially leading to erroneously interpreting noise as speech. Additionally, if the transmitter and receiver are blind to each other's presence, multiple instances of far- and near-end enhancement may occur and possibly work opposite directions, thus leading to degradations in the enhancement performance. In this paper, we perform a comprehensive exploration of our previously proposed joint far- and near-end minimum processing framework with systematic analysis and discussion. We derive a closed-form solution to the joint far- and near-end minimum processing optimization problem, with mean-square error processing penalty, a speech intelligibility constraint based on the approximated speech intelligibility index, and a noise power constraint. Performance was systematically studied using objective measures and listening tests for intelligibility, listening effort, and quality. We compared against relevant joint and blind methods with minimum and maximum processing. The results suggest that minimum processing achieves intelligibility comparable to maximum processing while preserving quality in higher SNRs, indicating its benefits in end-to-end communication. Joint processing provides advantages in objective estimated speech intelligibility for the minimum processing case, but not for maximum processing. However, no significant differences were observed in listening test results. This suggests that in certain speech and listening scenarios, it is feasible to optimize near- and far-end aspects separately, offering a more practical and convenient approach compared to joint optimization.

1 Introduction

Speech communication systems are used in various contexts, such as mobile phones, hearing aids, intercoms, and public address systems. Hence, they need to work in diverse situations, where background noise can significantly impact both Speech Intelligibility (SI) and Speech Quality (SQ).

In speech communication systems, we can distinguish between two separate environments, cf. Fig.D.1: The Far-End (FE) environment (the target

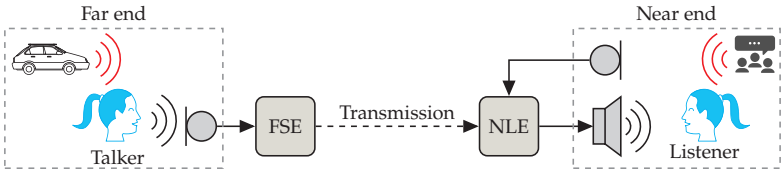


Fig. D.1: Speech communication system with Far-end Speech Enhancement (FSE) and Near-end Listening Enhancement(NLE).

talker's location) and the Near-End (NE) environment (the listener's location). Typically, both the FE and NE environments are subject to interfering acoustic noise sources, resulting in degradation of both SQ and SI for the listener. To mitigate this, noise reduction and speech enhancement algorithms can be applied in both FE and NE environments.

Based on the number of available microphones, Far-end Speech Enhancement (FSE) methods may employ either single- or multi-microphone noise reduction algorithms to remove noise from recorded signals [1–11] before transmission. While FSE methods can remove noise after it has been mixed with the target speech, Near-end Listening Enhancement (NLE) methods must process the signal received from the FE environment prior to playback in the noisy NE environment. Thus, FSE techniques cannot typically be used at the NE. Instead, NLE may utilize knowledge of the NE noise to pre-process the FSE signal coming from the FE to increase the SI and SQ in the NE background noise [12–24]. A particular type of NLE technique is Active Noise Control (ANC) [23, 25, 26], where the aim is to cancel the noise by adding an anti-phase noise component to the speech signal before playout. However, ANC performance is insufficient outside headsets and handheld mobile phone scenarios [23]. Therefore, because we are not only concerned with these scenarios, we only consider NLE based on speech modifications in this work. Recently Deep Neural Networks (DNNs) have shown good results when optimizing for advanced SQ and SI predictors in both FSE [7–9] and NLE [20–23]. However, increases in such predictions do not always translate to subjective performance gains [10]. Furthermore, DNNs have large memory usage, are difficult to interpret, and can have difficulties with generalizing to new acoustic scenarios. Hence, in this paper we take a more classic signal processing approach, providing closed-form and interpretable solutions.

Both SI and SQ are important in shaping the listening experience, and their significance varies depending on the acoustic setting [27–30]. However, traditionally, the aim of NLE methods has been to exclusively maximize SI. Similarly, FSE methods have been designed according to the inherent undesirability of noise, and thus, with the purpose of maximizing noise reduction, leaving only the clean speech signal. In noisy conditions, enhanced intelligibility may positively impact the perceived SQ [28, 29], thus making SI a crucial contributor to

SQ in noisy conditions [29]. To increase SI, NLE techniques potentially introduce speech distortions, which could diminish SQ; however, these distortions may be masked by more severe environmental NE noise [29, 30]. However, if the environmental noise subsides, speech becomes more intelligible and further SI enhancement is unattainable, and distortions become more disturbing [11, 19, 27–30]. Furthermore, aggressive noise reduction by FSE also leads to significant speech distortions and possibly to the loss of contextual noise from the FE environment [11, 31].

To remedy the effects of excessive FE processing, it was proposed in [11] to apply a *minimum processing principle* to multi-microphone FSE, such that the beamformer output is minimally processed with respect to a certain reference signal, provided that a given performance criterion is fulfilled [11]. In particular, two cases wee investigated in [11]: in the first case, the processing of the noisy signal is limited to the minimal amount necessary for fulfilling an SI requirement, and in the second case, noise is completely eliminated with an aggressive beamformer unless the resulting distortion of the clean speech violates the SI requirement. To remedy the effects of excessive processing at the NE, [19] applied the minimum processing principle to NLE. This provided an adaptive NLE that limits the processing of the signal received from the FE to the minimum required to achieve a target SI, thereby minimizing speech distortions in relation to the received signal.

Conventionally, the FE and NE scenarios have been considered separately [1, 16, 32]. However, several recent studies have proposed a joint approach to FE and NE speech enhancement in both single- [33–37] and multi-channel cases [38–40]. Particularly, [33] proposed a new training strategy for DNNs in cases where FSE is performed more than once. In [34], both FE noise reduction and modifications to limit speech distortions were added to an existing state-of-the-art NLE technique [24]. In [35], a DNN was trained to jointly remove FE noise and enhance SI using the method [24] as a "teacher". They showed improvements compared to the joint method in [34], but no comparison was made against blind approaches. In [37], a DNN was used to jointly enhance for multiple SQ and SI estimators, and showed improvements against versions of the proposed joint DNN method and blind concatenation of DNNs and classic signal processing methods, but no comparison was made against other joint approaches. The work of [36] proposed a classic signal processing approach to jointly control noise reduction and an NLE post-filter gain to increase SI, and improvements were reported against a blind signal processing approach in an informal preference test. For the multi-microphone case, [38] proposed a maximization of SI by closed-form optimization based on approximated mutual information, providing some small improvements in objective SI measures and an informal listening test against methods that were unaware of the remaining noise after FSE. It was then shown in [39] that similar performance could be obtained by a simpler closed-form optimization of the Approximated Speech

Intelligibility Index (ASII) [18, 41].

Apart from [34, 36, 37], the main goal of joint approaches has been to maximize SI. However, recently we proposed a *joint minimum processing beamforming and NLE* framework [40]. In contrast to existing joint processing works, this framework processes the signal the minimum amount required to achieve a desired target SI while preserving SQ in favorable noise condition [40]. Additionally, it extends the existing single-ended minimum processing frameworks [11, 19] to jointly consider all FSE, NLE, and environmental noises simultaneously. However, [40] only solved the optimization problem numerically, and a comparison was only made against the blind concatenation of the single-ended minimum processing frameworks of [11] and [19]. Thus, the joint minimum processing problem warrants further theoretical and experimental investigation.

In this paper, we conduct a comprehensive exploration of the joint FE and NE minimum processing framework initially introduced in [40]. The core contribution involves deriving a closed-form analytical solution for the joint FE and NE minimum processing optimization problem with a Mean-Square Error (MSE) processing penalty, an estimated SI constraint represented by the ASII, and a noise power SQ constraint. A systematic performance study encompassing objective measures and listening tests for SI, listening effort, and SQ, is conducted. We evaluate the effects of two aspects: minimum processing versus SI maximization and joint processing versus blind processing. Therefore, we compare against several methods: The joint ASII maximization of [39]; the blind concatenation of minimum processing FSE [11] and minimum processing NLE [19]; and blind concatenation classic "maximum" processing, i.e., a Minimum Variance Distortionless Response (MVDR) beamformer [1] and NLE ASII maximization [18]. The results suggest that minimum processing achieves a comparable SI to maximum processing while preserving good SQ in higher Signal-to-Noise Ratio (SNR) settings, emphasizing the benefits of applying the minimum processing principle in end-to-end communication scenarios. However, joint processing was only advantageous for estimated SI in the minimum processing case, but not in the maximum processing case. Finally, the subjective listening tests showed no significant differences between any of the tested methods. This leads to overall inconclusive but interesting results suggesting that in certain speech and listening scenarios, it is feasible to optimize the near- and far-end aspects separately, offering a more practical and convenient approach compared to joint optimization.

The remainder of this paper is organized as follows. In Section 2 we present our signal model. Section 3 introduces the minimum processing concept and details the differences between joint and blind approaches. In Section 4, we derive the case study optimization problem and its solution. Section 5 presents the experimental evaluation. Sections 6 and 7 present objective and subjective performance results, respectively. Finally, we discuss and conclude the paper

in Sections 8 and 9.

1.1 Abbreviations

For convenience Table D.1 lists the abbreviations used in this paper.

Table D.1: Abbreviations used in this paper.

Abbreviation	Description
ANC	Active Noise Control
ASII	Approximated Speech Intelligibility Index
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
ESTOI	Extended Short-Time Objective Intelligibility
FE	Far-End
FSE	Far-End Speech Enhancement
KKT	Karush-Kuhn-Tucker
NE	Near-End
NLE	Near-End Listening Enhancement
SI	Speech Intelligibility
SII	Speech Intelligibility Index
SQ	Speech Quality
MSE	Mean-Square Error
MUSHRA	Multiple Stimuli with Hidden Reference and Anchor
MVDR	Minimum Variance Distortionless Response
MWF	Multichannel Wiener Filter
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
PESQ	Perceptual Evaluation of Speech Quality

2 Signal Model

We consider the following time-frequency domain signal model with frequency index k and time index i, cf. Fig. D.2,

$$Z_{k,i} = g_{k,i}\boldsymbol{w}_{k,i}^H\boldsymbol{d}_{k,i}S_{k,i} + g_{k,i}\boldsymbol{w}_{k,i}^H\boldsymbol{U}_{k,i} + N_{k,i} \tag{D.1}$$

$$= g_{k,i}\boldsymbol{w}_{k,i}^H\boldsymbol{X}_{k,i} + N_{k,i} \tag{D.2}$$

$$= g_{k,i}Y_{k,i} + N_{k,i}. \tag{D.3}$$

Here, $S_{k,i}$ is the clean speech signal at the source location, $\boldsymbol{U}_{k,i} \in \mathbb{C}^M$ is the additive FE noise, and $\boldsymbol{X}_{k,i} \in \mathbb{C}^M$ is the noisy multi-microphone signal picked

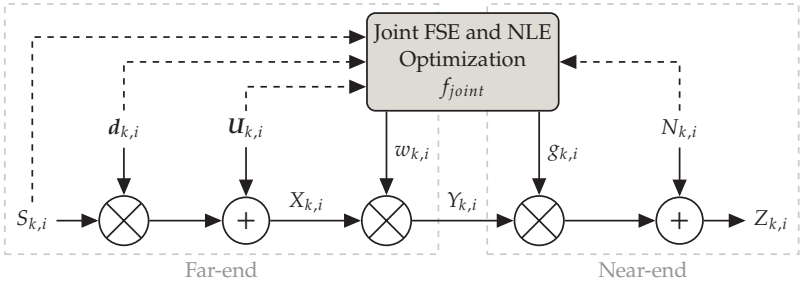


Fig. D.2: Signal model for joint FE and NE optimization.

up by M microphones, where $\boldsymbol{d}_{k,i} \in \mathbb{C}^M$ are acoustic transfer functions from the source to the microphones. First, to improve the SI and SQ of the noisy signal, $\boldsymbol{X}_{k,i}$, we employ a noise reduction beamformer, $\boldsymbol{w}_{k,i} \in \mathbb{C}^M$, as FSE with output signal $Y_{k,i}$. To further enhance SI and SQ, we apply an NLE gain, $g_{k,i} \in \mathbb{R}_+$, before the enhanced speech is played out in the noisy environment. The purpose of the beamformer $\boldsymbol{w}_{k,i}$ is to represent all processing focussed on the FE noise, whereas the purpose of the NLE gain is to amplify relevant speech regions over the NE noise. Finally, the signal received by the NE listener, $Z_{k,i}$, is contaminated by ambient noise, $N_{k,i}$ in the NE environment.

We model the speech and noise as complex random vector processes comprised of STFT coefficients. As is common in the literature, we assume that the speech and noise processes are uncorrelated and zero-mean random processes with independence across frequencies [32]. From these assumptions, we can obtain the speech distortion weighted covariance matrix $C^{(\mu)}_{\mathrm{X}_{k,i}}$ of $\boldsymbol{X}_{k,i}$ [11],

$$C^{(\mu)}_{\mathrm{X}_{k,i}} \triangleq C_{S_{k,i}} + \mu C_{U_{k,i}} = \sigma^2_{S_{k,i}} \boldsymbol{d}_{k,i} \boldsymbol{d}^H_{k,i} + \mu C_{U_{k,i}}, \tag{D.4}$$

where $\sigma^2_{S_{k,i}}$ is the clean speech power spectrum level, and $C_{U_{k,i}} \triangleq \mathrm{E}\left[\boldsymbol{U}_{k,i}\boldsymbol{U}^H_{k,i}\right]$ is the FE noise covariance matrix and $\mu \in \mathbb{R}_+$ is the speech distortion weight [2, 3], and $\mu = 1$ leads to the standard covariance matrix.

Similar to the existing minimum processing frameworks [11, 19, 40], we focus on signal processing within perceptually relevant subbands. That is, signals are analyzed and processed in, e.g., octave bands, fractional octave bands, or critical bands that all mimic aspects of human auditory perception. We define perceptually motivated subbands such that multiple frequency bins may be included in the same and/or more subbands, and denote subbands with index j and frequencies with index k. Hence, we can encompass the effect of non-rectangular auditory filters. Therefore, each frequency bin-subband pairing is assigned a weight. For the j'th subband, we denote the non-negative filter weights as $\omega_{j,k}$, and let $\mathbb{B}_j$ be the set of frequency bins that contribute to the j'th subband, where $j \in \{1, \ldots, J\}$ and J is the total number of subbands.

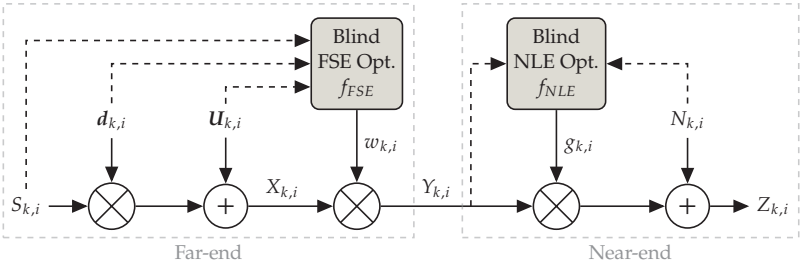


Fig. D.3: Block diagram of blind concatenation of FSE and NLE.

Thus, the NE noise spectrum level within one subband, j, and time-frame, i, is given as

$$\sigma^2_{N_{j,i}} \triangleq \sum_{k \in \mathbb{B}_j} \omega_{j,k} \sigma^2_{N_{k,i}}. \tag{D.5}$$

Any normalization for the filtering operation is already included within the subband filter weights, $\omega_{j,k}$. We provide additional details on the definition of $\omega_{j,k}$ and the connection between subbands and frequency bins in Appendix A.

In practical settings, joint FE and NE optimization requires sufficiently fast updating of signal properties and synchronization between the FE and NE for efficient processing. Similarly, the relevant signals must be both encoded, which leads to quantization noise, and transmitted across a possibly lossy channel. In this work, we are not concerned with these practical challenges and investigate performance under the assumption that they can be handled. Furthermore, with a similar motivation, we assume that no coupling issues exist between the microphones and loudspeakers at the FE and NE. Finally, in practice, the statistics of speech and noise processes must be estimated online, and we can apply the mathematical framework on a per time-frame basis. Hence, for brevity of notation, we disregard the time-index, i, and assume that we are working within a certain time frame, i, unless otherwise stated.

3 Concepts

3.1 Blind versus Joint Processing

With joint processing, cf. Fig. D.2, the beamformer, $\boldsymbol{w}_k$, and NLE gain, g_k, are determined as a function of all relevant signals, i.e.,

$$\left(\boldsymbol{w}^{joint},\ g^{joint}\right) = f_{joint}(S, U, N). \tag{D.6}$$

Hence, in joint processing all processing is optimized and derived jointly.

On the other hand, with blind processing, cf. Fig. D.3, the beamformer, $\boldsymbol{w}_k$, is first determined as a function of only the speech and FE noise. Then, the

NLE gain, g_k, is a function of the beamformer output signal and the NE noise, i.e,

$$\boldsymbol{w}^{blind} = f_{FSE}(S, U), \quad \text{and} \quad g^{blind} = f_{NLE}(Y, N) \tag{D.7}$$

Thus, the NLE is blind towards the effects of the FSE and cannot distinguish between speech and noise power in different parts of the spectrum. In particular, unlike joint processing, the signal after NLE is the output of a composite function of FSE and NLE.

