DOCTORAL THESIS

Online multichannel speech enhancement combining statistical signal processing and deep neural networks

University of Granada

Author:
Juan Manuel Martín Doñas

Thesis supervisors:
Antonio Miguel Peinado Herreros
Ángel Manuel Gómez García

Ph.D. Program in Information and Communication Technologies
Department of Signal Theory, Telematics and Communications

Granada, November 2020

Editor: Universidad de Granada. Tesis Doctorales
Autor: Juan Manuel Martín Doñas
ISBN: 978-84-1306-753-7
URI: http://hdl.handle.net/10481/66402

A mis padres, Juan e Isabel
por su apoyo incondicional y
por hacerme ser mejor persona.

Acknowledgements

En primer lugar, me gustaría agradecer a mis directores de Tesis, Antonio Peinado y Ángel Gómez, por toda su ayuda y guía durante el desarrollo de mi investigación y por su gran valor humano. Esta Tesis doctoral no habría sido posible sin todo su apoyo. También quiero expresar mi gratitud por los miembros del grupo SigMAT, entre ellos Victoria, José Luís, José Andrés y Juan Andrés, que siempre me han ayudado en todo lo que les he pedido. También a mis compañeros de laboratorio Alejandro y Amelia, a quienes les deseo mucho éxito en estos últimos años de doctorado que les quedan. Por último, quiero hacer una mención especial a mi gran amigo Iván López, quien ha sido un compañero y mentor durante todos estos años y que siempre ha estado dispuesto a echarme una mano, esta Tesis también te la debo a ti.

I want to express my gratitude to the Department of Communications Engineering, especially to professor Reinhold Haeb-Umbach, for their warm welcome during my research stay at Paderborn University. Also, I would like to thank Jens for his support in the project development, and Cristoph and Lukas for our interesting discussions, especially with a beer in our hands. In addition, I am very grateful to the Section for Signal and Information Processing of the Department of Electronic Systems, especially to the professors Zheng-Hua Tan and Jesper Jensen, for the support received in my research stay at Aalborg University. In particular, I would like to thank Morten Kolbaek for the brainstorming sessions on single-channel speech enhancement, and also to my lab partners Payam and Yuying for making the faculty my second home in Aalborg. Finally, a special mention to the other students that I met at Aalborg University, who made me enjoy my stay experience. To Cristian, Adria, Andrea, and Alex, I hope you will achieve all your goals.

A mis padres, Juan e Isabel, que tanto me han dado en todos estos años, por su apoyo, esfuerzo y consejos que tanto me han aportado en mi desarrollo personal como profesional. A mi hermano Alejandro, quien sabe que le tengo un especial cariño. También a mis otros familiares, especialmente a mis abuelos, y a los que ya no estáis con nosotros siempre os tendremos en el corazón. Estos años de dedicación a mi Tesis tampoco habrían sido posibles sin todos los amigos que me han acompañado. A Ernesto y David, dos grandes amigos y compañeros de carrera, y unos auténticos caballeros, por todo lo que me han aportado y todas las risas que les debo. A los hermanos Julián y Martín, amigos casi desde que tengo uso de razón, por todos esos momentos y los que nos quedan por disfrutar. También mencionar a Irene, Arturo y Lucía por sus ánimos y su confianza en que superaría esta etapa. En general, a todos esos amigos que tanto me han aportado durante toda esta etapa, aunque no os pueda mencionar a todos parte de esta Tesis también os pertenece. Y finalmente, pero no menos importante, a ti Raquel, por todo el cariño, apoyo y comprensión que me has dado en esta etapa final y que me ha permitido llegar hasta aquí, gracias de corazón.

Abstract

Speech-related applications on mobile devices require high-performance speech enhancement algorithms to tackle challenging real-world noisy environments. These speech processing techniques have to ensure good noise reduction capabilities with low speech distortion, thus improving the perceptual speech quality and intelligibility of the enhanced speech signal. In addition, current mobile devices often embed several microphones, allowing them to exploit the spatial information during the enhancement procedure. On the other hand, low latency and efficiency are requirements for extensive use of these technologies. Among the different speech processing paradigms, statistical signal processing offers limited performance under non-stationary noisy environments, while deep neural networks can lack generalization under real conditions.

The main goal of this Thesis is the development of online multichannel speech enhancement algorithms for speech services in mobile devices. The proposed techniques use multichannel signal processing to increase the noise reduction performance without degrading the quality of the speech signal. Moreover, deep neural networks are applied in specific parts of the algorithm where modeling by classical methods would be, otherwise, difficult or very limited. This allows for the use of more capable deep learning methods in real-time online processing algorithms. Our contributions focus on different noisy environments where these mobile speech technologies can be applied.

First, we develop a speech enhancement algorithm suitable for dual-microphone smartphones used in noisy and reverberant environments. The noisy speech signal is processed using a beamforming-plus-postfiltering strategy that exploits the dual-channel properties of the clean speech and noise signals to obtain more accurate acoustic parameters. Thus, the temporal variability of the relative transfer functions between acoustic channels is tracked by using an extended Kalman filter framework. Noise statistics are obtained by means of a recursive procedure using the speech presence probability. This speech presence is estimated through either statistical spatial models or deep neural network mask estimators, both exploiting dual-channel features from the noisy speech signal.

Then, we propose a recursive expectation-maximization framework for online multichannel speech enhancement. The goal is the joint estimation of the clean speech statistics

and the acoustic model parameters in order to increase robustness under non-stationary conditions. The noisy speech signal is first processed using a beamformer followed by a Kalman postfilter, which exploits the temporal correlations of the speech magnitude. The speech presence probability is then obtained using a deep neural network mask estimator, and its estimates are further refined through statistical spatial models defined for the noisy speech and noise signals. The resulting clean speech and speech presence estimates are then employed for maximum-likelihood estimation of beamformer and postfilter parameters. This also allows for an iterative procedure with positive feedback between the estimation of speech statistics and acoustic parameters.