3.2 The Minimum Processing concept

As in previous works [11, 19, 40], we assume that a designated target reference signal, S_k^R, is available, which could be the output signal from a beamformer with some specific desired characteristics; see [11] for more details. Focusing on a specific subband, j, we denote the number of frequency bins in this subband by $|\mathbb{B}_j|$. We then create the vectors $\boldsymbol{S}_j^R \in \mathbb{C}^{|\mathbb{B}_j|}$, $\boldsymbol{S}_j \in \mathbb{C}^{|\mathbb{B}_j|}$ and $\boldsymbol{Z}_j \in \mathbb{C}^{|\mathbb{B}_j|}$ by gathering all S_k^R, S_k and Z_k for $k \in \mathbb{B}_j$. Furthermore, we let $\mathcal{D}_j(\cdot,\cdot)$ and $\mathcal{I}_j(\cdot,\cdot)$ be two finite non-negative functionals indicating processing performance. Here, $\mathcal{D}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Z}_j\right)$ measures the distortion (processing penalty) between the target reference signal, S_k^R, and the signal perceived by the NE listener, Z_k, while $\mathcal{I}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Z}_j\right)$ is an intelligibility or performance estimator for the speech and listening enhancement in subband j. Consequently, the joint FE and NE minimum processing beamformer, $\boldsymbol{w}_k^{MP}$, and NLE gain, g_k^{MP}, for subband j, is defined as the solution to the optimization problem [40],

$$\underset{\{\boldsymbol{w}_k\},\{g_k\},k\in\mathbb{B}_j}{\arg\min} \quad \mathcal{D}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Z}_j\right) \quad \text{s.t.} \quad \mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right) \geq I_j', \tag{D.8}$$

where $I_j' \triangleq \min\left(I_j, I_j^{\max}\right)$ with I_j the desired minimum requirement on the SI performance $\mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right)$, and $I_j^{\max}$ the maximum achievable performance, when the performance $\mathcal{I}_j\left(\boldsymbol{S}_j, \boldsymbol{Z}_j\right)$ is maximized in an unconstrained manner.

In contrast to the existing minimum processing methods [11, 19], the joint approach considers the combined effects of all noise sources with FSE and NLE simultaneously.

Finally, we note that a straightforward approach to enhancing the NE output SNR involves increasing g_k indefinitely. However, this results in an infinite playback volume, and a substantial gap from the reference signal, thereby violating the fundamental principle of minimum processing.

4 Joint Far- and Near-end Minimum Processing

In this section, we introduce our previously proposed joint FE and NE minimum processing case study reported in [40] with preliminary results. We consider an MSE processing penalty, $\mathcal{D}_j$, and two performance criteria; an SI estimator, $\mathcal{I}_j^{SI}$, based on ASII [18, 19] and a noise power constraint, $\mathcal{I}_j^{NP}$, for SQ [40].

Inspired by the solutions in [11] and [36], we define a multichannel noise reduction vector (beamformer):

$$\boldsymbol{w}_k \triangleq \alpha_k \boldsymbol{v}_k^R + (1-\alpha_k)\boldsymbol{w}_k^{\mu\text{MWF}}, \tag{D.9}$$

where $\alpha_k \in [0,1]$, $\boldsymbol{v}_k^R$ is a pre-selected reference beamformer with a desired property, for example, low speech distortion (MVDR) or an ambient noise preserving beamformer [11]. Further, $\boldsymbol{w}_k^{\mu\text{MWF}}$ is the speech distortion weighted Multichannel Wiener Filter (MWF) MWF [2],

$$\boldsymbol{w}_k^{\mu\text{MWF}} \triangleq \left(C_{X_k}^{(\mu)}\right)^{-1} \sigma_{S_k}^2 \boldsymbol{d}_k \tag{D.10}$$

with pre-selected speech distortion weight, μ, e.g., $\mu \gg 1$ leading to high noise reduction and high speech distortion [11]. The parameter α_k provides a way to control the trade-off between the two beamformers, $\boldsymbol{v}_k^R$ and $\boldsymbol{w}_k^{\mu\text{MWF}}$, and their processing.

Early experiments show that solving (D.8) for our choice of $\mathcal{D}_j$ and $\mathcal{I}_j$ leads to solutions where only a single frequency within a subband is processed, leading to unpleasant artifacts. To avoid such solutions and ensure that we obtain more uniform processing across a subband, we assume that the NLE gains g_k and the combination weights α_k are fixed across an entire subband. That is,

$$\alpha_k = \alpha_i \quad \forall k, i \in \mathbb{B}_j \tag{D.11}$$

$$g_k = g_i \quad \forall k, i \in \mathbb{B}_j. \tag{D.12}$$

This also aligns with the results achieved in [11] and existing NLE studies [16, 19, 36, 38, 39], where the combination weights and NLE gains were derived to be fixed across subbands. Thus, we may write

$$\boldsymbol{w}_{j,k} \triangleq \alpha_j \boldsymbol{v}_k^R + (1-\alpha_j)\boldsymbol{w}_k^{\mu\text{MWF}}. \tag{D.13}$$

4.1 Processing Penalty

As suggested in [11, 19, 40], we consider a processing penalty, $\mathcal{D}_j(\cdot)$ based on an MSE criterion. We want to minimize the processing in relation to the reference

signal, S_k^R. The processing consists of two parts: the beamformer, $\boldsymbol{w}_{j,k}$, and the NLE gain, g_k. Therefore, we consider a processing penalty with two penalty terms: one that penalizes the processing caused by the beamformer, i.e., the distance between $\boldsymbol{S}_j^R$ and $\boldsymbol{Y}_j$, $\mathcal{D}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Y}_j\right)$, and another term that punishes the processing caused by the NLE gain, i.e., the distance between $\boldsymbol{Y}_j$ and $\boldsymbol{Z}_j$, $\mathcal{D}_j\left(\boldsymbol{Y}_j, \boldsymbol{Z}_j\right)$. That is,

$$\mathcal{D}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Z}_j\right) = \mathcal{D}_j\left(\boldsymbol{S}_j^R, \boldsymbol{Y}_j\right) + \mathcal{D}_j\left(\boldsymbol{Y}_j, \boldsymbol{Z}_j\right). \tag{D.14}$$

Since the reference signal, S_k^R, is the output of the reference beamformer, $\boldsymbol{v}_k^R$, the minimum processing solution to (D.8), i.e., $\boldsymbol{w}_{j,k}$ and g_k, should minimize the distance to $\boldsymbol{v}_k^R$ [40]. Therefore, we have that the processing penalty for j'th subband is,

$$\mathcal{D}_j(\boldsymbol{S}_j^R, \boldsymbol{Z}_j) = \left(1 - \alpha_j\right)^2 + \left(1 - g_j\right)^2, \tag{D.15}$$

where the details of the derivation are shown in Appendix B. The first term represents the processing penalty imposed on to the beamformer, urging $\boldsymbol{w}_{j,k}$ towards $\boldsymbol{v}_k^R$. The subsequent component signifies the penalty associated with the NLE gain, urging $g_k \boldsymbol{w}_{j,k}$ to approach $\boldsymbol{v}_k^R$ and mitigating potential speech distortions and excessive playback volume induced by the NLE gain [40].

4.2 Performance Criteria

For the performance criteria we consider both an intelligibility performance criterion, and a noise power criterion to increase quality performance [40].

4.2.1 Audibility constraint

The power of the processed speech within a subband for a given α_j is defined as

$$\delta_{S_j}(\alpha_j) \triangleq \sum_{k \in \mathbb{B}_j} \omega_{j,k} \boldsymbol{w}_k^H C_{S_k} \boldsymbol{w}_k \tag{D.16}$$

$$= \alpha_j^2 \delta_{S_j}^R + (1 - \alpha_j)^2 \delta_{S_j}^{\mu\text{MWF}} + \alpha_j (1 - \alpha_j) \delta_{S_j}^{cross} \tag{D.17}$$

where

$$\delta_{S_j}^{R} \triangleq \sum_{k\in\mathbb{B}_j} \omega_{j,k} \left(\boldsymbol{v}_k^R\right)^H C_{S_k} \boldsymbol{v}_k^R \tag{D.18}$$

$$\delta_{S_j}^{\mu\text{MWF}} \triangleq \sum_{k\in\mathbb{B}_j} \omega_{j,k} \left(\boldsymbol{w}_k^{\mu\text{MWF}}\right)^H C_{S_k} \boldsymbol{w}_k^{\mu\text{MWF}} \tag{D.19}$$

$$\delta_{S_j}^{cross} \triangleq \sum_{k\in\mathbb{B}_j} \omega_{j,k} 2\Re\left\{\left(\boldsymbol{w}_k^{\mu\text{MWF}}\right)^H C_{S_k} \boldsymbol{v}_k^R\right\}. \tag{D.20}$$

A similar definition applies to the processed FE noise subband power, $\delta_{U_j}(\alpha_j)$. We now define the processed NE and FE subband SNRs as

$$\xi_j^N \triangleq \frac{g_j^2 \delta_{S_j}(\alpha_j)}{g_j^2 \delta_{U_j}(\alpha_j) + \sigma_{N_j}^2}, \qquad \xi_j^F \triangleq \frac{\delta_{S_j}(\alpha_j)}{\delta_{U_j}(\alpha_j)}. \tag{D.21}$$

Thus, we define (processed) SNR as the ratio between the (processed) speech power and the (processed) noise power. This is in contrast to [11], where the signal-to-distortion ratio is used, i.e., the noise term is defined as the MSE between the clean speech and processed noise signal; hence all noise and speech distortions are considered as noise.

Remark 1 *We note that the processed NE SNR is upper bounded by the FE SNR, that is,*

$$\lim_{g_j\to\infty} \xi_j^N = \lim_{g_j\to\infty} \frac{g_j^2 \delta_{S_j}(\alpha_j)}{g_j^2 \delta_{U_j}(\alpha_j) + \sigma_{N_j}^2} = \frac{\delta_{S_j}(\alpha_j)}{\delta_{U_j}(\alpha_j)} = \xi_j^F. \tag{D.22}$$

Hence, the SI at the NE, as determined by the SNR, is upper bounded by the SI of the signal coming from the FE, since the NE noise can only lower the SI and we must compensate in the best possible way for this using the NLE gain.

Similar to [40] and [19], we derive optimal processing in relation to a performance criterion based on the ASII [18]. The original FSE minimum processing method [11] used SII. Both ASII [18] and SII [41] define SI as a weighted sum of intermediate subband audibility measures, where specifically for the ASII the subband audibility is given as a sigmoidal function of the NE subband SNR, ξ_j^N. We let I_j be a given minimum requirement on the ASII subband audibility performance [18] in subband, j. Then, by definition of the ASII audibility measures [18], it was shown in [19, App. C], that the SI constraint for the j'th

subband in terms of the NE subband SNR, ξ_j^N, is [40],

$$\frac{g_j^2 \delta_{S_j}(\alpha_j)}{g_j^2 \delta_{U_j}(\alpha_j) + \sigma_{N_j}^2} \geq I_j^{\xi} \tag{D.23}$$

$$\mathcal{I}_j^{SI} = g_j^2 \left(\delta_{S_j}(\alpha_j) - \delta_{U_j}(\alpha_j) I_j^{\xi} \right) \geq \sigma_{N_j}^2 I_j^{\xi}, \tag{D.24}$$

where $I_j^{\xi} \triangleq \frac{I_j}{1-I_j}$. As stated above, the NE subband SI is upper bounded by the FE subband SI. Considering the terms inside the parenthesis in (D.24), we see how the subband SI constraint is only feasible if an α_j exists such that $\xi_j^F > I_j^{\xi}$, i.e., if the parameterized beamformer can provide a feasible FE SNR. Therefore, as proposed in [40], we define the following parameter,

$$D_j^R \triangleq \delta_{S_j}^R - \delta_{U_j}^R I_j^{\xi}, \tag{D.25}$$

which indicates the ability of the reference beamformer, $\boldsymbol{v}_k^R$, to provide a feasible FE SNR. That is, D_j^R is positive only if the processed FE SNR resulting from the reference beamformer is above the desired audibility limit. Similarly, we can define the parameters $D_j^{\mu\text{MWF}}$ and D_j^{cross} which indicate the ability of $\boldsymbol{w}_k^{\mu MWF}$ and the cross combination of beamformers to provide a feasible FE SNR. Expanding the terms inside the parenthesis in (D.24) and using the above defined parameters, we can define the polynomial,

$$\begin{aligned} p_{FSE}(\alpha_j) \triangleq\ & \alpha_j^2 D_j^R + (1-\alpha_j)^2 D_j^{\mu\text{MWF}} \\ & + \alpha_j (1-\alpha_j) D_j^{cross} \end{aligned} \tag{D.26}$$

which represents the ability of the parameterized beamformer to remove sufficient FE noise for various values of α_j. For example, for $\alpha_j = 1$ the polynomial is equal to D_j^R and is positive only if the reference beamformer can remove sufficient FE noise. Using this polynomial, we can write the audibility constraint as

$$g_j^2 p_{FSE}(\alpha_j) \geq \sigma_{N_j}^2 I_j^{\xi}. \tag{D.27}$$

4.2.2 Noise power criterion

The joint approach has the advantage of having knowledge about the noise situation at both the FE and NE [40]. Thus, in contrast to the blind minimum processing of [11] and [19], we have the ability to try to control the processing in relation to noise in both environments [40].

From (D.23), it becomes apparent that, to enhance the SNR and meet the audibility requirement, it may be necessary for the processed FE noise to

surpass the NE noise. However, the improvement in SI could lead to an undesirable decline in SQ due to elevated overall noise levels, depending on the noise powers. Consequently, to mitigate distortions resulting from excessive noise levels, the joint minimum processing approach introduces a constraint, $\mathcal{I}_j^{NP}$, on the power of the processed FE noise [40],

$$10\log_{10}\left(g_j^2\delta_{U_j}(\alpha_j)\right) \leq 10\log_{10}\sigma_{N_j}^2 + \Delta_{U_j}. \tag{D.28}$$

Here, the parameter Δ_{U_j} is used to regulate the amount of dB the processed FE noise power can overpower or must stay below the NE noise power in subband j [40].

4.3 Optimization Problem and Solution

Combining the cost function and performance constraints, we have that the joint FE and NE minimum processing speech enhancement problem (D.8) with MSE processing penalty (D.15), ASII performance constraint (D.27) and noise power constraint (D.28) is [40],

$$\begin{aligned} &\underset{\alpha_j, g_j \in \mathbb{R}_+}{\arg\min} \quad \left(1-\alpha_j\right)^2 + \left(1-g_j\right)^2 && (P_0)\\ \text{s.t. } &C_1 : g_j^2 p_{FSE}(\alpha_j) \geq \sigma_{N_j}^2 I_j^{\xi}, \qquad C_3 : 0 \leq \alpha_j \leq 1,\\ &C_2 : g_j^2\delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}, \qquad\quad C_4 : 1 \leq g_j, \end{aligned}$$

where $c_{U_j} = 10^{\Delta_{U_j}/10}$.

Remark 2 *The interaction between the two performance constraints C_1 and C_2 and specifically the parameters I_j^{ξ} and Δ_{U_j} determine the feasibility of the optimization problem. For example, as the SI target increases, the audibility target, I_j^{ξ}, also increases, and it becomes more difficult to satisfy constraint C_1. Similarly, as we lower how much the FE noise is allowed to overpower the NE noise by lowering Δ_{U_j} it becomes increasingly difficult to satisfy constraint C_2.*

Thus, if the beamformers are sufficiently good, such that there exists an α_j where $p_{FSE}(\alpha_j) > 0$, then as the audibility constraint increases, it requires a larger NLE gain, g_j to satisfy C_1. The larger NLE gain then leads to increased boosting of the processed FE noise power, $\delta_{U_j}(\alpha_j)$. However, if Δ_{U_j} is chosen such that the FE noise is not allowed to sufficiently overpower the NE noise, the optimization may not have a feasible solution even though the intelligibility constraint is satisfied.

We show that the solution to the optimization problem is found at the boundary of the feasible set or at stationary points. Therefore, we identify the

following sets from the constraints,

$$\mathcal{F}^{C_1} \triangleq \{\alpha \in [0,1] : p_{FSE}(\alpha_j) \geq \sigma^2_{N_j} I^{\xi}_j\} \tag{D.29a}$$

$$\mathcal{F}^{C_2} \triangleq \{\alpha \in [0,1] : \delta_{U_j}(\alpha_j) \leq \sigma^2_{N_j} c_{U_j}\} \tag{D.29b}$$

$$\mathcal{F}^{SI} \triangleq \left\{\alpha \in [0,1] : \sigma^2_{N_j} I^{\xi}_j > p_{FSE}(\alpha_j) > 0\right\} \tag{D.29c}$$

$$\mathcal{F}^{NP} \triangleq \left\{\alpha \in [0,1] : I^{\xi}_j \delta_{U_j}(\alpha) - c_{U_j} p_{FSE}(\alpha) \leq 0\right\}. \tag{D.29d}$$

Finally, we have the set,

$$\mathcal{F}_S \triangleq \{\alpha \in [0,1] : h'(\alpha) = 0 \text{ and } h''(\alpha) \geq 0\}, \tag{D.30}$$

containing the stationary points of the convex regions of the following helper function,

$$h(\alpha_j) \triangleq \sqrt{\frac{\sigma^2_{N_j} I^{\xi}_j}{p_{FSE}(\alpha_j)}} - \alpha_j. \tag{D.31}$$

Details about the first and second derivatives of $h(\alpha_j)$ are provided in Appendix C.1.4. Denoting the boundary of a set $\mathcal{F}$ by $\partial\mathcal{F}$, we have the following theorem, stating the solution to the optimization problem.

Theorem 2 *The optimal minimum processing beamformer weight, α^*_j, and NLE gain, g^*_j, solution to the optimization problem (P_0) are:*

If $\mathcal{F}^{C_1} \cap \mathcal{F}^{C_2} \neq \emptyset$ or $\mathcal{F}^{SI} \cap \mathcal{F}^{NP} \neq \emptyset$: α^*_j is the minimum of the stationary and boundary solutions, i.e.,

$$\begin{aligned} \alpha^*_j = \arg\min_{\alpha}\ & \pi(\alpha), \\ s.t.\ \alpha \in\ & \left(\mathcal{F}_S \cap \mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right) \\ & \cup\, \partial\left(\mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right) \cup \partial\left(\mathcal{F}^{C_1} \cap \mathcal{F}^{C_2}\right), \end{aligned} \tag{D.32}$$

where

$$\pi(\alpha) = \begin{cases} 1-\alpha & \text{if } \alpha \in \partial\left(\mathcal{F}^{C_1} \cap \mathcal{F}^{C_2}\right) \\ h(\alpha) & \text{if } \alpha \in \left(\mathcal{F}_S \cap \mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right) \\ & \text{or if } \alpha \in \partial\left(\mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right). \end{cases} \tag{D.33}$$

and

$$g^*_j = \begin{cases} 1 & \text{if } \alpha^*_j \in \partial\left(\mathcal{F}^{C_1} \cap \mathcal{F}^{C_2}\right) \\ \sqrt{\frac{\sigma^2_{N_j} I^{\xi}_j}{p_{FSE}(\alpha^*_j)}} & \text{if } \alpha^*_j \in \left(\mathcal{F}_S \cap \mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right) \\ & \text{or if } \alpha^*_j \in \partial\left(\mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right). \end{cases} \tag{D.34}$$

If $\mathcal{F}^{C_1} \cap \mathcal{F}^{C_2} = \emptyset$ and $\mathcal{F}^{SI} \cap \mathcal{F}^{NP} = \emptyset$: No feasible solution exists.

The proof of Theorem 2 is found in Appendix C.

Remark 3 *We note that the set of stationary points, $\mathcal{F}_S$ is a finite discrete set. Similarly, the boundary of an interval in $\mathbb{R}$ is a finite discrete set, and hence, the boundaries of the intersections $\partial\left(\mathcal{F}^{C_1} \cap \mathcal{F}^{C_2}\right)$ and $\partial\left(\mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right)$ are finite discrete sets. Thus, to find the optimal solution, we avoid searching over a continuum of points and instead only need to compare a small finite number of points. In fact, from the proof in Appendix C it can be seen that, $\|\mathcal{F}_S\|_0 \leq 2$ and $\|\mathcal{F}^i\|_0 \leq 4$, for $i \in \{C_1, C_2, SI, NP\}$.*

With the optimal solution (α_j^*, g_j^*) the optimal minimum processing beamformer is then expressed as

$$\boldsymbol{w}_{j,k}^* = \alpha_j^* \boldsymbol{v}_k^R + (1 - \alpha_j^*)\boldsymbol{w}_k^{\mu MWF}. \tag{D.35}$$

Based on the subband definition, various frequencies may contribute to multiple subbands indexed by $\mathbb{F}_k$. Consequently, the optimal beamformer $\boldsymbol{w}j, k^*$ and NLE gain g_j^* can influence multiple subbands. Denoting the weight that reflects the influence of this contribution as $\eta_{j,k}$, the optimal beamformer and NLE gain for each frequency are,

$$\boldsymbol{w}_k^* = \sum_{j \in \mathbb{F}_k} \eta_{j,k} \boldsymbol{w}_{j,k}^* \quad \text{and} \quad g_k^* = \sum_{j \in \mathbb{F}_k} \eta_{j,k} g_j^*. \tag{D.36}$$

See Appendix A for definition of the weights $\eta_{j,k}$.

4.3.1 Infeasible cases

For the sets in (D.29), $\mathcal{F}^{C_1}$ represents the feasible α for which $g = 1$ satisfies C_1, $\mathcal{F}^{C_2}$ represents the feasible α for which $g = 1$ satisfies C_2, $\mathcal{F}^{SI}$ represents the feasible α for which we must have $g > 1$ to satisfy C_1, and $\mathcal{F}^{NP}$ represents the feasible α for which $g^2 = \sigma_{N_j}^2 I_j^{\xi} / p_{FSE}(\alpha_j) > 1$ satisfies C_2. Particularly, this means, $\mathcal{F}^{C_1} = \varnothing$ implies no α exists for which the beamformer can provide a feasible FE SNR for the optimal $g = 1$; $\mathcal{F}^{C_2} = \varnothing$ implies no α exists for which the beamformer can provide a feasible FE noise power relative to NE noise for the optimal $g = 1$; $\mathcal{F}^{SI} = \varnothing$ implies no α exists for which the beamformer provides a feasible FE SNR where the optimal NLE gain is $g > 1$; and finally $\mathcal{F}^{NP} = \varnothing$ implies no α exists for which the beamformer provides a feasible FE noise power relative to the NE noise for the optimal $g^2 = \sigma_{N_j}^2 I_j^{\xi} / p_{FSE}(\alpha_j) > 1$.

For a feasible solution, the constraints C_1 and C_2 must both be satisfied simultaneously. Thus, we have that $\mathcal{F}^{C_1} \cap \mathcal{F}^{C_2} = \varnothing$ indicates that no α exists for which the optimal NLE gain is $g = 1$, and $\mathcal{F}^{SI} \cap \mathcal{F}^{NP} = \varnothing$ indicates that no α exists for which the optimal NLE gain is $g^2 = \sigma_{N_j}^2 I_j^{\xi} / p_{FSE}(\alpha_j) > 1$. If either constraint cannot be satisfied individually, then the constraints cannot

Category	$\mathcal{F}^{C_1}$	$\mathcal{F}^{C_2}$	$\mathcal{F}^{SI}$	$\mathcal{F}^{NP}$	$\mathcal{F}^{C_1} \cap \mathcal{F}^{C_2}$	$\mathcal{F}^{SI} \cap \mathcal{F}^{NP}$
Intersection	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
Intersection	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
Intersection	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
Intersection	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
Intersection	$\neq \varnothing$	$\varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
Intersection	$\varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
Intersection	$\varnothing$	$\varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
FE SI	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
FE SI	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
FE SI	$\varnothing$	$\varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
FE Noise power	$\neq \varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
FE Noise power	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
FE Noise power	$\varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
SI+NP	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
SI+NP	$\varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
SI+NP	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$

Table D.2: Categorization of infeasible cases for the optimal joint minimum processing optimization problem.

be jointly feasible. Therefore, and because the intersection between an empty set and another set is always empty, we have a total of 16 different combinations of the sets in (D.29) that can result in an infeasible solution, cf. Table D.2. For simplicity, we combine these configurations into four categories, as indicated in Table D.2: (1) infeasibility due to empty intersections, i.e., the beamformer can provide a feasible FE SI and a feasible FE noise power but not for the same α; (2) infeasibility because the FE beamformer cannot provide a sufficient SI for any choice of g; (3) infeasibility because the processed FE noise power is too high in relation to the NE noise power for any choice of g; and (4) infeasibility because the beamformer cannot provide either a feasible FE SI or noise power for any choice of g. In Section 6.1, we further investigate how the number of feasible and infeasible subbands changes with the FE and NE SNRs.

5 Experimental Evaluation

We investigate performance in two noise scenarios: (Babble-Car) where the target talker is in a babble noise setting and the NE listener is inside a car, and (Car-Babble) the reverse situation where the target talker is in car noise and the NE listener is in babble noise. Additional investigations, not reported here, have shown that the results generalize to other noise types.

5.1 Reference methods

We compare performance against three relevant reference methods: (1) **Blind Min**; The blind concatenation of minimum processing FSE [11] and minimum processing NLE [19]. The FSE beamformer of [11] is parameterized according to a reference beamformer and a μ-MWF. Therefore, these beamformers are selected to be the same as those used in the proposed joint method. This allows us to investigate the effect of combining joint processing with minimum processing in the end-to-end speech enhancement setting. (2) **Joint Max**: The joint FE and NE optimization method based on maximizing ASII [39], since this is also a joint approach based on ASII but without minimum processing. Thus, we investigate effects of adding minimum processing to the joint setting. (3) **Blind Max**: The blind concatenation of an MVDR beamformer at the FE (as this is the beamformer used in [39]) and optimal ASII maximization at the NE [18]. Thus, we investigate the performance of the simple classic maximum processing, and the effect of adding minimum processing to the classic blind end-to-end approach. Furthermore, we investigate the effect of using a joint approach [39] against its classic blind counter part without any relation to minimum processing.

The reference methods [39] and [18] are based on a target speech power equality constraint. That is, the output power should match the power of the input signal, P_{ref}. However, the proposed method is allowed to gain the input signal to overpower the NE noise, resulting in a total processed power level of P_{prop}. Therefore, for a fair comparison, the power constrained methods [39] and [18] are implemented with the same power gain to their input signals such that $P_{ref} = P_{prop}$. The blind NLE reference of [19] was, similar to the proposed method, designed to gain the input signal a sufficient amount to overpower the NE noise. However, we cannot control the amount of total applied gain, because the method was designed specifically without this option.

Finally, the blind concatenation methods are implemented such that the NLE parts [19] and [18] interpret the signal coming from the FE, Y_k, as clean speech.

5.2 Handling Infeasible Subbands

The joint minimum processing optimization problem (P_0) is solved per subband. However, as shown in Theorem 2, the existence of a feasible solution is determined by the subband noise and speech powers. Therefore, to investigate performance using real speech and noise signals, infeasible subbands must be handled. In [40], various heuristic approaches were proposed to find good solutions in infeasible subbands. However, to better investigate the performance differences between the proposed joint minimum processing and the blind concatenated minimum processing, in this paper we instead default

to using the blind minimum processing concatenation within these infeasible subbands. Thus, all differences between the proposed and blind minimum processing method are due to our joint solution in the feasible subbands.

5.3 Estimating statistics

The statistics of speech and non-stationary noisy signals change over time, and must therefore be estimated and updated in time. However, updating too fast may lead to abrupt changes in the processing between time frames, leading to audible distortions. Therefore, slow time-varying processing is commonly employed. Hence, in this paper, for simplicity, we estimate the average speech and noise power per DFT bin using a long-term average over several short-time frames,

$$\sigma_{S_k}^2 \triangleq \frac{1}{I}\sum_i |S_{k,i}|^2, \tag{D.37}$$

$$C_{U_k} \triangleq \frac{1}{I}\sum_i \boldsymbol{u}_{k,i}\boldsymbol{u}_{k,i}^H, \tag{D.38}$$

where I denotes the total number of frames. An expression similar to (D.37) holds for the NE noise signal, $N_{k,i}$. Thus, the estimated statistics do not change with time, and we process signals in a time-invariant manner. For time-varying processing in practice, the statistics must be updated with, for example, a recursive average. Furthermore, in the simulations, we assume that the speech and noise spectra are known. However, in practical scenarios, the speech and noise spectra must be estimated from noisy microphone recordings [32].

5.4 Experimental Setup

Unless otherwise stated, the target speech material used for the evaluations are English sentences from the TIMIT-database [42] test set sampled at 16 kHz. We pad each target speech excerpt with 1.5 s of silence at the beginning and end to allow ramping up the noise level before the speech segment starts and down after it ends, ensuring a more pleasant listening experience in subjective listening tests.

The babble noise is created by mixing talkers from the TIMIT training set. In the (Babble-Car) scenario, we use six talkers per FE noise source position, and in the (Car-Babble) setting, we use six competing talkers at the NE. The car noise is generated by taking a random excerpt of an appropriate length from noise recorded inside a car traveling at 130 km/h.

We consider an FE room with dimensions $3 \times 4 \times 3\,\mathrm{m}^3$, four noise sources at $[0.50, 1.00, 1]$ m, $[0.75, 3.00, 1]$ m, $[3.00, 2.40, 1]$ m, and $[2.70, 1.30, 1]$ m, and a target talker at $[1.50, 3.00, 1]$ m. The FE beamformer has three microphones at

$[1.50, 2.00, 1]$ m, $[1.50, 2.02, 1]$ m and $[1.50, 1.98, 1]$ m, where each microphone is subject to a 60 dB SNR white noise. We assume that the room transfer functions are known and generated without reverberation using [43]. The speech and noise signals are converted to the time-frequency domain using an STFT with 32 ms Hann windows with 50% overlap and a sampling rate of 16 kHz. We consider a total of $J = 30$ auditory subband filters with center frequencies linearly spaced on the equivalent rectangular bandwidth scale from 150 Hz to 8000 Hz [44].

We investigate performance with a focus on achieving high SI with minimal speech distortion. Therefore, we select the reference beamformer, v^R, as the MVDR beamformer, and $\mu = 5$ for the μ-MWF. The per-band audibility targets, I_j, and noise power constraints, Δ_{U_j}, are derived from the overall parameters A^* and Δ_U, using the band importance functions of the SII as weights, cf. [19, Sec. IV.B]. Through informal listening and objective scoring we tuned the parameters of the proposed method, such that the target total ASII is $A^* = 0.9$ for all scenarios, while $\Delta_U = 0$ dB in the (Babble-Car) scenario and $\Delta_U = -5$ dB in the (Car-Babble) scenario. We note that these are not generally applicable values, and a new tuning should be made when working with other scenarios.

Finally, to further control the gains and limit excessive sound levels, we set an additional maximum gain limit on the NLE gain, g_j^*, of 60 dB.

6 Objective Performance

Performance is evaluated and averaged across 10 trials[1] from the TIMIT dataset. We consider performance in terms of estimated SI with the ESTOI metric and estimated SQ using the PESQ metric. Furthermore, we consider how the number of feasible bands changes in (P_0) with varying FE and NE SNRs.

6.1 Feasible and infeasible subbands

We consider 30 auditory subband filters and solve the joint minimum processing optimization problem (P_0) for each subband. However, as shown in Theorem 2, a feasible solution may not exist for all subbands depending on the noise situation at the FE and NE. As, these infeasible situations must be handled, it is of great interest to investigate how often we are in an infeasible case (and which one) or in a feasible case. Table D.3 shows the average number of feasible and infeasible subbands for each infeasibility category (Table D.2) for various SNR and noise combinations, where the average is taken across the 10 TIMIT trials.