Scenarios with multiple overlapped speakers are also analyzed in this Thesis. Thus, beamforming with the model parameters obtained from deep neural network mask estimators is also explored. To deal with interfering speakers, we study the use of adapted mask estimators that exploit spectral and spatial information, obtained through auxiliary information, to focus on a target speaker. Therefore, additional speech processing blocks are integrated into the mask estimators so that the network can discriminate among different speakers. As an application, we consider the problem of automatic speech recognition in meeting scenarios, where our proposal can be used as a front-end processing.

Finally, we study the training of deep learning methods for speech processing using perceptual considerations. Thus, we propose a loss function based on a perceptual quality objective metric. We evaluate the proposed loss for training deep neural network-based single-channel speech enhancement algorithms in order to improve the speech quality perceived by human listeners. The two most common approaches for single-channel processing using these networks are considered: spectral mapping and spectral masking. We also explore the combination of different objective metric-related loss functions in a multi-objective learning training approach.

To conclude, we would like to highlight that our contributions successfully integrate signal processing and deep learning methods to jointly exploit spectral, spatial, and temporal speech features. As a result, the set of proposed techniques provides us with a manifold framework for robust speech processing under very challenging acoustic environments, thus allowing us to improve perceptual quality, intelligibility, and distortion measures.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

This chapter serves as an introduction to the topic covered in this Thesis, which is intended for the research on online multichannel speech enhancement algorithms that combines the use of statistical signal processing and deep neural networks. The motivation of this Thesis and an overview of the current state of this area are first presented in Section 1.1. Then, the objectives pursued in this Thesis are summarized in Section 1.2. Section 1.3 describes the structure organization on the different chapters. Finally, in Section 1.4 we enumerate the different publications derived from this Thesis.

1.1 Motivation and overview

Speech is surely the main and most relevant method of communication between humans. It allows us to express our ideas, interchange information with other people, and it is a fundamental tool in our society. Speech communications have been favored in the last decades thanks to the advent of the information and communication technology era. The broadcasting technologies, such as the radio, television, or the Internet, make it possible to quickly access the information, which in great part is given through speech. Telephone communications allow to converse with people at long distances, and mobile devices have spread human communications ubiquitously and pervasively. Communication services using the Internet have also gained importance in the last years using computer applications such as Skype or Discord, or even mobile applications like WhatsApp or Telegram are commonly employed, keeping a continued interconnection with people across the world.

Technology improvements have also propitiated advances in the area of human-machine interactions. Good examples are the digital assistant devices that can interact with humans to accomplish different tasks. These assistants are included in the majority of our mobile devices, as well as new smart speakers have been developed by different companies as in the case of Amazon Alexa, Apple Siri, or Google Home. These human-machine communications require speech technologies such as automatic speech recognition and text-to-speech synthesis. Another important aspect is the security in speech technologies, where robust speaker verification and anti-spoofing technologies are needed to ensure the user identity (e.g. operations in an electronic bank using voice biometrics). Moreover, speech technologies have found great utility in different health services, such as hearing-aids devices or silent speech interfaces. As we can see, speech technologies can be found in different aspects of our life and they are expected to grow up in the following years.

One of the main challenges for these speech technologies is their use in conditions where the speech signal is affected by different kinds of distortion. These distortions can come from different sources, such as environmental noise, which is the main problem in the use of mobile devices, or interfering speech from other speakers. Another kind of distortions are those due to the acoustics properties of the environment, as in the case of echoes or reverberation. These different distortion sources can severely degrade the performance of the previously described technologies. For example, they make mobile communications between humans less intelligible and affect the perceptual speech quality, which is especially problematic for hearing-impaired listeners. Moreover, the performance of recognition and verification systems drops under severe speech distortions, hindering its use in challenging noisy conditions.

The use of speech technologies in noisy conditions demands high-performance speech processing algorithms capable of improving speech quality and intelligibility. This is the goal of speech enhancement, which deals with the design of techniques to estimate clean speech from noisy and distorted speech. Speech enhancement is essential in many speech-related technologies in order to provide good performance in real environments. Therefore, research in different methods and algorithms for speech enhancement is a crucial and still challenging field. The first works in this area range from the design of heuristics algorithms to the use of statistical frameworks to model the properties of the underlying signals involved. The use of statistical estimators, along with assumptions about the noise, allowed the design of techniques with competitive performance, especially in stationary conditions. The research of more powerful statistical models offered solutions with a high potential, which could be integrated into speech technologies thanks to the increased computational capabilities. In addition, current devices have started to incorporate arrays of multiple microphones to capture the speech signal. This propitiates the interest in multichannel speech processing algorithms that exploit the spatial information from the different microphones, thus improving the noise reduction while achieving low speech distortion. The combination of multichannel techniques with statistical frameworks gave state-of-the-art results in multiple applications, especially in automatic speech recognition systems. Nevertheless, classical signal processing faces additional limitations in very challenging scenarios, such as those involving non-stationary noises, interfering speakers, and severe distortion due to reverberant environments. In this case, the assumptions about the signals are not accurate enough so the modeling and, therefore, the final performance, degrades. This results in enhanced speech signals with poor quality and low intelligibility for both humans and machines.