[1] Audio samples available: `https://afugls.github.io/Joint_MinProc_FSE_and_NLE/`

SNR FE	NE Band Category	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Feasible	2.3	1.9	1.1	1.0	1.0	1.0	0.9	0.8	0.5	0.1	0.0
	Infeas. NP	0.0	0.0	0.0	0.1	0.6	0.7	0.7	0.7	0.7	0.7	0.7
−12.5	Infeas. SI	27.0	27.0	26.9	24.7	17.8	8.9	4.9	3.0	1.8	0.8	0.4
	Infeas. SI+NP	0.5	0.5	1.0	3.8	10.3	19.1	23.1	25.1	26.4	27.7	28.3
	Infeas. intersec.	0.2	0.6	1.0	0.4	0.3	0.3	0.4	0.4	0.6	0.7	0.6
	Feasible	6.1	6.1	4.8	3.1	3.0	2.8	2.4	1.7	0.6	0.2	0.0
	Infeas. NP	0.0	0.0	0.0	0.1	0.7	1.5	1.7	1.9	2.0	2.1	2.1
−5.0	Infeas. SI	21.6	21.6	21.6	20.0	11.8	3.2	1.2	0.3	0.0	0.0	0.0
	Infeas. SI+NP	1.2	1.2	1.3	4.4	13.3	21.3	23.2	24.6	25.6	26.6	27.0
	Infeas. intersec.	1.1	1.1	2.3	2.4	1.2	1.2	1.5	1.5	1.8	1.1	0.9
	Feasible	9.8	9.8	9.6	6.0	4.5	4.0	3.1	1.9	0.9	0.2	0.0
	Infeas. NP	0.0	0.0	0.0	0.2	1.2	1.7	1.9	2.1	2.1	2.1	2.1
0.0	Infeas. SI	17.7	17.7	17.7	17.6	11.5	2.1	0.4	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.7	1.7	1.7	2.2	11.7	21.3	23.2	24.7	26.0	27.0	27.7
	Infeas. intersec.	0.8	0.8	1.0	4.0	1.1	0.9	1.4	1.3	1.0	0.7	0.2
	Feasible	25.6	25.6	25.6	25.6	24.2	9.9	4.7	3.6	2.4	1.0	0.3
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.4	1.4	1.4	1.4
15.0	Infeas. SI	0.7	0.7	0.7	0.7	0.7	0.6	0.5	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.3	1.3	1.3	1.3	1.3	7.0	18.7	23.7	25.0	26.5	27.6
	Infeas. intersec.	2.4	2.4	2.4	2.4	3.8	12.5	5.3	1.3	1.2	1.1	0.7

SNR NE	FE Band Category	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Feasible	0.0	0.0	0.0	1.1	3.6	7.9	12.8	20.8	29.4	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−40.0	Infeas. SI	30.0	30.0	30.0	28.7	25.6	19.7	14.6	3.1	0.5	0.0	0.0
	Infeas. SI+NP	0.0	0.0	0.0	0.1	0.4	1.1	1.6	3.5	0.1	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.0	0.1	0.4	1.3	1.0	2.6	0.0	0.0	0.0
	Feasible	0.0	0.0	0.0	0.4	2.9	7.9	12.8	20.8	29.4	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−30.0	Infeas. SI	30.0	30.0	30.0	28.7	25.6	19.7	14.6	3.1	0.5	0.0	0.0
	Infeas. SI+NP	0.0	0.0	0.0	0.1	0.4	1.1	1.6	3.5	0.1	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.0	0.8	1.1	1.3	1.0	2.6	0.0	0.0	0.0
	Feasible	0.0	0.0	0.0	0.3	1.7	3.9	5.7	16.5	29.4	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.1	0.2	1.0	0.4	0.0	0.0	0.0	0.0
−10.0	Infeas. SI	29.7	28.3	26.1	21.5	14.8	11.3	12.7	3.1	0.5	0.0	0.0
	Infeas. SI+NP	0.3	1.7	3.9	8.1	13.0	12.5	7.5	4.0	0.1	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.0	0.0	0.3	1.3	3.7	6.4	0.0	0.0	0.0
	Feasible	0.0	0.0	0.0	0.3	1.6	3.0	3.7	4.3	6.8	22.8	29.9
	Infeas. NP	0.0	0.0	0.0	0.2	0.7	1.4	2.4	3.7	0.0	0.0	0.0
5.0	Infeas. SI	25.9	21.1	14.5	7.3	3.3	0.7	0.3	0.7	0.5	0.0	0.0
	Infeas. SI+NP	4.1	8.9	15.5	22.1	23.9	23.7	22.7	19.0	9.8	0.9	0.0
	Infeas. intersec.	0.0	0.0	0.0	0.1	0.5	1.2	0.9	2.3	12.9	6.3	0.1

(a) Band counts for far-end Babble and near-end Car noise

Table D.3: Out of a total of 30 subbands the tables show the average number of subbands that were either feasible or infeasible according to the four categories in Table D.2 for (a) the FE babble and NE car noise scenario and (b) the FE car and NE babble noise scenario. Background color scales with the score, brighter colors correspond to higher values.

SNR FE	NE Band Category	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Feasible	3.8	3.8	3.8	3.8	3.8	3.8	3.8	2.6	0.0	0.0	0.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	1.5	1.5	1.5
−40.0	Infeas. SI	24.7	24.7	24.7	24.7	24.7	24.7	24.4	19.7	10.1	4.6	3.2
	Infeas. SI+NP	1.5	1.5	1.5	1.5	1.5	1.5	1.8	6.8	18.3	23.9	25.3
	Infeas. intersec.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0
	Feasible	26.7	26.7	26.7	26.7	26.7	26.7	26.7	25.3	22.1	13.7	8.5
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.3	2.5
−10.0	Infeas. SI	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.0
	Infeas. SI+NP	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.7	5.2	12.9	18.5
	Infeas. intersec.	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.5	2.2	0.6	0.5
	Feasible	29.5	29.5	29.5	29.5	29.5	29.5	29.5	28.9	25.7	21.3	10.9
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
−5.0	Infeas. SI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.7	2.9	8.4	18.6
	Infeas. intersec.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.4	0.3	0.0
	Feasible	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	28.7	26.1	24.5
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5.0	Infeas. SI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	3.8	5.5
	Infeas. intersec.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.1	0.0

SNR NE	FE Band Category	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Feasible	3.8	5.6	7.8	15.2	26.7	29.5	30.0	30.0	30.0	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−20.0	Infeas. SI	24.7	22.3	16.5	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.5	2.1	5.6	9.2	2.5	0.5	0.0	0.0	0.0	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.1	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
	Feasible	3.8	5.6	7.8	15.2	26.7	29.5	30.0	30.0	30.0	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−10.0	Infeas. SI	24.7	22.3	16.5	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.5	2.1	5.6	9.2	2.5	0.5	0.0	0.0	0.0	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.1	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
	Feasible	3.8	5.6	7.8	15.2	26.7	29.5	30.0	30.0	30.0	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
−5.0	Infeas. SI	24.7	22.3	16.5	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.5	2.1	5.6	9.2	2.5	0.5	0.0	0.0	0.0	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.1	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
	Feasible	1.5	5.5	7.4	13.5	24.6	28.4	29.8	30.0	30.0	30.0	30.0
	Infeas. NP	1.0	0.9	2.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15.0	Infeas. SI	16.4	13.7	12.8	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	10.8	9.9	7.8	9.2	3.3	1.0	0.1	0.0	0.0	0.0	0.0
	Infeas. intersec.	0.3	0.0	0.0	2.0	1.6	0.6	0.1	0.0	0.0	0.0	0.0

(b) Band counts for far-end Car and near-end Babble noise

Table D.3: (Continued).

The results show that, for fixed NE SNRs, as the FE SNR increases the number of feasible subbands increases, as the number of infeasible subbands due to insufficient FE SI decreases. For higher fixed NE SNRs, the number of infeasible subbands that satisfy neither the SI nor the noise power constraint first increases and then decreases. In particular, in (Babble-car) scenario (Table D.3a) the number of infeasible subbands owing to empty intersections first increases and then decreases. Thus, as the FE SNR increases, it generally becomes easier for the beamformer to remove sufficient FE noise to satisfy the subband SI and noise power constraints. However, for higher NE SNRs, the remaining FE noise power is still so high that the beamformer cannot maintain the processed FE noise below the NE noise until the FE SNR is sufficiently high and the FE noise power subsides.

For fixed FE SNRs, as the NE SNR increases, the number of feasible subbands decreases. This seems counterintuitive at first, because the number of infeasible subbands due to insufficient FE SI decreases. However, this is caused by an increase in the number of infeasible subbands satisfying neither the SI nor the noise power constraint. Furthermore, for the (Babble-Car) scenario, the number of infeasible subbands owing to empty intersections first increases and then decreases. Thus, as the NE SNR increases, satisfying the SI constraint becomes easier. However, too much FE noise still remains and causes infeasibility because the NE noise is so low that the FE noise still overpowers it.

6.2 Estimated Intelligibility

Tables D.4 and D.5 show the ESTOI performance of the proposed and reference methods in the (Babble-Car) and (Car-Babble) scenario, respectively. At low SNRs, all methods demonstrate an improvement in performance compared to the unprocessed scenario, except for instances of FE babble noise with an SNR less than $-25\,\mathrm{dB}$ for all NE car noise SNRs. In FE car noise scenarios, performance improvement is more substantial than in babble noise situations because car noise is less detrimental to SI and easier to remove, thus providing better conditions for the NLE component to operate effectively. As the SNRs increase, the unprocessed performance naturally improves, making it challenging for the various processing methods to exhibit significant enhancements. Additionally, at higher SNRs, objective scores may penalize processing artifacts, even though the speech is highly intelligible. Thus, the differences in scores at this level may not accurately reflect real-world perceptual improvements.

Comparing processing methods, the maximum processing techniques generally exhibit slightly higher performance than minimum processing methods, with the difference becoming more pronounced at higher SNRs. This is expected because the maximum processing methods are designed to maximize

SNR FE	NE Method	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0
	Unproc.	.123	.178	.214	.241	**.267**	**.285**	**.292**	**.294**	**.295**
	Joint Min	**.259**	**.260**	**.260**	**.260**	.261	.263	.267	.273	.279
−12.5	Blind Min	.250	.249	.250	.250	.250	.253	.259	.269	.278
	Joint Max	.251	.251	.251	.251	.252	.255	.260	.262	.262
	Blind Max	.252	.253	.253	.253	.254	.257	.262	.264	.264
	Unproc.	.155	.256	.330	.386	.444	**.497**	**.529**	**.542**	**.545**
	Joint Min	.442	.442	.442	.443	.446	.458	.478	.498	.511
−5.0	Blind Min	.425	.425	.425	.426	.429	.440	.466	.494	.511
	Joint Max	**.448**	**.448**	**.448**	**.449**	**.453**	.470	.486	.490	.491
	Blind Max	.446	.446	.446	.447	.451	.471	.488	.493	.493
	Unproc.	.164	.282	.382	.458	.541	**.621**	**.677**	**.703**	**.710**
	Joint Min	.548	.548	.548	.549	.554	.583	.626	.655	.670
0.0	Blind Min	.530	.530	.531	.532	.538	.561	.611	.650	.669
	Joint Max	**.571**	**.571**	**.571**	**.571**	**.581**	.618	.643	.650	.651
	Blind Max	.566	.566	.566	.566	.577	.618	.645	.652	.653
	Unproc.	.170	.301	.429	.537	.658	.790	**.895**	**.945**	**.963**
	Joint Min	.670	.670	.670	.672	.687	.746	.836	.897	.926
15.0	Blind Min	.667	.667	.667	.669	.685	.743	.833	.896	.926
	Joint Max	**.736**	**.736**	**.736**	**.737**	**.760**	**.840**	.895	.913	.917
	Blind Max	.731	.731	.731	.733	.757	.840	.895	.913	.917

SNR NE	FE Method	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0
	Unproc.	.006	.014	.041	.089	.137	.161	.168	.170	.171
	Joint Min	.007	.015	**.057**	**.158**	**.318**	.500	.615	.663	.676
−40.0	Blind Min	.007	.015	**.057**	.153	.306	.481	.605	.659	.675
	Joint Max	.011	.019	.053	.148	.312	**.513**	**.654**	**.721**	**.737**
	Blind Max	**.011**	**.020**	.054	.150	.313	.509	.648	.716	.732
	Unproc.	.004	.015	.046	.124	.226	.305	.336	.345	.347
	Joint Min	.004	.017	**.058**	**.157**	**.318**	.500	.615	.663	.676
−30.0	Blind Min	.004	.017	**.058**	.153	.306	.481	.605	.659	.675
	Joint Max	.012	.020	.054	.148	.313	**.513**	**.654**	**.721**	**.737**
	Blind Max	**.013**	**.021**	.055	.150	.314	.509	.648	.716	.732
	Unproc.	.003	.012	.053	**.160**	**.326**	.496	.605	.651	.664
	Joint Min	.006	.017	**.057**	.158	.319	.505	.627	.679	.694
−10.0	Blind Min	.006	.017	**.057**	.154	.307	.486	.616	.676	.693
	Joint Max	.013	.021	.054	.149	.314	**.520**	**.674**	**.746**	**.766**
	Blind Max	**.013**	**.021**	.055	.150	.316	.517	.670	.742	.763
	Unproc.	.004	.013	.054	**.167**	**.367**	**.607**	**.787**	**.880**	.911
	Joint Min	.006	.018	**.058**	.160	.332	.556	.736	.822	.850
5.0	Blind Min	.006	.018	**.058**	.157	.322	.541	.724	.818	.850
	Joint Max	.013	.021	.054	.151	.329	.567	.769	.877	**.914**
	Blind Max	**.013**	**.021**	.055	.152	.331	.569	.770	.877	**.914**

Table D.4: Average ESTOI scores in the (Babble-Car) scenario. Best performance is highlighted in bold for each SNR and noise pair. Background color scales with the score. Brighter colors correspond to higher values. Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

SNR FE	NE Method	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0
	Unproc.	.005	.009	.023	.063	.137	.212	.254	.268	.271
	Joint Min	.259	.271	.277	.277	.277	.277	.277	.277	.279
−40.0	Blind Min	.260	.272	.278	.278	.278	.278	.278	.278	.279
	Joint Max	.275	.281	.282	.283	.283	.283	.283	.283	.283
	Blind Max	**.284**	**.285**	**.285**	**.285**	**.285**	**.285**	**.285**	**.285**	**.285**
	Unproc.	.006	.009	.025	.081	.206	.391	.577	.702	.756
	Joint Min	.675	.685	.685	.685	.685	.688	.699	**.727**	**.760**
−10.0	Blind Min	.649	.657	.657	.657	.657	.659	.671	.718	.759
	Joint Max	.710	.722	.723	.723	.723	.723	.723	.724	.729
	Blind Max	**.713**	**.725**	**.726**	**.726**	**.726**	**.726**	**.726**	.727	.732
	Unproc.	.006	.009	.025	.082	.213	.411	.621	.769	**.837**
	Joint Min	.701	.711	.712	.712	.712	.717	.744	.795	.830
−5.0	Blind Min	.683	.693	.694	.694	.694	.697	.721	.786	.829
	Joint Max	.778	.792	.794	.794	.794	.794	.794	.795	.802
	Blind Max	**.780**	**.795**	**.796**	**.796**	**.796**	**.796**	**.796**	**.797**	.805
	Unproc.	.006	.009	.025	.082	.217	.429	.667	.845	**.934**
	Joint Min	.719	.731	.732	.732	.732	.740	.787	.868	.910
5.0	Blind Min	.715	.726	.727	.727	.727	.734	.777	.863	.909
	Joint Max	.865	.882	.884	.884	.884	.884	.884	.885	.896
	Blind Max	**.865**	**.882**	**.884**	**.884**	**.884**	**.884**	**.884**	**.886**	.897

SNR NE	FE Method	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0
	Unproc.	.046	.049	.050	.052	.055	.056	.056	.056	.056
	Joint Min	.277	.392	.491	.587	.685	.721	.732	.734	.735
−20.0	Blind Min	.278	.393	.490	.579	.657	.706	.727	.733	.735
	Joint Max	.283	.401	.502	.607	.723	.824	.884	**.906**	**.911**
	Blind Max	**.285**	**.405**	**.506**	**.610**	**.726**	**.825**	**.884**	**.906**	**.911**
	Unproc.	.137	.164	.181	.194	.206	.215	.217	.218	.218
	Joint Min	.277	.392	.491	.587	.685	.721	.732	.734	.735
−10.0	Blind Min	.278	.393	.490	.579	.657	.706	.727	.733	.735
	Joint Max	.283	.401	.502	.607	.723	.824	.884	**.906**	**.911**
	Blind Max	**.285**	**.406**	**.506**	**.610**	**.726**	**.825**	**.884**	**.906**	**.911**
	Unproc.	.190	.239	.271	.300	.326	.345	.352	.353	.354
	Joint Min	.277	.392	.491	.587	.686	.724	.735	.738	.738
−5.0	Blind Min	.278	.393	.490	.579	.658	.708	.730	.736	.738
	Joint Max	.283	.401	.502	.607	.723	.824	.884	**.906**	**.911**
	Blind Max	**.285**	**.406**	**.506**	**.610**	**.726**	**.825**	**.884**	**.906**	**.911**
	Unproc.	.270	.395	.500	.610	.727	.829	.884	.901	.905
	Joint Min	.278	.394	.496	.610	**.739**	**.836**	.886	.901	.905
15.0	Blind Min	.279	.395	.497	.608	.735	.831	.882	.900	.904
	Joint Max	.283	.401	.503	.607	.724	.826	.887	.909	.915
	Blind Max	**.285**	**.406**	**.507**	**.610**	.727	.828	**.887**	**.909**	**.915**

Table D.5: Average ESTOI scores in the (Car-Babble) scenario. Best performance is highlighted in bold for each SNR and noise pair. Background color scales with the score. Brighter colors correspond to higher values. Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

SI. However, because we provided a high SI target the minimum processing methods also yield high ESTOI scores. The high SI target also leads to a larger power increase, which the maximum processing methods utilize to achieve a slightly higher ESTOI score.

When comparing the joint minimum processing and blind minimum processing methods, their performances are relatively similar. However, joint minimum processing tends to have slightly better ESTOI performance than blind minimum processing when the FE SNR is at or above -20 dB for all noise scenarios and NE SNRs. Looking at Table D.3, we see that this corresponds to the number of feasible bands starting to rise (if not before depending on NE noise scenario). Because we chose to let the joint minimum processing default to the blind approach in the infeasible subbands, we expect joint minimum processing to behave similar to blind processing if there are no feasible bands. Thus, the differences in performance are caused by the joint solution in the feasible subbands, and joint processing can, for certain SNRs, outperform blind processing in the minimum processing setting. However, for the maximum processing methods, their performances remain largely similar across various noise and SNR combinations, with only marginal differences. In cases where differences exist, it is inconsistent whether joint or blind maximum processing is superior, with blind processing seemingly outperforming more frequently. This is surprising since the joint maximum processing method has no feasibility issues; thus, it should always be able to utilize joint knowledge to outperform blind maximum processing. Hence, based on objective SI estimation using ESTOI, it is inconclusive whether joint processing can consistently enhance performance over blind processing .

Finally, as either noise type vanishes (very high SNRs), we see, as expected, that there are no differences between the joint and blind methods. This is expected because if there is little to no noise at either the FE or NE, the end-to-end communication scenario tends towards a single-end scenario. Hence, both the joint and blind methods operate on the same terms, and it does not matter if the joint method can utilize joint knowledge if there is not much to be knowledgeable about. Thus, using the proposed joint minimum processing method, we automatically obtain the performance and behavior seen in the single-sided minimum processing [11, 19].

6.3 Estimated Quality

Tables D.6 and D.7 show the PESQ performance of the proposed and reference methods in the (Babble-Car) and (Car-Babble) scenario, respectively. We note that PESQ is very sensitive to noise, and PESQ scores tend to be dominated by noise at lower SNRs, thus making it very difficult to discern any differences when noise is the dominant component [11]. As the SNRs increase, the environmental FE noise and environmental NE noise become less dominant, and

SNR FE	NE Method	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Unproc.	1.06	**1.11**	**1.15**	1.12	1.13	1.09	1.08	1.13
	Joint Min	1.09	1.09	1.10	1.13	1.27	1.13	**1.14**	1.15
−12.5	Blind Min	**1.10**	1.10	1.11	**1.14**	**1.27**	**1.13**	**1.14**	1.15
	Joint Max	1.05	1.05	1.05	1.13	1.03	1.03	1.05	1.05
	Blind Max	1.04	1.05	1.06	1.06	1.06	1.07	1.06	**1.19**
	Unproc.	1.04	**1.10**	**1.15**	**1.18**	1.20	1.20	1.20	1.20
	Joint Min	1.09	1.08	1.10	1.16	1.19	1.23	1.23	**1.23**
−5.0	Blind Min	**1.10**	1.09	1.11	1.17	**1.20**	**1.23**	**1.23**	**1.23**
	Joint Max	1.04	1.04	1.05	1.07	1.08	1.08	1.08	1.08
	Blind Max	1.03	1.04	1.05	1.07	1.07	1.08	1.08	1.08
	Unproc.	1.04	**1.12**	**1.24**	**1.31**	**1.35**	1.36	1.37	1.37
	Joint Min	1.11	1.12	1.18	1.27	1.33	1.37	1.38	**1.38**
0.0	Blind Min	**1.12**	1.12	1.18	1.28	1.34	**1.38**	**1.38**	**1.38**
	Joint Max	1.04	1.04	1.07	1.12	1.14	1.14	1.14	1.14
	Blind Max	1.04	1.04	1.07	1.12	1.14	1.14	1.14	1.14
	Unproc.	1.05	1.18	**1.54**	**2.03**	**2.45**	**2.64**	**2.70**	**2.72**
	Joint Min	**1.18**	**1.20**	1.43	1.86	2.22	2.46	2.53	2.56
15.0	Blind Min	1.15	1.18	1.42	1.86	2.22	2.46	2.54	2.56
	Joint Max	1.05	1.06	1.18	1.52	1.87	2.01	2.04	2.04
	Blind Max	1.05	1.07	1.18	1.52	1.87	2.01	2.04	2.05

SNR NE	FE Method	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Unproc.	1.10	**1.13**	1.02	1.02	1.03	1.03	1.03	1.02
	Joint Min	1.16	1.07	1.10	**1.14**	**1.17**	**1.17**	**1.16**	**1.15**
−40.0	Blind Min	1.18	1.08	**1.11**	1.14	1.15	1.15	1.15	1.15
	Joint Max	1.06	1.04	1.04	1.04	1.05	1.05	1.05	1.05
	Blind Max	**1.22**	1.03	1.04	1.04	1.05	1.05	1.05	1.05
	Unproc.	1.08	1.02	1.02	1.02	1.02	1.02	1.02	1.02
	Joint Min	1.17	1.07	1.10	**1.14**	**1.17**	**1.17**	**1.16**	**1.16**
−30.0	Blind Min	**1.18**	**1.08**	**1.11**	1.14	1.15	1.16	1.16	1.16
	Joint Max	1.05	1.03	1.04	1.04	1.05	1.05	1.05	1.05
	Blind Max	1.11	1.03	1.04	1.04	1.05	1.05	1.05	1.05
	Unproc.	1.16	**1.07**	**1.11**	**1.15**	1.17	1.18	1.18	1.18
	Joint Min	**1.18**	1.06	1.10	1.15	**1.19**	**1.20**	**1.19**	**1.19**
−10.0	Blind Min	1.16	1.07	1.10	1.15	1.18	1.18	1.19	1.19
	Joint Max	1.04	1.04	1.04	1.05	1.06	1.06	1.06	1.06
	Blind Max	1.11	1.04	1.04	1.05	1.06	1.06	1.06	1.06
	Unproc.	1.11	1.10	**1.23**	**1.52**	**1.92**	**2.18**	**2.27**	**2.29**
	Joint Min	1.14	1.10	1.21	1.45	1.77	1.96	2.03	2.05
5.0	Blind Min	1.14	**1.11**	1.22	1.46	1.77	1.96	2.03	2.05
	Joint Max	1.08	1.05	1.09	1.22	1.45	1.61	1.66	1.67
	Blind Max	**1.19**	1.05	1.09	1.22	1.45	1.61	1.66	1.67

Table D.6: Average PESQ scores in the (Babble-Car) scenario. Best performance is highlighted in bold for each SNR and noise pair. Background color scales with the score. Brighter colors correspond to higher values. Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

SNR FE	NE Method	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
−40.0	Unproc.	1.03	1.03	1.02	1.02	1.02	1.02	1.02	1.02
	Joint Min	1.20	1.20	1.20	1.20	1.19	**1.17**	**1.14**	**1.14**
	Blind Min	**1.20**	**1.20**	**1.20**	**1.20**	**1.20**	**1.17**	**1.14**	**1.14**
	Joint Max	1.07	1.07	1.07	1.07	1.07	1.07	1.07	1.07
	Blind Max	1.07	1.07	1.07	1.07	1.07	1.07	1.07	1.07
−10.0	Unproc.	1.04	1.06	1.11	1.19	1.33	1.45	1.50	1.51
	Joint Min	2.04	2.04	2.06	2.07	**2.10**	**2.11**	2.31	2.41
	Blind Min	**2.06**	**2.06**	**2.07**	**2.09**	2.10	2.10	**2.31**	**2.42**
	Joint Max	1.50	1.50	1.50	1.50	1.50	1.53	1.65	1.71
	Blind Max	1.54	1.54	1.54	1.54	1.54	1.58	1.72	1.78
−5.0	Unproc.	1.04	1.06	1.12	1.23	1.49	1.77	1.91	1.95
	Joint Min	**2.12**	**2.12**	**2.13**	**2.15**	2.18	**2.23**	2.54	2.71
	Blind Min	2.11	2.11	2.13	2.15	**2.18**	2.22	**2.54**	**2.71**
	Joint Max	1.63	1.63	1.63	1.63	1.64	1.69	1.91	2.04
	Blind Max	1.65	1.65	1.65	1.65	1.66	1.72	1.96	2.09
5.0	Unproc.	1.04	1.06	1.13	1.27	1.65	2.23	2.70	2.92
	Joint Min	**2.14**	**2.14**	**2.16**	**2.19**	2.25	2.39	2.93	3.34
	Blind Min	2.14	2.14	2.16	2.19	**2.25**	**2.39**	**2.93**	**3.34**
	Joint Max	1.82	1.82	1.82	1.82	1.83	1.93	2.43	2.79
	Blind Max	1.82	1.82	1.82	1.82	1.84	1.94	2.43	2.80

SNR NE	FE Method	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
−20.0	Unproc.	1.04	1.05	1.05	1.05	1.04	1.05	1.05	1.05
	Joint Min	1.84	2.04	**2.13**	**2.14**	**2.14**	**2.14**	**2.14**	**2.14**
	Blind Min	**1.85**	**2.06**	2.13	2.14	2.14	2.14	2.14	**2.14**
	Joint Max	1.31	1.50	1.69	1.82	1.87	1.88	1.89	1.89
	Blind Max	1.35	1.54	1.71	1.82	1.87	1.88	1.89	1.89
−10.0	Unproc.	1.05	1.06	1.06	1.06	1.06	1.06	1.06	1.06
	Joint Min	1.84	2.04	**2.13**	**2.14**	**2.14**	**2.14**	**2.14**	**2.14**
	Blind Min	**1.85**	**2.06**	2.13	2.14	2.14	2.14	**2.14**	**2.14**
	Joint Max	1.31	1.50	1.69	1.82	1.87	1.88	1.89	1.89
	Blind Max	1.35	1.54	1.71	1.82	1.87	1.88	1.89	1.89
−5.0	Unproc.	1.07	1.09	1.09	1.10	1.10	1.10	1.10	1.10
	Joint Min	1.84	2.05	**2.14**	**2.15**	**2.15**	**2.15**	**2.15**	**2.15**
	Blind Min	**1.86**	**2.06**	2.13	2.15	2.15	2.15	2.15	**2.15**
	Joint Max	1.31	1.50	1.69	1.82	1.87	1.88	1.89	1.89
	Blind Max	1.35	1.54	1.71	1.82	1.87	1.88	1.89	1.89
15.0	Unproc.	1.12	1.38	1.67	1.84	1.94	1.96	1.96	1.96
	Joint Min	1.85	**2.11**	2.22	2.28	**2.31**	**2.32**	**2.32**	**2.32**
	Blind Min	**1.85**	2.09	**2.22**	**2.28**	2.31	2.32	**2.32**	**2.32**
	Joint Max	1.31	1.50	1.71	1.85	1.90	1.91	1.92	1.92
	Blind Max	1.36	1.55	1.72	1.85	1.90	1.91	1.92	1.92

Table D.7: Average PESQ scores in the (Car-Babble) scenario. Best performance is highlighted in bold for each SNR and noise pair. Background color scales with the score. Brighter colors correspond to higher values. Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

the PESQ scores increase, and improvements become evident. Furthermore, PESQ compares clean speech to the (processed) noisy signal, thus the unprocessed scenarios may receive the highest score in higher SNRs where speech distortions exceed the noise level.

In the (Babble-Car) noise scenario, with FE SNRs below 0 dB and NE SNRs below -5 dB, PESQ scores remain consistently low, and differences are indiscernible between processed and unprocessed signals. For higher SNRs, the performance of the processing methods improves. However, the unprocessed performance remains the highest, with the minimum processing methods being able to better maintain a low distortion and higher speech quality than the maximum processing methods. For the (Car-Babble) noise scenario, all methods can improve performance over the unprocessed signal, even at low SNRs, where PESQ can be dominated by the noise signal [11], and increases in SI lead to increases in SQ as the speech signal becomes clearer within the noise.

Comparing processing methods, the minimum processing methods consistently outperform the maximum processing methods, particularly at high SNRs, where they exhibit improved speech quality compared to their maximum processing counterparts. At lower SNRs, the flexibility of the minimum processing methods to adopt a more aggressive MWF beamformer allows them to remove more noise than the MVDR beamformer employed in the maximum processing methods, resulting in a more substantial increase in PESQ. Thus, minimum processing methods achieve competitive estimated objective SI on par with maximum processing methods while concurrently achieving superior estimated objective SQ.

Comparing between joint and blind methods, there is no difference in performance in either the minimum or the maximum processing case. As mentioned earlier, at high FE SNRs, there are many feasible bands, cf. Table D.3, but not much to gain using a joint approach method because there is not much noise to consider and the terms are the same for both joint and blind methods. However, surprisingly we see no difference at the medium SNRs where there are still many feasible bands. Again it is interesting, that there is no difference between joint and blind maximum processing, because there are no feasibility issues, so joint should always be able to utilize its joint knowledge to increase performance.

7 Subjective Performance

In this section, we evaluate the performance of the proposed joint minimum processing along with the reference methods in subjective listening tests.

The objective results showed that the joint minimum processing method obtained a slightly better estimated SI than the blind minimum processing.

However, the results were inconclusive regarding the effects of joint processing in the maximum processing setting. Although there was a small advantage to maximum processing over minimum processing in the estimated SI. Additionally, there were no differences between joint and blind processing in terms of estimated SQ, but there was a clearer difference between minimum and maximum processing in estimated SQ. However, subjective intelligibility and quality may not always be well represented by objective measures, and the differences observed in these methods may not reflect realistic performance [10]. Furthermore, we conducted informal listening tests, which indicated that it was difficult to distinguish between the intelligibility of the evaluated methods, even when there was a large difference in objective performance.

Therefore, to further evaluate the performance differences between the joint and blind methods, and the minimum processing and SI maximization methods, we conducted subjective listening tests. We perform a listening test for SI in combination with listening effort, and a separate listening test for perceptual SQ. We investigate self-reported listening effort, since SI also affects listening effort [27], and we only found small differences in estimated SI. Thus, it is interesting to see if joint processing has a clearer effect on listening effort than on SI.

The SNRs used in the listening tests were chosen such that there is noise present at both ends since we are interested in the joint setup with end-to-end noisy communication. If we consider very high SNRs at either end, we would repeat the studies of [11] and [19].

7.1 Shared setup and procedure

Both listening tests were performed in a silent room. A Lenovo T460s laptop connected to an external monitor was used for reporting and displaying the user interface. The laptop was also equipped with a USB sound card (DragonFly Black) and a set of closed over-ear headphones (Beyerdynamic DT-770 Pro 32 ohm) for audio playback. All audio stimuli in both tests were normalized to a perceived loudness of −31 LUFS according to the EBU R128 [45] recommendation for loudness normalization as implemented in `ffmpeg-normalize`[2]. The test participants were allowed to adjust the overall volume to a comfortable level during a prior training session of each test. Both tests included a short training session to familiarize the participants with the test procedure, audio stimuli, and user interface and limit learning bias. The training scores were not included in the final test results. Finally, to limit listening fatigue, participants were not allowed to participate in both tests on the same day.

[2] `https://github.com/slhck/ffmpeg-normalize`

7.2 Speech Intelligibility and Listening Effort Test

7.2.1 Setup

A total of 22 (3 female, 19 male) native Danish speaking untrained listeners, with an age span of 25 to 65 years and an average age of 39.9 years, volunteered for participation. All participants had self-reported normal hearing. The average test time, including the training, was 52 minutes. The user interface was based on [46] and was modified to also enable self-reported listening effort.

7.2.2 Procedure

We performed a closed-vocabulary matrix test combined with an additional rating of listening effort. The speech material used for this test was the Danish Dantale II corpus [47]. The utterances in the Dantale II corpus were recorded by a single female native Danish speaker in silent conditions. Each utterance has a syntactical structure of `name + verb + numeral + adjective + object` and was generated by randomly choosing a word from a set of 10 different candidate words for each word class [47].

For a series of trials, the user interface presented a matrix of all the candidate words for each of the five word classes. Participants initiated audio playback via a mouse click, and the audio was played *only once*. Subsequently, participants selected the words heard from the matrix of candidate words using the mouse. When participants were satisfied with their word selection, a second window automatically opened where participants were asked to rate, using a slider interface, how much listening effort they spent on understanding the words on a scale from 0 (no effort) to 10 (maximum effort). After rating the listening effort, the interface returned to the word matrix and the stimulus of the next trial was automatically played. This procedure was repeated until the end of the test. Intelligibility is measured as the percentage of correctly identified words.

Each test consisted of 2 noise scenarios $\times$ 2 SNR pairs $\times$ 5 processing types (including unprocessed) $\times$ 6 sentences = 120 trials. The target speech and the order of the trials were random for each participant. For the (Babble-Car) scenario, performance was evaluated at FE and NE SNR pairs (-12.5 dB, -40 dB) and (-7.5 dB, -30 dB). In the (Car-Babble) scenario, performance was evaluated at FE and NE SNR pairs (-40 dB, -20 dB) and (-10 dB, -10 dB). The SNR that showed the largest difference in ESTOI scores did not show differences in SI in informal listening, and indicated full intelligibility. Therefore, we performed the listening test at more severe SNRs where it might be possible to detect differences in SI.

The training session consisted of 20 trials (1 sentence per noise/SNR/processing pair).

7.3 Speech Quality Test

7.3.1 Setup

The speech quality listening test was conducted by 25 (4 female, 21 male) volunteer untrained listeners, with an age span of 24 to 65 years and an average age of 39.3 years. All participants had self-reported normal hearing. The average test time, including the training, was 46 minutes. The speech material used for the speech quality listening test was sentences from the English TIMIT test set. The test was conducted using a user interface that was slightly modified from [48].