In recent years, the deep learning revolution has changed most of the current human technologies. Deep learning allows for the design of algorithms that can be trained to learn directly from the data on how to perform their tasks. This shares a similarity, in a general view, with the way humans learn from their surroundings. Nowadays, deep neural networks are complex models that include several layers of non-linearities and millions of parameters that have to be learned. Two main factors have favored the boom of these machine learning algorithms. The first one is the increase of computational resources, especially with the improvements in graphical processing units or GPUs, that allows for parallel training of these models in less amount of time. The second, and probably the most important, is the availability and variety of a huge amount of training data, which increase these networks' generalization capabilities. Deep neural networks have certainly become a state-of-the-art technology in many different areas, including speech processing, where they have outperformed classical approaches. Current speech recognition, verification, and synthesis algorithms are designed using deep neural networks due to its astounding performance. Deep learning has also been applied to speech enhancement, providing enhanced speech signals with high perceptual quality, low distortion, and great noise removal capabilities. However, one of the main criticism of deep neural networks is that they act as black-boxes, where it is almost impossible to know how the signals are processed and how the algorithm learns. The need for a high amount of parameters and large databases is another limitation of these techniques, as we do not have control over how we are sizing our problem and the way these models will generalize in real conditions. Finally, these models do not need engineering knowledge and specific modeling of the problem, meaning we are wasting the accumulated experience in speech research. This experience can be still useful to certain problems or for a better understanding of the problem to be solved.

A final important aspect of speech technologies in current smart devices is that they also need to ensure computational efficiency and online processing (i.e. using current and past information) with low-latency. While the requirement of efficient algorithms is needed for the integration of these technologies in a wide variety of devices, online processing is still needed for real-time applications with a suitable quality service. The design of online processing algorithms is generally more difficult and its performance is usually lower than offline techniques. Therefore, the research on online speech enhancement and the use of low-complexity algorithms is yet another key point to be studied by the speech community.

1.2 Objectives of this Thesis

As we have introduced, speech enhancement algorithms are needed in mobile devices applications to improve the perceptual speech quality and intelligibility in non-stationary noisy conditions. Current devices embed microphone arrays, so multichannel information can also be exploited. On the other hand, speech-related applications in mobile devices have to ensure online processing with low-latency and computational efficiency. Among the speech enhancement algorithms in the literature, classical signal processing is limited due to the assumptions made about the signal statistics, which are often unrealistic, while deep neural networks are black-boxes that require large amounts of data and parameters, and can lack generalization. This Thesis is focused on the development of techniques for online multichannel speech enhancement suitable for mobile devices. The proposed algorithms are designed to integrate the use of statistical signal processing and deep neural networks in specific parts of the algorithm pipeline. Thus, we can take advantage of multichannel signal processing to develop powerful speech enhancement techniques with high performance and low distortion. In addition, more efficient deep neural networks can be used in parts of the algorithm where assumptions about the statistics and properties of the signals are weak. This can increase the robustness under challenging real-world non-stationary noisy environments while allowing for online processing. More precisely, we highlight the following objectives, each focusing on a different scenario to apply these integrated techniques:

1. To develop speech enhancement algorithms suitable for dual-microphone smartphones in noisy and reverberant environments. Our goal is to exploit the particular relationship between the clean speech and the noise signals at both sensors, achieving a more accurate estimation of the acoustic channels and noise statistics.

2. To study the joint estimation of the clean speech signal and the different speech statistics and acoustic parameters in an online multichannel speech enhancement framework. The idea is to increase the robustness under non-stationary noises by jointly exploiting the spectral, spatial, and temporal characteristics of the speech signal.

3. To improve the performance of speech enhancement algorithms in scenarios with multiple overlapped speakers. Thus, the goal is to focus on a target speaker using auxiliary information from him/her, then simplifying the problem to a noisy environment scenario.

4. To analyze and evaluate the training of deep neural networks for speech enhancement using perceptual considerations of the human auditory system. Thus, our objective is to evaluate well-known objective quality metrics as training functions, improving the quality perceived by human listeners.

1.3 Thesis organization

This Thesis is comprised of a total of eight chapters, including this introductory chapter. A comprehensive summary in Spanish is also included in order to fulfill the requirements of the University of Granada regarding the drafting of doctoral dissertations. The theoretical foundations of this Thesis and a review of the state-of-the-art are developed in Chapter 2, while the experimental framework is described in Chapter 3. Then, Chapters 4, 5, 6 and 7 are devoted to describe our contributions on online multichannel speech enhancement. Each chapter develops one of the previously enumerated objectives of this Thesis. Finally, in Chapter 8 the final conclusions are summarized. More specifically:

- In Chapter 2, a review of the speech enhancement literature is carried out to present the theoretical fundamentals of this Thesis. First, we introduce the analysis and processing of the noisy speech signal in the time-frequency domain using the short-time Fourier transform. Then, the single-channel algorithms based on classical signal processing are reviewed, remarking the problem of noise estimation. Next, multichannel speech enhancement approaches based on beamforming algorithms are explained along with the use of postfiltering techniques and the estimation of the needed acoustic parameters. To conclude, we overview the use of deep neural networks for speech enhancement. The most common network architectures are summarized, and the use of these models for single-channel and multichannel speech enhancement is discussed.

- The experimental framework used in this Thesis is described in Chapter 3. This includes the noisy speech databases and the objective quality metrics used for the training and evaluation of the proposed contributions. In addition, we detail the setup followed in the training of the deep neural networks that are integrated into our proposals.

- In Chapter 4, a speech enhancement algorithm intended for dual-microphone smartphones is proposed. This approach exploits the dual-channel information and the mode of use of the smartphone to obtain more accurate acoustic model parameters. We make first a general description of our approach, which is based on a beamforming-plus-postfiltering architecture. Then, we describe our extended Kalman filter framework to track the time-variability of the acoustic function between microphones. Finally, noise estimation is addressed using speech presence probability. Two approaches are considered: statistical spatial models and deep neural networks mask estimators. The proposals are evaluated in a dual-channel noisy and reverberated database obtained from a smartphone used in close-talk and far-talk positions.