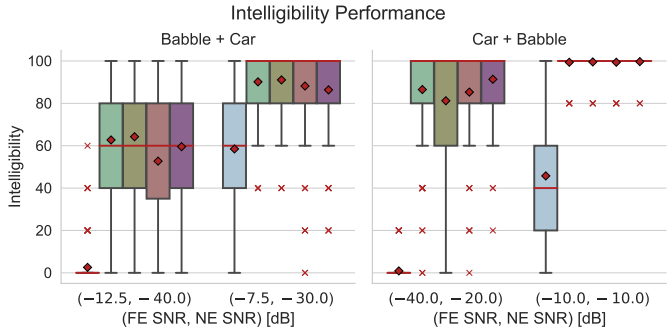
7.3.2 Procedure

We carried out a listening test using the MUlti Stimulus with Hidden Reference and Anchor (MUSHRA) paradigm [49]. Participants evaluated audio quality on a scale ranging from 0 to 100, segmented into five equal intervals denoted as *bad, poor, fair, good*, and *excellent*. Participants were specifically directed to assess the *basic audio quality* in comparison to a known reference signal, with no additional specifications provided for the definition of audio quality. Each participant was presented with a sequence of 2 noise scenarios $\times$ 2 SNR pairs $\times$ 4 sentences = 16 trials. Both the reference sentences and the order of the trials were random for each participant. Each trial consisted of a clean reference signal and 7 other signals to be rated: 1 hidden reference, the 4 systems under test, 1 unprocessed signal, and 1 hidden anchor (unprocessed signal at lower FE and NE SNRs). For the (Babble-Car) scenario, performance was evaluated at the FE and NE SNR pairs (0 dB, -10 dB) and (15 dB, 5 dB), and the anchor SNR pair was (-10 dB, -20 dB). In the (Car-Babble) scenario, performance was evaluated at the FE and NE SNR pairs (-5 dB, -5 dB) and (5 dB, 20 dB), and the anchor SNR pair was (-25 dB, -10 dB).

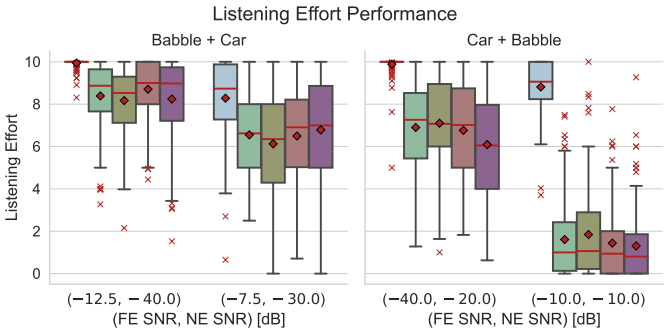
The training session consisted of four trials (1 sentence per noise-SNR pair).

7.4 Listening test results

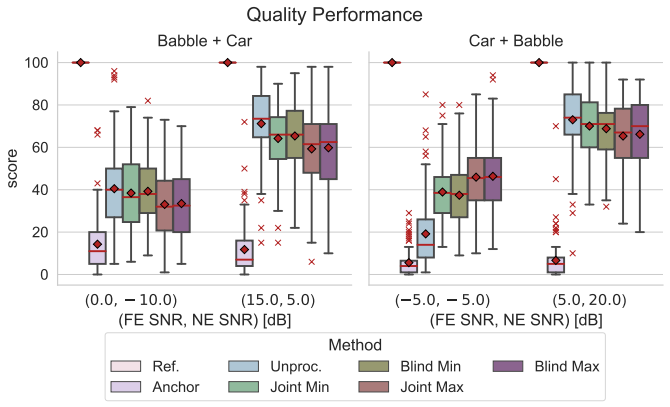
Figure D.4 shows box plots of the results of the three listening tests for each noise, processing, and SNR condition. For statistical significance tests of the results, we consider the nonparametric Kruskal-Wallis H test [50], since the assumption of normal distribution of the data is invalid according to the Kolmogorov-Smirnov test [51], and given the number of participants and their different interpretations of the scales [52]. The p-values for the comparisons considered in this paper are listed in Table D.8. p-values are considered significant and are marked in bold if $p < 0.05/m = 0.005$ ($m = 10$), where we have corrected the significance level with the Bonferroni method [53], and m is the number of tested hypotheses.



(a) Speech intelligibility test results. Higher is better.



(b) Listening effort test results. Lower is better.



(c) MUSHRA test results. Higher is better.

Fig. D.4: Boxplot of the (a) speech intelligibility, (b) listening effort, and (c) speech quality test results for FE babble and NE car noise (left) and FE car and NE babble noise (right). Medians and means are indicated by red horizontal lines and diamonds, respectively. Outliers are indicated by red crosses. Legend from bottom plot applies to all figures (no reference or anchor was used in the (a) intelligibility and (b) listening effort tests). Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

p-values	Babble + Car		Car + Babble	
Comparison	$(-12.5, -40)$	$(-7.5, -30)$	$(-40, -20)$	$(-10, -10)$
Joint Min - Blind Min	0.6657	0.6102	0.0426	0.7022
Joint Min - Joint Max	0.0075	0.9908	0.2195	1.0000
Joint Min - Blind Max	0.4083	0.1598	0.1328	0.4097
Joint Min - Unproc.	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Blind Min - Joint Max	**0.0023**	0.6160	0.3242	0.7022
Blind Min - Blind Max	0.2230	0.0585	**0.0003**	0.6522
Blind Min - Unproc.	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Joint Max - Blind Max	0.0600	0.1878	0.0051	0.4097
Joint Max - Unproc.	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Blind Max - Unproc.	**0.0000**	**0.0000**	**0.0000**	**0.0000**

(a) p-values of the intelligibility test.

p-values	Babble + Car		Car + Babble	
Comparison	$(-12.5, -40)$	$(-7.5, -30)$	$(-40, -20)$	$(-10, -10)$
Joint Min - Blind Min	0.22380	0.30091	0.49752	0.43898
Joint Min - Joint Max	0.08729	0.80512	0.73846	0.44787
Joint Min - Blind Max	0.96382	0.25050	**0.00382**	0.12780
Joint Min - Unproc.	**0.00000**	**0.00000**	**0.00000**	**0.00000**
Blind Min - Joint Max	**0.00373**	0.24302	0.25136	0.12982
Blind Min - Blind Max	0.28745	0.03407	**0.00051**	0.02486
Blind Min - Unproc.	**0.00000**	**0.00000**	**0.00000**	**0.00000**
Joint Max - Blind Max	0.09525	0.29859	0.02168	0.41195
Joint Max - Unproc.	**0.00000**	**0.00000**	**0.00000**	**0.00000**
Blind Max - Unproc.	**0.00000**	**0.00000**	**0.00000**	**0.00000**

(b) p-values of the listening effort test.

p-values	Babble + Car		Car + Babble	
Comparison	$(0, -10)$	$(15, 5)$	$(-5, -5)$	$(5, 20)$
Joint Min - Blind Min	0.70117	0.62662	0.46110	0.51945
Joint Min - Joint Max	0.03357	0.03194	**0.00063**	0.04592
Joint Min - Blind Max	0.04324	0.06199	**0.00066**	0.11483
Joint Min - Unproc.	0.58908	**0.00079**	**0.00000**	0.13086
Blind Min - Joint Max	0.00937	0.00950	**0.00013**	0.16240
Blind Min - Blind Max	0.01275	0.02265	**0.00013**	0.34973
Blind Min - Unproc.	0.92308	**0.00371**	**0.00000**	0.02867
Joint Max - Blind Max	0.89596	0.89014	0.98147	0.67942
Joint Max - Unproc.	0.00659	**0.00000**	**0.00000**	**0.00055**
Blind Max - Unproc.	0.01126	**0.00000**	**0.00000**	**0.00253**

(c) p-values of the MUSHRA test.

Table D.8: p-values for (a) the intelligibility, (b) listening effort, and (c) MUSHRA test. p-values below a Bonferroni corrected significance level of 0.005 are marked in bold. Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

7.4.1 Speech Intelligibility

Considering Fig. D.4a and Table D.8a jointly, the results show across both noise scenarios and all SNR pairs that all processing methods significantly enhance SI compared to the unprocessed condition, a result consistent with the objective ESTOI scores. However, generally, there are large variations in the data, and it is difficult to determine a particular trend. Notably, no statistically significant differences were observed between the maximum and minimum processing methods. Similarly, no significant differences were found between joint and blind processing, with the exception of the (Car-Babble) scenario at very low SNRs. Surprisingly, in this case, blind minimum processing demonstrated better performance than joint and blind maximum processing. In the (Babble-Car) scenario, there appears to be a slight, insignificant advantage favoring minimum processing methods over maximum processing; this trend is also reflected in the ESTOI scores. Despite the expectation that maximum processing should outperform minimum processing and joint processing outperform blind processing, the limited participant pool precluded the determination of statistically significant performance differences.

7.4.2 Listening Effort

Considering the listening effort results in Fig. D.4b and Table D.8b, the results show that all methods significantly alleviate listening effort compared to the unprocessed condition across both noise scenarios and all SNR pairs. However, no discernible significant differences were observed between the maximum and minimum processing methods. Similarly, no clear distinctions were observed between joint and blind processing, with the exception of the low SNR case in the (Car-Babble) noise scenario, where blind maximum processing exhibited a significantly better outcome than the minimum processing methods. However, the data exhibits substantial variations, and no clear tendencies can be discerned.

7.4.3 Speech Quality

Considering Fig.D.4c and TableD.8c, we see no trends or significant differences in SQ between the blind and joint processing methods. However, some trend is seen aligning with PESQ scores at high SNRs, where the minimum processing methods outperform maximum processing. However, no significant differences could be determined between the minimum and maximum processing methods. Although, in the low SNR case of the (Car-Babble) scenario, the maximum processing methods significantly outperform minimum processing methods. This is in contrast with the PESQ results, where minimum processing was notably superior to maximum processing. However, the maximum processing methods had higher ESTOI scores, although minimum processing

also achieved high ESTOI scores, cf. Table D.5. In addition, for the high SNR case of (Car-Babble), maximum processing methods are significantly worse than the unprocessed performance, while no significant differences are observed between minimum processing methods and unprocessed or between minimum and maximum processing. Hence, the optimal balance between SI, speech distortions, and noise suppression remains unclear in scenarios with noise at both ends. In the high SNR case of the (Babble-Car) scenario, all methods were significantly worse than the unprocessed, aligning with the PESQ results. This indicates that the increase in estimated SI comes at the cost of speech distortion. Although not significant, the minimum processing methods appear to exhibit better quality than the maximum processing methods, which is also seen in the PESQ results.

7.4.4 Joint versus blind

The number of feasible bands is worth noting when comparing joint and blind minimum processing. There were only a limited number of feasible bands in both the SI and listening effort tests, and also in the (Babble-Car) scenario for the quality test. Hence, there is only slight variations between the processing of the joint and blind minimum processing methods. This can partially explain why we see only a slight performance difference between these methodologies. Notably, while higher SNRs presented more feasible bands, lower SNRs levels were specifically chosen for the listening tests because ESTOI and informal listening tests suggested that unprocessed SI was so high that all processing would lead to maximum SI and the absence of observable differences at higher SNRs. In contrast, in the (Car-Babble) scenario in the speech quality test, there were many feasible bands; however, no significant differences were observed between the joint and blind minimum processing. Furthermore, it is interesting that there were only very small variations between the joint and blind maximum processing despite the absence of feasibility constraints.

In summary, the subjective listening tests yielded inconclusive results regarding the superiority of joint processing over blind processing in both the minimum and maximum processing scenarios. Additionally, we did not establish statistically significant improvements in speech quality with minimum processing over maximum processing in high SNR cases, despite trends aligned with objective measures. The inherent challenge of small sample sizes underscores the need for future studies with larger participant pools to draw more definitive conclusions.

8 Discussion

The results of this study revealed more distinct differences between minimum and maximum processing than between joint and blind processing, although the disparities remained statistically insignificant in subjective listening tests. Notably, this showed that the performance of minimum processing was not significantly inferior to that of maximum processing, since both the maximum and minimum processing methods significantly enhanced SI over unprocessed signals. This highlights the effectiveness of minimum processing approaches in SI enhancement.

The proposed joint minimum processing approach converges, by design, towards the performance of single-ended minimum processing methods in scenarios with high SNRs. Specifically, our method achieves comparable performance to the single-ended minimum processing methods [11] and [19] when there is minimal to no noise at either the NE or the FE. The adaptability of joint minimum processing is further emphasized by its ability to provide SI performance similar to joint maximum processing at low SNRs and when noise is present in both the FE and NE. In addition, for higher SNRs, where noise is absent at either end, our approach achieves a higher SQ. Consequently, the joint minimum processing method minimizes speech distortions while preserving a high SI. This indicates the beneficial application of the minimum processing principle in end-to-end communication scenarios

Existing studies on joint FE and NE enhancement [33–40] show varied results regarding the superiority of joint processing over blind methods. Although these studies compare a wide array of methods, they sometimes omit comparisons against blind methods or other blind and joint approaches of the same nature, leading to a lack of clarity regarding their implementations under consistent conditions. In our approach, we compared similar methods to investigate the specific gains from using a joint approach versus a blind approach, with the aim of eliminating the confounding effects of vastly different processing types. Hence, this work provides a clearer indication of the dynamics between joint and blind processing, as well as between minimum and maximum processing. Therefore, it is interesting that our results were inconclusive regarding the superiority of joint processing over blind processing. Despite conducting experiments in oracle situations, the distinctions between the two approaches were very subtle. Although trends showed that joint processing outperformed blind processing in objective SI, the differences were statistically insignificant in the subjective listening tests.

The limitation in increasing the joint performance over blind minimum processing can, in certain cases, be attributed to the number of feasible bands in the optimization problem. Hence, the performance constraints must be carefully considered in practical implementations. However, the absence of

feasibility issues in the maximum processing case indicates that the lack of performance difference between joint and blind maximum processing cannot be solely attributed to the optimization constraints.

We investigated the effect of joint processing when noise is present at both ends, because this is the only time that it makes sense to use a joint approach. However, conducting listening tests with noise at both the FE and NE is not very common. In [37] and [36], preference tests were used for both SI and SQ. [38] conducted an informal closed-matrix SI test. [35] and [34] conducted SI tests in which the words heard were typed in a computer interface. In [34] SQ was assessed only in quiet NE conditions. Hence, it is not clear what best practice is to judge especially SQ in the presence of both FE and NE noise. However, by using the MUSHRA test for SQ, we allowed the participants to freely judge what they deemed important in terms of SQ, under the assumption that they were familiar with the target speech, i.e., the reference signal. Hence, participants judged the importance of noise and speech distortions simultaneously. However, future studies may benefit from a clearer focus on which SQ degradations are caused by either (processed) environmental noise or speech distortions.

The results also show the impressive performance of the reference methods, suggesting limited potential for substantial improvement in terms of SI and SQ. Additionally, as SNRs improve at either end, the demand for joint FE and NE speech enhancement diminishes as the situation converges towards the single-ended cases. Generally, the better FSE methods are at removing noise in a distortionless manner by, for example, using DNNs [7–9] or an increased number of microphones [5], there is less need for joint SI and SQ enhancement.

9 Conclusion

We extensively explored the joint FE and NE minimum processing framework introduced in [40]. The primary contribution lies in deriving a closed-form analytical solution for the optimization problem, with an MSE processing penalty, an estimated SI constraint represented by ASII and an SQ noise power constraint. We provided a thorough explanation of the key elements and conducted a systematic performance study, including objective measures and listening tests for SI, listening effort, and SQ. Performance was compared to joint ASII maximization, the blind concatenation of minimum processing FSE and NLE, and the blind concatenation of classic maximum processing, and revealed nuanced results.

For estimated SI measured by ESTOI, maximum processing methods generally exhibit slightly superior performance, with the proposed joint minimum processing framework showing a slight edge over blind minimum processing. However, the results were inconclusive regarding the consistent superiority

of joint maximum processing over blind maximum processing in ESTOI. The PESQ results consistently show that minimum processing outperforms maximum processing, especially in high SNRs, but no significant differences were observed between the joint and blind methods. All subjective listening tests yielded inconclusive results. Subjective listening test results align with trends observed in objective measures, but fail to establish significant differences between maximum and minimum processing or between joint and blind processing. Hence, minimum processing performs on-par in SI with maximum processing while preserving a good SQ in higher SNR settings when noise is present at both ends. Additionally, because the joint minimum processing method has the single-ended solutions as special cases, the results and performance from the single-end minimum processing works extend to the joint case. This shows that it is also beneficial to apply the minimum processing principle in the context of end-to-end communication scenarios.

In essence, our work sheds light on the intricate relationship between SI, SQ, and the joint, blind, maximum and minimum processing methods. We provide in-depth insights into the optimization problem of joint minimum processing and underscore the importance of future investigations concerning optimization at both FE and NE in a joint context.

9.1 Future work

Our results provide valuable insights into this field, emphasizing the need for more thorough investigations with controlled implementations of joint and blind methods under identical conditions, for a deeper understanding of the behavioral aspects of blind and joint processing and a combined review of the performance of existing results.

Future work, directly pertaining to our method, includes investigating the conditions for feasibility in the optimization problem and how to derive an optimal performance in infeasible cases. Additionally, interesting extensions include other optimization targets, ANC, non-linear processing, and multiple FE environments with only one NE, together with a single FE broadcasting to multiple NE environments as in online meetings. It is also very interesting to investigate performance under more real-world conditions, with estimated speech and noise statistics, and in dimensions other than SI and SQ, such as complexity, synchronization requirements, and the need for bidirectional side-information transfer, are crucial. Particularly, when considering speech coding and how it can be incorporated into the joint end-to-end communication, where joint knowledge might be utilized in allocating limited bit rates to frequencies that are inaudible due to NE noise. Furthermore, it is interesting to investigate the impact of nonlinear processes between FE and NE, especially in the context of speech coding in real-life systems. Exploring the robustness of methods to coding added after the beamformer is also of interest. Similarly, the interplay

between the dependence of the proposed solution on frequency-band energy and automatic gain control warrants further exploration. In addition, the potential increase in bit rates for carrying enhanced signals needs consideration, and adjustments to the solutions may be necessary.

A Subband Filtering

Filtering frequency bins into subbands and determining the filter weights can be performed in various ways, cf. [11, 19, App. A]. In this paper, we consider auditory critical band filters [41] based on the gammatone filter bank model of [44].

As in [19], we let h_j be the impulse response of the j'th subband auditory filter. Now, for the j'th subband, the energy of the clean speech signal, $\mathcal{S}_{j,i}$, is given as the convolution between s and h_j, which in the time-subband domain is

$$\mathcal{S}_{j,i}^2 \triangleq \sum_{k\in\mathbb{B}_j} |S_{k,i}|^2 |H_j(k)|^2, \tag{D.39}$$

where $H_j(k)$ represents the DFT of h_j in frequency-bin k. We generate the frequency domain gammatone filters, $H_j(k)$, according to [44], and normalize them according to the mean total weight per frequency, that is.,

$$H_j'(k) = \frac{H_j(k)}{\frac{1}{K}\sum_m^K \sum_l^J H_l(m)}, \quad \forall j,k. \tag{D.40}$$

We then let the subband filter weights, $\omega_{j,k}$, be the normalized squared magnitude response of h_j, i.e., $\omega_{j,k} = |H_j'(k)|^2$.

When producing the final beamformers and NLE gains for each frequency bin, we need to apply a combination formula, such as that in (D.36). Because the weights, $\eta_{j,k}$, are applied to beamformer vectors and NLE gains, and not power spectra, we let the weights be given according to the normalized subband filter amplitudes, i.e., $\eta_{j,k} = |H_j'(k)|$.

B Processing penalty

B.1 Beamforming Cost

For far-end-only minimum processing speech enhancement, it was shown in [11, Sec. IV.A] that the minimum processing beamforming processing penalty is

$$\mathcal{D}_j(\boldsymbol{S}_j^R, \boldsymbol{Y}_j) = \sum_{k\in\mathbb{B}_j} \omega_{j,k} \left(\boldsymbol{v}_k^R - \boldsymbol{w}_k\right)^H C_{\mathrm{X}_k}^{(\mu)} \left(\boldsymbol{v}_k^R - \boldsymbol{w}_k\right). \tag{D.41}$$

Considering the difference to the reference beamformer in the cost function we have

$$\boldsymbol{v}_k^R - \boldsymbol{w}_{j,k} = \boldsymbol{v}_k^R - \left(\alpha_j \boldsymbol{v}_k^R + (1-\alpha_j)\boldsymbol{w}_k^{\mu\text{MWF}}\right) \tag{D.42}$$

$$= \left(1-\alpha_j\right)\left(\boldsymbol{v}_k^R - \boldsymbol{w}_k^{\mu\text{MWF}}\right). \tag{D.43}$$

Inserting this into the minimum processing beamforming processing penalty gives

$$\mathcal{D}_j(\boldsymbol{S}_j^R, \boldsymbol{Y}_j) = \sum_{k\in\mathbb{B}_j}\left[\omega_{j,k}\left(1-\alpha_j\right)^2\left(\boldsymbol{v}_k^R - \boldsymbol{w}_k^{\mu\text{MWF}}\right)^H C_{\text{X}_k}^{(\mu)}\left(\boldsymbol{v}_k^R - \boldsymbol{w}_k^{\mu\text{MWF}}\right)\right] \tag{D.44}$$

$$\propto \left(1-\alpha_j\right)^2. \tag{D.45}$$

B.2 Listening Enhancement Cost

The minimum processing NLE processing penalty was shown in [19] to be

$$\mathcal{D}_j(\boldsymbol{Y}_j, \boldsymbol{Z}_j) = \sum_{k\in\mathbb{B}_j}\omega_{j,k}\left(1-g_k\right)^2\sigma_{Y_k}^2. \tag{D.46}$$

Following the results of [19], where it was shown to be optimal to have fixed gains across the entire subband and to avoid comb filtering [40], we assume that the gains are equal across the subband. Therefore, we have that

$$\mathcal{D}_j(\boldsymbol{Y}_j, \boldsymbol{Z}_j) = \left(1-g_j\right)^2\sum_{k\in\mathbb{B}_j}\omega_{j,k}\sigma_{Y_k}^2 \propto \left(1-g_j\right)^2. \tag{D.47}$$

C Proof of Theorem 2

We reformulate and solve the optimization problem as a maximization problem with a simpler cost function. That is,

$$\begin{aligned} &\underset{\alpha_j, g_j\in\mathbb{R}_+}{\arg\max} \quad \alpha_j - g_j \\ \text{s.t. } &C_1 : g_j^2 p_{FSE}(\alpha_j) \geq \sigma_{N_j}^2 I_j^{\xi}, \qquad C_3 : 0 \leq \alpha_j \leq 1, \\ &C_2 : g_j^2 \delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}, \qquad C_4 : 1 \leq g_j. \end{aligned} \tag{D.48}$$

Looking at C_1 and C_2, we see that the optimization problem is only feasible if

$$\exists \alpha_j \in [0,1] \ : \ p_{FSE}(\alpha_j) > 0 \text{ and } \delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}. \tag{D.49}$$

That is, when there exists an α_j such that the FE SNR is above the desired audibility limit and the processed FE noise is below the upper noise limit. For convex optimization problems, the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient conditions that determine the global optimal solution, cf. [54]. However, we see from the constraints that we are optimizing over a non-convex set; hence, the optimization problem is non-convex. Therefore, the KKT conditions are not sufficient conditions for a global optimum to our optimization problem, but they are still necessary conditions [54]. Therefore, we investigate the KKT conditions to determine an optimum point.

Firstly, we formulate the Lagrangian,

$$\begin{aligned}\mathcal{L} = \alpha_j - g_j + \lambda_1\alpha_j + \lambda_2(1-\alpha_j) + \lambda_3(g_j - 1) \\ + \lambda_4\left(g_j^2 p_{FSE}(\alpha_j) - \sigma^2_{N_j} I_j^{\xi}\right) + \lambda_5\left(\sigma^2_{N_j} c_{U_j} - g_j^2 \delta_{U_j}(\alpha_j)\right).\end{aligned} \tag{D.50}$$

Determining the gradient and writing up the KKT conditions we get

$$g_j^2 p_{FSE}(\alpha_j) \geq \sigma^2_{N_j} I_j^{\xi}, \tag{D.51a}$$

$$g_j^2 \delta_{U_j}(\alpha_j) \leq \sigma^2_{N_j} c_{U_j}, \tag{D.51b}$$

$$\alpha_j \geq 0, \quad \alpha_j \leq 1, \quad g_j \geq 1, \tag{D.51c}$$

$$\lambda_1 \geq 0, \quad \lambda_2 \geq 0, \quad \lambda_3 \geq 0, \quad \lambda_4 \geq 0, \quad \lambda_5 \geq 0, \tag{D.51d}$$

$$\lambda_1\alpha_j = 0, \quad \lambda_2(\alpha_j - 1) = 0, \quad \lambda_3(g_j - 1) = 0, \tag{D.51e}$$

$$\lambda_4\left(g_j^2 p_{FSE}(\alpha_j) - \sigma^2_{N_j} I_j^{\xi}\right) = 0, \tag{D.51f}$$

$$\lambda_5\left(\sigma^2_{N_j} c_{U_j} - g_j^2 \delta_{U_j}(\alpha_j)\right) = 0, \tag{D.51g}$$

$$\begin{bmatrix} 1 + \lambda_1 - \lambda_2 + \lambda_4 g_j^2 p'_{FSE}(\alpha_j) - \lambda_5 g_j^2 \delta'_{U_j}(\alpha_j) \\ -1 + \lambda_3 + 2\lambda_4 g_j p_{FSE}(\alpha_j) - 2\lambda_5 g_j \delta_{U_j}(\alpha_j) \end{bmatrix} = \mathbf{0} \tag{D.51h}$$

To solve the optimization problem, we begin by determining the boundary and stationary solutions. Subsequently, we compare the different feasible solutions and select the one with the optimal cost function value as the optimal solution [54].

C.1 Solving KKT conditions

C.1.1 If $\lambda_1 > 0,\ \lambda_2 = 0,\ \lambda_3 > 0,\ \lambda_4 \geq 0,\ \lambda_5 \geq 0$

Then we must have $\alpha_j^* = 0$ and $g_j^* = 1$ and the cost function value is $f_{cost} = -1$. This is only a feasible solution if $p_{FSE}(0) = D_j^{\mu\text{MWF}} \geq \sigma^2_{N_j} I_j^{\xi}$ and $\delta_{U_j}(0) = \delta_{U_j}^{\mu\text{MWF}} \leq \sigma^2_{N_j} c_{U_j}$, i.e., the μMWF beamformer, $\boldsymbol{w}_k^{\mu\text{MWF}}$, provides a feasible FE SNR with a sufficiently low FE noise power.

C.1.2 If $\lambda_1 = 0,\ \lambda_2 > 0,\ \lambda_3 > 0,\ \lambda_4 \geq 0,\ \lambda_5 \geq 0$

Then we must have $\alpha_j^* = 1$ and $g_j^* = 1$ and the cost function value is globally maximized at $f_{cost} = 0$. This is only a feasible solution if $p_{FSE}(1) = D_j^R \geq \sigma_{N_j}^2 I_j^\xi$ and $\delta_{U_j}(1) = \delta_{U_j}^R \leq \sigma_{N_j}^2 c_{U_j}$, i.e., the reference beamformer $\boldsymbol{v}_k^R$ provides a feasible FE SNR with sufficiently low FE noise power.

C.1.3 If $\lambda_1 = 0,\ \lambda_2 = 0,\ \lambda_3 > 0,\ \lambda_4 \geq 0,\ \lambda_5 \geq 0$

Then we must have $g_j^* = 1$, and the cost function is $f_{cost} = \alpha_j - 1$. To ensure feasibility, it is necessary that the optimal α_j satisfies $p_{FSE}(\alpha_j) \geq \sigma_{N_j}^2 I_j^\xi$ and $\delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}$.

The shape of the polynomial $p_{FSE}(\alpha_j)$ defines a set of points

$$\mathcal{F}^{C_1} \triangleq \{\alpha \in [0,1] : p_{FSE}(\alpha_j) \geq \sigma_{N_j}^2 I_j^\xi\}, \tag{D.52}$$

that satisfy C_1. The polynomial may define the set to have one of five possible shapes, (a)–(e), as illustrated in Fig. D.5, which are

$$\begin{aligned}
&\mathcal{F}_{(a)}^{C_1} = [0,1], && \mathcal{F}_{(b)}^{C_1} = [0, \tilde{\alpha}_{b_0}] \cup [\tilde{\alpha}_{b_1}, 1], \\
&\mathcal{F}_{(c)}^{C_1} = [0, \tilde{\alpha}_{c_0}], && \mathcal{F}_{(d)}^{C_1} = [\tilde{\alpha}_{d_0}, \tilde{\alpha}_{d_1}], \\
&\mathcal{F}_{(e)}^{C_1} = [\tilde{\alpha}_{e_0}, 1].
\end{aligned} \tag{D.53}$$

Similarly, the shape of the polynomial $\delta_{U_j}(\alpha_j)$ defines a set of points,

$$\mathcal{F}^{C_2} \triangleq \{\alpha \in [0,1] : \delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}\}, \tag{D.54}$$

that satisfy C_2. Again there are five possible shapes to the set,

$$\begin{aligned}
&\mathcal{F}_{(1)}^{C_2} = [0,1], && \mathcal{F}_{(2)}^{C_2} = [0, \hat{\alpha}_{b_0}] \cup [\hat{\alpha}_{b_1}, 1] \\
&\mathcal{F}_{(3)}^{C_2} = [0, \hat{\alpha}_{c_0}], && \mathcal{F}_{(4)}^{C_2} = [\hat{\alpha}_{d_0}, \hat{\alpha}_{d_1}], \\
&\mathcal{F}_{(5)}^{C_2} = [\hat{\alpha}_{e_0}, 1].
\end{aligned} \tag{D.55}$$

Then,

$$\mathcal{F}^{g=1} = \mathcal{F}^{C_1} \cap \mathcal{F}^{C_2}, \tag{D.56}$$

is the jointly feasible region for both C_1 and C_2, and we have joint feasibility if $\mathcal{F}^{g=1} \neq \emptyset$, i.e, the feasible region is non-empty. Since $p_{FSE}(\alpha_j)$ and $\delta_{U_j}(\alpha_j)$ are second-order polynomials, we note that the feasible set, $\mathcal{F}^{g=1}$, is not necessarily

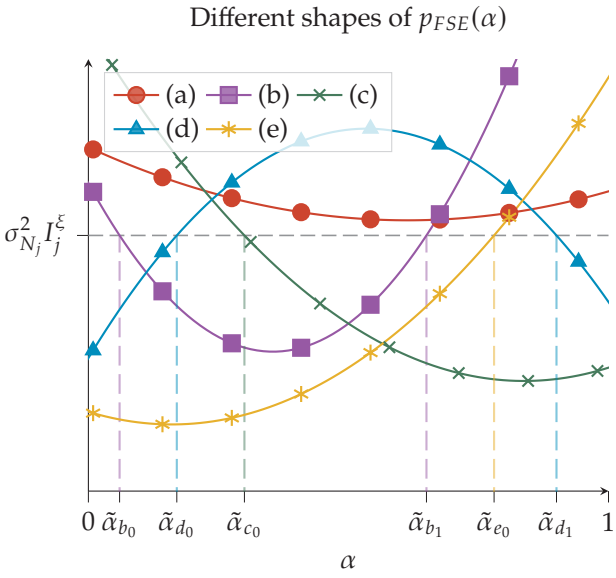


Fig. D.5: The figure illustrates the five different shapes and feasible regions that may be created by $p_{FSE}(\alpha)$.

a convex set. Now since the cost function, $f_{cost} = \alpha_j - 1$, is strictly monotonically increasing with α_j, the optimal α_j^* is the largest α in $\mathcal{F}^{g=1}$, i.e., the rightmost boundary of the feasible region, $\partial\mathcal{F}^{g=1}$. For example, if $\mathcal{F}^{g=1} = \mathcal{F}^{C_1}_{(d)} \cap \mathcal{F}^{C_2}_{(4)}$ and $\tilde{\alpha}_{d_0} \leq \hat{\alpha}_{d_0} \leq \tilde{\alpha}_{d_1} \leq \hat{\alpha}_{d_1}$, then $\mathcal{F}^{g=1} = [\hat{\alpha}_{d_0}, \tilde{\alpha}_{d_1}]$ and the optimal value is $\alpha_j^* = \tilde{\alpha}_{d_1}$.

C.1.4 If $\lambda_1 \geq 0,\ \lambda_2 \geq 0,\ \lambda_3 = 0,\ \lambda_4 \geq 0,\ \lambda_5 \geq 0$

Finally, we consider the interior stationary solution. Isolating g_j in the audibility constraint, C_1, we have

$$g_j \geq \sqrt{\frac{\sigma^2_{N_j} I^{\xi}_j}{p_{FSE}(\alpha_j)}}, \tag{D.57}$$

where $p_{FSE}(\alpha) > 0$. The only feasible root is the positive principal root since we must have $g_j \geq 1$. Combining the above limit with $g_j \geq 1$ we obtain

$$g_j = \max\left\{\sqrt{\frac{\sigma^2_{N_j} I^{\xi}_j}{p_{FSE}(\alpha_j)}}, 1\right\}. \tag{D.58}$$

We see that $g_j = 1$ if for the optimal α_j we have $p_{FSE}(\alpha_j^*) \geq \sigma^2_{N_j} I_j^{\xi}$, which we haved already solved in the previous case. Therefore, we focus on the case where $p_{FSE}(\alpha_j) < \sigma^2_{N_j} I_j^{\xi}\ \forall \alpha_j \in [0,1]$. In this case, the optimal g_j is $g_j^* = \sqrt{\frac{\sigma^2_{N_j} I_j^{\xi}}{p_{FSE}(\alpha_j^*)}}$. Inserting this into (P_0) and rearranging the terms of C_2, we find the optimal α_j^* by solving the optimization problem

$$\begin{aligned} &\underset{\alpha_j \in [0,1]}{\arg\max}\ \alpha_j - \sqrt{\frac{\sigma^2_{N_j} I_j^{\xi}}{p_{FSE}(\alpha_j)}} \\ \text{s.t.} \quad &C_{SI} : \sigma^2_{N_j} I_j^{\xi} > p_{FSE}(\alpha_j) > 0, \\ &C_{NP} : I_j^{\xi} \delta_{U_j}(\alpha_j) - c_{U_j} p_{FSE}(\alpha_j) \leq 0. \end{aligned} \tag{D.59}$$

First, we consider the conditions for feasibility. Similar to the above case, we can define

$$\mathcal{F}^{SI} \triangleq \left\{ \alpha \in [0,1] : \sigma^2_{N_j} I_j^{\xi} > p_{FSE}(\alpha) > 0 \right\} \tag{D.60}$$

as the set of points that satisfy C_{SI}. Additionally, for the second-order polynomial in the noise power constraint, we define the set of points that satisfy C_{NP},

$$\mathcal{F}^{NP} \triangleq \left\{ \alpha \in [0,1] : I_j^{\xi} \delta_{U_j}(\alpha) - c_{U_j} p_{FSE}(\alpha) \leq 0 \right\}. \tag{D.61}$$

Letting $\dot{\alpha}_l, \dot{\alpha}_r \in \mathbb{R}$ be the real roots of the polynomial, we have four possible regions for the noise power constraint:

$$\mathcal{F}^{NP}_{(i)} = [0,1], \qquad \mathcal{F}^{NP}_{(ii)} = [\dot{\alpha}_l, \dot{\alpha}_r], \tag{D.62}$$

$$\mathcal{F}^{NP}_{(iii)} = [0, \dot{\alpha}_l] \cup [\dot{\alpha}_r, 1], \qquad \mathcal{F}^{NP}_{(iv)} = \varnothing. \tag{D.63}$$

The jointly feasible set is then $\mathcal{F}^{g(\alpha)} = \mathcal{F}^{SI} \cap \mathcal{F}^{NP}$. As we have seen previously, this is not necessarily a convex set.