- A recursive expectation-maximization framework for online multichannel speech enhancement is proposed in Chapter 5. This framework allows the joint estimation of the clean speech signal, the speech presence probability, and the different acoustic parameters in an iterative way, increasing the robustness in non-stationary noisy environments. A beamformer is first used to exploit the spatial information of the noisy speech signals. Then, a Kalman postfilter uses the temporal correlations in the clean speech signal to further enhance the noise reduction performance. The speech presence probability is estimated using a model that combines a statistical spatial model with a deep neural network mask estimator. Finally, the estimated statistics are used for maximum-likelihood estimation of the acoustic model parameters. Our proposal is evaluated in a multichannel noisy speech database recorded with a table in different real-world environments.

- In Chapter 6, an approach for target speaker separation in a multiple speaker scenario is proposed. This approach allows focusing on one speaker using a deep neural network mask estimator that integrates auxiliary information for the desired speaker. To this end, the network is improved with additional blocks that exploit the spectral and spatial characteristics of the speaker. The mask estimator is used along with block-online beamforming, which is initialized with the contextual information to improve the system convergence. The proposal is evaluated for automatic speech recognition in meeting scenarios with overlapped speakers.

- A deep learning loss function for the perceptual evaluation of the speech quality is proposed in Chapter 7. This novel loss function is derived from the perceptual evaluation of the speech quality algorithm, which is a well-known objective quality metric. The approach is intended for the training of deep neural networks using perceptual considerations, improving the speech quality perceived by human listeners. Our proposal is evaluated for deep neural network-based single-channel speech enhancement, considering the two most common approaches: spectral mapping and spectral masking.

- Finally, the conclusions of this Thesis are presented in Chapter 8 along with a summary of our contributions and future work.

1.4 List of publications

The following publications have been produced as a result of the work in this Thesis:

1. **J. M. Martín-Doñas**, J. Jensen, Z. H. Tan, A. M. Gomez and A. M. Peinado. *Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation*. IEEE/ACM Transactions on Audio, Speech, and Language Processing. *Early Access*, 2020.
   IF (JCR 2019): 3.398. Acoustics. Rank 5/32 (Q1).

2. **J. M. Martín-Doñas**, A. M. Peinado, I. López-Espejo and A. M. Gomez. *Dual-channel speech enhancement based on extended Kalman filter relative transfer function estimation*. Applied Sciences, 9:2520, 2019.
   IF (JCR 2019): 2.474. Engineering, Multidisciplinary. Rank 32/91 (Q2).

3. **J. M. Martín-Doñas**, A. M. Gomez, J. A. Gonzalez and A. M. Peinado. *A deep learning loss function based on the perceptual evaluation of the speech quality*. IEEE Signal Processing Letters, 25(11):1680-1684, 2018.
   IF (JCR 2018): 3.268. Engineering, Electrical & Electronic. Rank 79/266 (Q2).
   Note: This work was also presented in the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 12-17, Brighton (United Kingdom), 2019.

4. **J. M. Martín-Doñas**, A. M. Peinado, I. López-Espejo and A. M. Gomez. *Dual-channel eKF-RTF framework with DNN-based speech presence estimation*. Submitted to IberSpeech 2020.

5. **J. M. Martín-Doñas**, J. Heitkaemper, R. Haeb-Umbach, A. M. Gomez and A. M. Peinado. *Multi-channel block-online source extraction based on utterance adaptation*. In Proceedings of 20th Annual Conference of the International Speech Communication (InterSpeech), pp. 96-100, September 15-19, Graz (Austria), 2019.

6. **J. M. Martín-Doñas**, I. López-Espejo, A. M. Gomez and A. M. Peinado. *A postfiltering approach for dual-microphone smartphones*. In Proceedings of IberSpeech 2018, pp. 142-146, November 21-23, Barcelona (Spain), 2018.

7. **J. M. Martín-Doñas**, I. López-Espejo, A. M. Gomez and A. M. Peinado. *An extended Kalman filter for RTF estimation in dual-microphone smartphones*. In Proceedings of 2018 26th European Signal Processing Conference (EUSIPCO), pp. 2474-2478, September 3-7, Rome (Italy), 2018.

Moreover, other additional publications and collaborations have been done during the development of this Thesis, whose topics are directly related to the research on online multichannel speech enhancement. These publications are not directly included in this Thesis, but they have contributed to the doctoral research:

1. M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F. R. Stöter, M. Hu, **J. M. Martín-Doñas**, D. Ditter, A. Frank, A. Deleforge, E. Vincent. *Asteroid: the PyTorch-based audio source separation toolkit for researchers*. Accepted at 21st Annual Conference of the International Speech Communication (InterSpeech), October 25-29, Shangai (China), 2020.

2. I. López-Espejo, **J. M. Martín-Doñas**, A. M. Gomez and A. M. Peinado.
Unscented transform-based dual-channel noise estimation: Application to speech enhancement on smartphones. In Proceedings of 2018 41st International Conference on Telecommunications and Signal Processing (TSP), pp. 88-91, July 4-6, Athens (Greece), 2018.

3. **J. M. Martín-Doñas**, A. M. Gomez, I. López-Espejo and A. M. Peinado.
Dual-channel DNN-based speech enhancement for smartphones. In Proceedings of 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), pp. 1-6, October 16-18, Luton (United kingdom), 2017.

Chapter 2

Fundamentals of speech enhancement in the STFT domain

Speech enhancement is a fundamental element of many different speech-related applications used in our daily life. The main objective of speech enhancement is the improvement of the perceptual quality and intelligibility of the speech signal which has been degraded due to distortions. Many sources of distortion can affect the clean speech signal: environmental noises that contaminate the speech signal, echoes and reverberations that deform the original waveform, or interfering speakers that overlap with the target speaker. These degradations directly affect the spectral characteristics of the speech signal, making it less intelligible for human listeners or difficult to process for computer programs.

Most of the research on speech enhancement focus on the problem of corrupted speech by additive environmental noise, which is commonly known as the noise reduction problem. Even though this topic has been extensively studied in the past last decades, noise reduction remains a difficult task due to many reasons. In the first place, the noise is highly variable across different scenarios or even during a period of time. For example, street noise and restaurant noise have different spectral characteristics. Moreover, most environments present highly non-stationary noise signals whose statistics can be difficult to estimate. This makes even more difficult the design of speech processing algorithms that provide good performance in a wide variety of environments and conditions. On the other hand, there are no clear criteria on how to evaluate the performance of the enhancement techniques. This performance can be evaluated in terms of the improved speech quality, or the final intelligibility perceived by humans, or even in the recognition accuracy achieved by a machine. These criteria do not necessarily correlate among themselves. For example, the design of algorithms that improve a specific aspect, as the perceptual quality, can yield to degradation in another aspect, as speech intelligibility. Therefore, speech enhancement is one of the main topics in the speech processing research area and a notorious element in most of the speech processing pipelines.

A wide variety of speech enhancement algorithms deal with single-channel scenarios. In this case, the noisy speech signal, composed by the clean speech signal and the different degradations and interferences, is captured by only one microphone. This noisy speech signal is presented as a time sequence of samples, converted from an analog signal at a given sampling rate. Despite its simplicity, this representation does not provide refined information about the spectral characteristics of the underlying signals, hindering the estimation of the clean speech signal. Therefore, speech enhancement algorithms are usually implemented in a transformed domain [125]. Among the different existing transformations proposed in the speech processing field, the use of the frequency domain via the Fourier transform is the most common one for speech enhancement [162] because of its simple implementation and its resemblance to the human auditory system processing. Moreover, these algorithms usually work with segmented frames of the original time signal, which are processed sequentially and transformed back again to the time domain. Different approaches have been proposed for single-channel speech enhancement in the frequency transformed domain [125, 162]. From the first approaches based on heuristics, the techniques evolved to the use of filtering methods and then the emerge of statistical model-based approaches, most of them following a Bayesian estimation framework. These methods require knowledge and assumptions about the statistics of the clean speech and noise signals, that must be estimated in advance. The availability of devices embedding arrays with two or more microphones auspicious the development of multichannel speech enhancement methods [57]. These approaches also allow us to exploit the spatial information of the signals: localization of the target speaker, characteristics of the acoustic channel, or the statistical properties of the noise field across the microphones. These approaches can provide increased performance in noise reduction without distorting the speech signal, especially when the target speaker and the interfering sources are spatially-separated. Nevertheless, the performance of these methods depends on the array configuration, the spatial characteristics of the signals, and the accurate estimation of the acoustic parameters involved. In addition, these techniques suffer from additional distortion in reverberant environments.

A more recent approach considers speech enhancement as a supervised learning problem. Thus, a machine learning algorithm is trained using a data-driven approach to estimate the clean speech signal or other target features used for its estimation. Supervised speech enhancement approaches have benefited in the last years from the quick progress of deep learning [224], which has achieved astounding results in different technology areas, including speech processing. Deep neural networks (DNN) have shown good results in single-channel speech enhancement, outperforming classical statistical approaches, and standing as state-of-the-art algorithms. These methods have moved towards the use of more powerful network architectures that considers the different aspects of the processing pipeline, from significant feature extraction to the reconstruction of the time signal. The DNN-based approaches have also been explored in multichannel speech enhancement from different perspectives: the direct estimation of the clean speech signal using multichannel features, the estimation of important acoustic parameters for the multichannel techniques, or the integration in most complex multichannel algorithms. Currently, the DNN research focuses on increasing the performance in more challenging scenarios with low SNRs, highly non-stationary noises or interfering sources, and real-world applications.

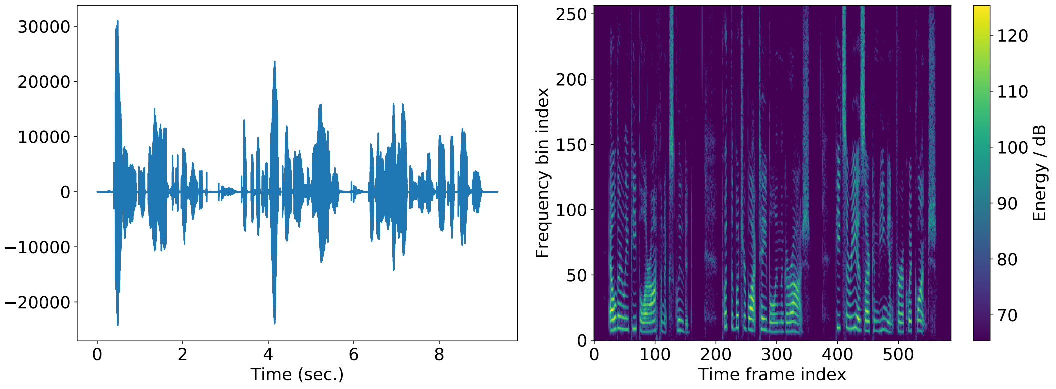
This chapter reviews the literature in speech enhancement processing, bringing up the main concepts involved in the proposals presented in the following chapters. The remainder of the chapter is structured as follows. Section 2.1 introduces the short-time Fourier transform (STFT), that is used for the analysis and processing of the noisy speech signal. Then, the classical single-channel speech enhancement approaches in the STFT domain are reviewed in Section 2.2. These techniques require knowledge about the noise statistics, so different noise estimation techniques will be also presented in that section. Section 2.3 covers the multichannel approaches based on beamforming algorithms, which are the most common techniques for array processing. Different beamforming approaches and the estimation of the needed acoustic parameters are reviewed in that section. Finally, the use of DNNs for speech enhancement will be presented in Section 2.4, including an overview of the main architectures, and the most common DNN-based single-channel and multichannel speech enhancement approaches.