Secondly, letting

$$h(\alpha_j) \triangleq \alpha_j - \sqrt{\frac{\sigma^2_{N_j} I_j^{\xi}}{p_{FSE}(\alpha_j)}} \tag{D.64}$$

be the cost function, we investigate the concavity/convexity of the cost func-

tion. It can be shown that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}h(\alpha) = 1 + \frac{\sqrt{\sigma_{N_j}^2 I_j^{\xi}}\, p'_{FSE}(\alpha_j)}{2\big(p_{FSE}(\alpha_j)\big)^{3/2}} \tag{D.65}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}h(\alpha) = -\frac{3\sqrt{\sigma_{N_j}^2 I_j^{\xi}}\big(p'_{FSE}(\alpha_j)\big)^2}{4\big(p_{FSE}(\alpha_j)\big)^{5/2}} + \frac{\sqrt{\sigma_{N_j}^2 I_j^{\xi}}\, p''_{FSE}(\alpha_j)}{2\big(p_{FSE}(\alpha_j)\big)^{3/2}} \tag{D.66}$$

$$= -\frac{2\sqrt{\sigma_{N_j}^2 I_j^{\xi}}\,\phi(\alpha_j)}{4\big(p_{FSE}(\alpha_j)\big)^{5/2}}, \tag{D.67}$$

where

$$\begin{aligned}\phi(\alpha_j) \triangleq{}& \left(D_j^{cross} - D_j^R - D_j^{\mu\mathrm{MWF}}\right)^2 \alpha_j^2 \\ &- \left(D_j^{cross} - 2D_j^{\mu\mathrm{MWF}}\right)\left(D_j^{cross} - D_j^R - D_j^{\mu\mathrm{MWF}}\right)\alpha_j \\ &+ \left(D_j^{\mu\mathrm{MWF}}\right)^2 - \left(D_j^{cross} + \frac{D_j^R}{2}\right) D_j^{\mu\mathrm{MWF}} + \frac{3\left(D_j^{cross}\right)^2}{8}\end{aligned} \tag{D.68}$$

is a second-order polynomial in α_j. Because the optimization problem is only feasible for $p_{FSE}(\alpha_j) > 0$, the sign of the second derivative depends on the sign of $\phi(\alpha_j)$. Let $\overline{\alpha}_l \in \mathbb{R}$ and $\overline{\alpha}_r \in \mathbb{R}$ be the real roots of $\phi(\alpha_j)$, if they exist, where we assume $\overline{\alpha}_l \leq \overline{\alpha}_r$. Then, we may consider four different scenarios:

(*a*) If $\phi(\alpha) \geq 0\ \forall\ \alpha \in [0,1]$ then $h(\alpha)$ is concave in the entire interval $[0,1]$.

(*b*) If $\phi(\alpha) \geq 0\ \forall\ \alpha \in [0,\overline{\alpha}_l] \cup [\overline{\alpha}_r,1]$, and $\phi(\alpha) < 0\ \forall\ \alpha \in (\overline{\alpha}_l,\overline{\alpha}_r)$, then $h(\alpha)$ is concave in the intervals $[0,\overline{\alpha}_l]$ and $[\overline{\alpha}_r,1]$, and $h(\alpha)$ is convex in the interval $(\overline{\alpha}_l,\overline{\alpha}_r)$.

(*c*) If $\phi(\alpha) \geq 0\ \forall\ \alpha \in [\overline{\alpha}_l,\overline{\alpha}_r]$, and $\phi(\alpha) < 0\ \forall\ \alpha \in [0,\overline{\alpha}_l) \cup (\overline{\alpha}_r,1]$, then $h(\alpha)$ is concave in the interval $[\overline{\alpha}_l,\overline{\alpha}_r]$, and $h(\alpha)$ is convex in the intervals$[0,\overline{\alpha}_l)$ and $(\overline{\alpha}_r,1]$.

(*d*) If $\phi(\alpha) \leq 0\ \forall\ \alpha \in [0,1]$, then $h(\alpha)$ is convex in the entire interval $[0,1]$.

We note that the stationary points in the intervals where $h(\alpha)$ is concave are maxima, and the stationary points in the intervals where $h(\alpha)$ is convex are minima. Thus, the optimal points are either at the stationary points of the concave regions or at the boundary of the convex intervals. However, these optimal points may not necessarily be feasible.

Let α_s be a stationary point in a concave region of $h(\alpha)$. The stationary points can be determined by explicitly solving $h'(\alpha) = 0$ or via a simple bisection of $h'(\alpha)$ on the concave regions. If $\alpha_s \in \mathcal{F}^{g(\alpha)} \cap [0,1]$, then α_s is an optimal and feasible point. On the other hand, if $\alpha_s \notin \mathcal{F}^{g(\alpha)}$, i.e., the stationary point

is not feasible, then the optimal solution is at the boundary of the feasible set close to α_s. For example, if we are in the case of (a) and $\mathcal{F}_{(iii)}^{NP}$, then we might have $\dot{\alpha}_l < \alpha_s < \dot{\alpha}_r$, and α_s is not feasible. Therefore, the optimal feasible value in this case is found to be either $\dot{\alpha}_l$ or $\dot{\alpha}_r$.

Finally, let

$$\mathcal{F}_S^{g(\alpha)} = \{\alpha : \alpha \in \mathcal{F}^{g(\alpha)},\ h'(\alpha) = 0,\ h''(\alpha) \leq 0\} \tag{D.69}$$

be the set of feasible stationary points in the concave regions of $h(\alpha)$, and denote the boundary of $\mathcal{F}^{g(\alpha)}$ by $\partial\mathcal{F}^{g(\alpha)}$. Then, the optimal α_j is given as

$$\alpha_j^* = \arg\max_{\alpha}\ h(\alpha), \quad \text{s.t. } \alpha \in \mathcal{F}_S^{g(\alpha)} \cup \partial\mathcal{F}^{g(\alpha)}, \tag{D.70}$$

which is a simple combinatorial problem.

C.2 Comparing Cost functions

We now combine all the above cases such that we find the optimal solution by comparing the cost function values for the various optimum points:

$$f_{cost}(1,1) = 0 \tag{D.71}$$

$$f_{cost}(0,1) = -1 \tag{D.72}$$

$$f_{cost}(\alpha_j,1) = \alpha_j - 1 \tag{D.73}$$

$$f_{cost}\left(\alpha_j, \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{f_{con}(\alpha_j)}}\right) = \alpha_j - \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{f_{con}(\alpha_j)}} \tag{D.74}$$

First, if $D_j^R \geq \sigma_{N_j}^2 I_j^{\xi}$ and $\delta_{U_j}^R \leq \sigma_{N_j}^2 c_{U_j}$ then the optimal feasible cost function is $f_{cost}(1,1) = 0$, which is clearly the global optimum value. Therefore, in this case $(\alpha_j^*, g_j^*) = (1,1)$ regardless of the value of any of the other parameters.

We can cover all other cases by first solving the following combinatorial problem to determine the optimal α_j,

$$\alpha_j^* = \arg\max_{\alpha}\ \pi(\alpha), \quad \text{s.t. } \alpha \in \mathcal{F}_S^{g(\alpha)} \cup \partial\mathcal{F}^{g(\alpha)} \cup \partial\mathcal{F}^{g=1}, \tag{D.75}$$

where

$$\pi(\alpha) = \begin{cases} \alpha - 1 & \text{if } \alpha \in \partial\mathcal{F}^{g=1} \\ h(\alpha) & \text{if } \alpha \in \mathcal{F}_S^{g(\alpha)} \text{ or } \alpha \in \partial\mathcal{F}^{g(\alpha)}. \end{cases} \tag{D.76}$$

Finally, the optimal g_j is

$$g_j^* = \begin{cases} 1 & \text{if } \alpha_j^* \in \partial\mathcal{F}^{g=1} \\ \sqrt{\frac{\sigma_{N_j}^2 I_j^\xi}{f_{con}(\alpha_j^*)}} & \text{if } \alpha_j^* \in \mathcal{F}_S^{g(\alpha)} \text{ or } \alpha_j^* \in \partial\mathcal{F}^{g(\alpha)}. \end{cases} \tag{D.77}$$

This completes the proof.

References

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[2] M. Brandstein and D. Ward, Eds., *Microphone arrays: signal processing techniques and applications*, ser. Digital signal processing. New York: Springer, 2001.

[3] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic Beamforming for Hearing Aid Applications," in *Handbook on Array Processing and Sensor Networks*. John Wiley & Sons, Ltd, 2010, pp. 269–302.

[4] K. Eneman, H. Luts, J. Wouters, M. Büchler, N. Dillier, W. Dreschler, M. Froehlich, G. Grimm, V. Hohmann, R. Houben, A. Leijon, A. Lombard, D. Mauler, M. Moonen, H. Puder, M. Schulte, A. Spriet, and M. Vormann, "Evaluation of signal enhancement algorithms for hearing instruments," in *2008 16th European Signal Processing Conference*, Aug. 2008, pp. 1–5.

[5] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New Insights Into the MVDR Beamformer in Room Acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[6] Yiteng Huang, J. Benesty, and Jingdong Chen, "Analysis and Comparison of Multichannel Noise Reduction Methods in a Common Framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 957–968, Jul. 2008.

[7] K. Tesch and T. Gerkmann, "Insights Into Deep Non-Linear Filters for Improved Multi-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023.

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

[9] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[10] I. López-Espejo, A. Edraki, W.-Y. Chan, Z.-H. Tan, and J. Jensen, "On the deficiency of intelligibility metrics as proxies for subjective intelligibility," *Speech Communication*, vol. 150, pp. 9–22, May 2023.

[11] A. Zahedi, M. S. Pedersen, J. Østergaard, T. U. Christiansen, L. Bramsløw, and J. Jensen, "Minimum Processing Beamforming," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2710–2724, 2021.

[12] C. Chermaz, C. Valentini-Botinhao, H. Schepker, and S. King, "Near End Listening Enhancement in Realistic Environments," *Proceedings of the ICA 2019 and EAA Euroregio : 23rd International Congress on Acoustics*, vol. integrating 4th EAA Euroregio 2019 : 9-13 September 2019, pp. 5731–5735, 2019.

[13] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, May 2013.

[14] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[15] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Interspeech*, Lyon, France, 2013, pp. 3552–3556.

[16] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. N. Petkov, B. Sauert, and P. Vary, "Optimizing Speech Intelligibility in a Noisy Environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, Mar. 2015.

[17] B. Sauert, *Near-End Listening Enhancement: Theory and Application*, 1st ed., ser. Aachener Beiträge zu digitalen Nachrichtensystemen. Aachen: Wissenschaftsverlag Mainz, 2014, no. 36, oCLC: 880393716.

[18] C. H. Taal, J. Jensen, and A. Leijon, "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[19] A. J. Fuglsig, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof, and J. Østergaard, "Minimum Processing Near-End Listening Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2233–2245, 2023.

[20] H. Li and J. Yamagishi, "Multi-Metric Optimization Using Generative Adversarial Networks for Near-End Speech Intelligibility Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3000–3011, 2021.

[21] G. Li, H. Ruimin, R. Zhang, and X. Wang, "A mapping model of spectral tilt in normal-to-Lombard speech conversion for intelligibility enhancement," *Multimedia Tools and Applications*, vol. 79, no. 27-28, pp. 19 471–19 491, Jul. 2020.

[22] J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-Enhancing Speech Modifications — The Hurricane Challenge 2.0," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 1341–1345.

[23] G. Li, R. Hu, X. Wang, and R. Zhang, "A near-end listening enhancement system by RNN-based noise cancellation and speech modification," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 483–15 505, Jun. 2019.

[24] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise Intelligibility Improvement Based on Power Recovery and Dynamic Range Compression," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Portland, USA, Aug. 2012, pp. 2075–2079.

[25] N. V. George and G. Panda, "Advances in active noise control: A survey, with emphasis on recent nonlinear techniques," *Signal Processing*, vol. 93, no. 2, pp. 363–377, Feb. 2013.

[26] S. Kuo and D. Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87, no. 6, pp. 943–973, Jun. 1999.

[27] J. Rennies, A. Pusch, H. Schepker, and S. Doclo, "Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements," *The Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. EL315–EL321, Oct. 2018.

[28] R. Pricken, M. Wältermann, E. Parotat, M. Soloducha, and A. Raake, "Quality Aspects of Near-End Listening Enhancement Approaches in

Telecommunication Applications," in *Proceedings of DAGA 2017*. Kiel: German Acoustical Society (DEGA), 2017, pp. 872–875.

[29] Y. Tang, C. Arnold, and T. J. Cox, "A Study on the Relationship between the Intelligibility and Quality of Algorithmically-Modified Speech for Normal Hearing Listeners," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, vol. 1, no. 1, p. 10, Jun. 2018.

[30] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech & Language*, vol. 28, no. 4, pp. 858–872, Jul. 2014.

[31] B. Hect, J. Teevan, and Sellen, "The "Leaf Blower Problem" and the importance of common ground," Sep. 2021, Microsoft Research.

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

[33] K. Tan and D. Wang, "Improving Robustness of Deep Learning Based Monaural Speech Enhancement Against Processing Artifacts," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 6914–6918.

[34] T.-C. Zorilă and Y. Stylianou, "On the Quality and Intelligibility of Noisy Speech Processed for Near-End Listening Enhancement," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 2023–2027.

[35] M. P. Shifas, C. Zorilă, and Y. Stylianou, "End-to-End Neural Based Modification of Noisy Speech for Speech-in-Noise Intelligibility Improvement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 162–173, 2022.

[36] M. Niermann, P. Jax, and P. Vary, "Joint Near-End Listening Enhancement and far-end noise reduction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, pp. 4970–4974.

[37] H. Li, Y. Liu, and J. Yamagishi, "Joint Noise Reduction and Listening Enhancement for Full-End Speech Enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2023, pp. 1–5.

[38] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Intelligibility Enhancement Based on Mutual Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1694–1708, Aug. 2017.

[39] A. J. Fuglsig, J. Østergaard, J. Jensen, L. S. Bertelsen, P. Mariager, and Z.-H. Tan, "Joint Far- and Near-End Speech Intelligibility Enhancement Based on the Approximated Speech Intelligibility Index," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7752–7756.

[40] A. J. Fuglsig, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof, and J. Østergaard, "Joint Minimum Processing Beamforming and Near-end Listening Enhancement," in *IEEE 2024 Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. Seoul: IEEE, May 2024, p. 5.

[41] American National Standards Institute, *Methods for Calculation of the Speech Intelligibility Index*. New York, N.Y: Acoustical Society of America, 2017, vol. ANSI S.35-1997.

[42] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

[43] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.

[44] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A Perceptual Model for Sinusoidal Audio Coding Based on Spectral Integration," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1292–1304, Jun. 2005.

[45] EBU, "EBU recommendation R128 - Loudness Normalisation and Permitted Maximum Level of Audio Signals," European Broadcasting Union, Recommendation R-128, 2014.

[46] A. H. Andersen, "Speech Intelligibility Prediction for Hearing Aid Systems," PhD thesis, Aalborg University, 2017.

[47] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[48] E. Vincent, "MUSHRAM - A Matlab interface for MUSHRA listening tests," https://c4dm.eecs.qmul.ac.uk/downloads/#mushram, 2005, Accessed on 12-12-2023.

[49] ITU-R, "Recommendation BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Unition, Recommendation ITU-R BS.1534-3, Oct. 2015.

[50] W. Kruskal and W. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[51] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[52] C. Mendonça and S. Delikaris-Manias, "Statistical Tests with MUSHRA Data," *Journal of the Audio Engineering Society*, no. 10006, May 2018.

[53] A. Field, *Discovering statistics using IBM SPSS statistics*, 5th ed., ser. SAGE edge. SAGE, 2018.

[54] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge University Press, 2004.