2.1 Analysis and processing of the noisy speech signal

We first consider a scenario where the clean speech signal from a target speaker is affected by an additive noise signal, composed of different surrounded sounds from the environment. The distortion model in the discrete-time domain is simply described as,

$$y(m) = x(m) + n(m), \tag{2.1}$$

where $x(m)$, $n(m)$, and $y(m)$ are, respectively, the clean speech, the noise and the noisy speech signals, and m is the sample index. Traditionally, the temporal domain has not been convenient for the analysis of the speech signal, as its main characteristics are better represented in the frequency domain: the predominant frequencies, the formants due to the



(a) Time signal (b) STFT signal

Fig. 2.1 Example of the STFT processing applied to a clean speech signal. The spectrogram (magnitude of the STFT) is represented. The time signal is sampled at 16 kHz and the STFT is computed using a 512-point DFT with 50% overlap.

vocal tract shape, or the pitch are some examples. On the other hand, the speech signal is non-stationary and its spectral characteristics change on time. Then, it is better to analyze the speech in short segments (10 - 20 milliseconds) as, fortunately, these properties change slowly.

The previous properties indicate that a time-frequency analysis of the speech signal is more appropriate to represent its spectral and non-stationary characteristics. The short-time Fourier transform (STFT) [7, 9] is commonly used for different speech processing tasks. For example, the STFT of the clean speech signal is defined as

$$X(t,f) = \sum_{m=0}^{N_s-1} x(m+tl_s)w(m)e^{-j\frac{2\pi}{N_s}mf}, \tag{2.2}$$

where $w(m)$ is a window function, N_s is the frame length in samples, l_s the frame shift in samples, and t and f are the time frame and frequency indices, respectively. The idea is to divide the signal into overlapped frames which are multiplied by the window function $w(m)$. Then, the discrete Fourier transform (DFT) is applied to each frame. An example of a clean speech discrete-time domain signal and its STFT representation is depicted in Fig. 2.1.

Generally, we can reconstruct a time-sequence from its STFT using the inverse STFT (ISTFT) [9]. There are different ISTFT methods, but the most common is the overlapp-and-add (OLA) method, that can be mathematically described as,

$$\tilde{x}(m) = \sum_{t=0}^{T-1} w(m-tl_s)\left(\left[\frac{1}{N_s}\sum_{f=0}^{N_s-1} X(t,f)e^{j\frac{2\pi}{N_s}mf}\right] * \delta(m-tl_s)\right), \tag{2.3}$$

where T is the number of time frames, $\delta(m)$ is the unitary impulse function and $*$ represents the convolution operation. The term between the brackets represents the inverse DFT (IDFT) of each time trame, which are temporally shifted and multiplied by the window function. Finally, the shifted time frames are summed to obtain the temporal sequence.

The OLA method requires a proper window function $w(n)$ [7]. This window is chosen for its spectral properties, as its spectrum is convolved with the one of the speech signal. To understand this, we have to remember that the segmentation process is equivalent to the multiplication with a rectangular function, whose DFT has certain properties (as the width of its main lobe or the relative power of the sidelobes). A different window function can be chosen with better characteristics than the rectangular function. Also, we could use the window only for the STFT (analysis). Nevertheless, if the STFT is going to be processed, the use of a window in the ISTFT (synthesis) can provide a better performance, alleviating some speech artifacts. Considering the use of analysis/synthesis windows, the resulting time sequence $\widetilde{x}(m)$ is equivalent to the original sequence $x(m)$ if the next condition is fulfilled,

$$\sum_{t=0}^{T-1} w^2(m - tl_s) = 1, \forall m, \tag{2.4}$$

which is known as the perfect reconstruction condition. An example of window function that fulfills this condition using a 50% overlap ($l_s/N_s = 0.5$) is the square-root Hann function, that can be defined as

$$w^2(m) = \begin{cases} \frac{1}{2}\left(1 - \cos\left(\frac{2\pi m}{N_s}\right)\right) & \text{if } m < N_s - 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

The reader can notice that the condition does not entirely fulfill for the first and last samples of the signal. Nevertheless, we can ignore those samples if they do not contain speech, or we can use a proper zero padding at the beginning and end of the time sequence.

Finally, an important property of the STFT is that it is a linear operation, so the additive distortion model for the noisy speech signal remains in the transformed domain,

$$Y(t,f) = X(t,f) + N(t,f), \tag{2.6}$$

where $Y(t,f)$ and $N(t,f)$ are the noisy speech STFT and the noise STFT, respectively.

In the next sections, we will explore different speech enhancement techniques in the STFT domain. These techniques process the noisy speech STFT signal, $Y(t,f)$, to obtain an estimation of the clean speech STFT, $\widehat{X}(t,f)$. The estimated clean speech time sequence $\widehat{x}(m)$ is then obtained using the ISTFT transformation.

2.2 Classical single-channel speech enhancement

The classic speech enhancement approaches consider the single-channel additive distortion model for the noisy speech signal in the STFT domain. Their objective is to eliminate the noise component for the signal while the clean speech signal remains unaltered, increasing the speech quality and/or intelligibility. These techniques are based on different assumptions about the statistical properties of the underlying signal components, as the statistical independence between the clean speech and the noise, the slowly variant characteristics of the noise, or the modeling of these statistics using well-known probabilistic models. These assumptions are necessary to deal with a variety of noisy environments, but their accuracy yields limited performance in challenging non-stationary scenarios. In general, the use of these algorithms implies a trade-off between the noise reduction achieved and the speech distortion introduced. Moreover, their performance highly depends on a good estimation for the statistics of the clean speech and noise signals. As an advantage, these techniques are usually easy to implement in resource-efficient and low-latency applications, and they are the base of more complicated techniques that we will review in the next sections.

Many of these classical algorithms follow a gain-based approach, which means that the clean speech signal is estimated by applying a gain function to the noisy STFT as

$$\widehat{X}(t,f) = G(t,f)Y(t,f), \tag{2.7}$$

where $G(t,f)$ is the gain value at each time-frequency bin. Moreover, this gain function is usually real-valued, so the phase of the estimated signal remains the same as the noisy speech signal. This introduces phase distortion, especially at low signal-to-noise ratios (SNRs). The phase problem has been investigated in recent years, and the interested reader can find in [63] a review of phase processing techniques using classical speech processing. Nevertheless, the phase processing is out of the scope in this Thesis, as the phase effect in the enhancement procedure is limited [225] and requires additional treatment.

Next, we will review the most widely known approaches for classical single-channel speech enhancement. These approaches mainly differ in the statistical framework and the definition of the gain function used.

2.2.1 Spectral subtraction algorithms

The spectral subtraction (SS) is one of the first algorithms proposed for speech enhancement in the literature [14]. The basic assumption is the availability of an estimate for the noise amplitude STFT, $|N(t,f)|$. Assuming that the phases of the noisy and clean speech are similar, the clean speech STFT is approximated as

$$\widehat{X}(t,f) = \left[|Y(t,f)| - |N(t,f)|\right] e^{j\theta_y(t,f)}, \tag{2.8}$$

where θ_y represents the noisy phase. As we will see later, the noise is a random signal that cannot be predicted, so most of the noise estimation methods seek an estimate of the noise spectral variance, defined as

$$\sigma_n^2(t,f) = E\left\{|N(t,f)|^2\right\}, \tag{2.9}$$

where $E\{\cdot\}$ represents the expected value of a random variable. The subtraction is then performed in the power spectrum domain,

$$\left|\widehat{X}(t,f)\right|^2 = |Y(t,f)|^2 - \sigma_n^2(t,f), \tag{2.10}$$

and the amplitude spectrum is obtained from the power spectrum. The subtraction could take negative values, so the obtained estimate is lower-bounded by zero. This algorithm can be also written in terms of a gain function applied to the noisy speech STFT,

$$G_{SS}(t,f) = \sqrt{1 - \frac{\sigma_n^2(t,f)}{|Y(t,f)|^2}}. \tag{2.11}$$

Although easy to implement, the SS algorithm is based on heuristics instead of an optimization criterion. In addition, this approach considers the cross-terms in the computation of the power spectrum equals to zero, which is not necessarily true. To overcome these limitations, different alternatives have been proposed for the SS algorithm: the use of the masking properties of the human auditory [222], a geometric approach for the subtraction rule [133], over-subtraction [215], non-linear subtraction [124] or multi-band subtraction [100] are some examples.

2.2.2 Wiener filtering

The Wiener filter (WF) [21] is based on a statistically optimal criterion with well-defined statistical assumptions, in contrast to the previously presented SS algorithm. The WF filter is a linear minimum mean square estimator (MMSE). Thus, its gain function in the STFT domain is computed by solving the following minimization problem,

$$G_{WF}(t,f) = \underset{G(t,f)}{\operatorname{argmin}}\, E\left\{|X(t,f) - G(t,f)Y(t,f)|^2\right\}. \tag{2.12}$$

Therefore, the WF filter is given for that gain function that minimizes the mean square error between the clean and estimated speech spectra. Assuming that the clean speech and noise STFT coefficients are uncorrelated zero-mean random variables, and their second-order moments, or variances, are respectively $\sigma_x^2(t,f)$ and $\sigma_n^2(t,f)$, the WF gain is obtained as

$$G_{WF}(t,f) = \frac{\xi(t,f)}{\xi(t,f)+1}, \tag{2.13}$$

where $\xi(t,f) = \sigma_x^2(t,f)/\sigma_n^2(t,f)$ is the a priori SNR. As can be observed, the gain function is close to zero for those bins with a low SNR (noise-dominant bins), suppressing that frequencies, while for those bins with high SNR (speech dominant) the gain function is close to one, preserving those components.

Apart from the classical WF, other alternative Wiener-like functions have been proposed in the literature, as the square-root or parametric Wiener filter [125], the codebook-driven filter [197], or the use of psychoacoustic properties of the human auditory system [89, 90].

2.2.3 Model-based Bayesian estimators

The Bayesian estimators do not assume a linear relationship between the estimator and the noisy speech signal, as in the case of WF. Moreover, these methods define optimal estimators for the clean speech amplitude, while using the phase of the noisy speech signal. This is inspired by preliminary works as [225, 42], which showed that the amplitude of the speech signal carried most of the information, and its estimation is easier than the clean speech phase one.

The general Bayesian optimization criteria are based on the minimization of the expected value of a given cost function provided a noisy spectrum $Y(t,f)$. This cost function depends on the clean speech amplitude $S(t,f) = \left|\widehat{X}(t,f)\right|$ and its estimation, which is obtained as

$$\widehat{S}(t,f) = \underset{S(t,f)}{\operatorname{argmin}}\, E\left\{C\left(S(t,f),\widehat{S}(t,f)\right)\middle|\,Y(t,f)\right\}, \tag{2.14}$$

where $C(\cdot,\cdot)$ is a particular Bayesian cost function. A first proposal for the cost function was the mean square error [42],

$$C_{SA}\left(S(t,f),\widehat{S}(t,f)\right) = \left(S(t,f) - \widehat{S}(t,f)\right)^2. \tag{2.15}$$

To solve the minimization problem, the probabilistic distribution of the STFT coefficients must be defined. Assuming that these coefficients follow a zero-mean complex circularly symmetric Gaussian distribution, the amplitudes follow a Rayleigh distribution and the phase a uniform distribution. Moreover, the amplitude and phase are supposed to be independent. Based on these assumptions, the MMSE estimator of the clean speech amplitude (SA), defined as in [42], can be expressed as the following gain function,

$$G_{SA}(t,f) = \frac{\sqrt{\nu(t,f)}}{\gamma(t,f)}\Gamma(1.5)\mathscr{M}\left(-0.5,1;-\nu(t,f)\right), \tag{2.16}$$

where $\Gamma(\cdot)$ is the gamma function, $\mathscr{M}(\cdot,\cdot;\cdot)$ is the confluent hypergeometric function, $\gamma(t,f) = |Y(t,f)|^2/\sigma_n^2(t,f)$ is the a posteriori SNR, and

$$\nu(t,f) = \frac{\xi(t,f)}{\xi(t,f)+1}\gamma(t,f). \tag{2.17}$$

As demonstrated in [42], the MMSE estimator for the clean speech phase is the noisy speech phase. The same authors suggested in [43] a logarithmic version of the MMSE estimator, also known as the log-MMSE estimator of the speech amplitude (LSA), which defines a cost function in the log-spectra domain as

$$C_{LSA}\left(S(t,f),\widehat{S}(t,f)\right) = \left(\log S(t,f) - \log \widehat{S}(t,f)\right)^2. \tag{2.18}$$

The idea of this estimator is that the log-domain is closer to human acoustic perception. The log-MMSE estimator yields the following gain function,

$$G_{LSA}(t,f) = \frac{\nu(t,f)}{\gamma(t,f)}\exp\left(\frac{1}{2}\int_{\nu(t,f)}^{\infty}\frac{e^{-u}}{u}du\right). \tag{2.19}$$

This estimator was further improved in [29] with the optimally-modified LSA (OMLSA) estimator, which includes the speech presence probability (SPP) to increase the noise reduction performance.

Apart from the previous Bayesian estimators, other cost functions have been proposed [125], yielding different gain functions. Two classic Bayesian estimators are the maximum-likelihood (ML) and the maximum a posteriori (MAP) estimators, that are defined respectively as

$$\widehat{S}^{(\mathrm{ML})}(t,f) = \underset{S(t,f)}{\operatorname{argmax}}\ p\left(Y(t,f) | S(t,f), \theta_x(t,f)\right), \tag{2.20}$$

$$\widehat{S}^{(\mathrm{MAP})}(t,f) = \underset{S(t,f)}{\operatorname{argmax}}\ p\left(S(t,f), \theta_x(t,f) | Y(t,f)\right), \tag{2.21}$$

where $\theta_x(t,f)$ is the clean speech phase. The main difference between these estimators is that the ML approach considers that the clean speech is a deterministic value, while the MAP approach treats the clean speech as a random variable with a given speech prior model.

The different Bayesian estimators yield a gain function that mainly depends on the a priori SNR $\xi(t,f)$ and the a posteriori SNR $\gamma(t,f)$. These SNRs are the key values to be estimated, as their accurate estimation bounds the performance of these approaches. While the a posteriori SNR only depends on the noise estimation, the a priori SNR also needs a good clean speech variance estimate. An approach to recursively estimate the a priori SNR was proposed in [42], which is known as the decision-directed approach

$$\widehat{\xi}(t,f) = \alpha \frac{\left|\widehat{X}(t-1,f)\right|^2}{\sigma_n^2(t-1,f)} + (1-\alpha)\max\left(\gamma(t,f)-1, 0\right), \tag{2.22}$$

where $\alpha \in (0,1)$ is a weighting factor that controls the mixture between the estimation obtained in the last frame, based on the estimated clean speech, and an estimation of the SNR in the current frame using the a posteriori SNR.

Finally, recent works have explored the use of non-Gaussian distribution models for the clean speech signal [131]. These works are motivated by the fact that the shape of the clean speech histogram can be better approximated using other super-Gaussian distributions, as generalized gamma distributions [171, 189]. Moreover, certain distributions yield to closed-form gain functions that can provide a performance increase for speech enhancement.

2.2.4 Noise estimation

The definition of the gain function in the aforementioned speech enhancement methods requires an estimation of the noise spectral variance $\sigma_n^2(t,f)$. Moreover, the performance of these methods relies on the accuracy of this noise estimate. The topic of noise estimation has been relevant in the past decades, with different approaches addressing it from different perspectives. In this subsection, we describe some of the most relevant works in this area, particularly those closer to the research carried out in this Thesis.

Voice activity detection (VAD). This is one of the simplest and earliest approaches in the noise estimation field. The VAD methods exploit the fact that speech is absent in short periods between the speaker's activity. Therefore, noise statistics can be updated in speech absent frames, while keeping the last estimate during speech present ones. Consider a $\mathrm{VAD}(t)$ as a binary decisor which yields one when speech is active in a certain frame and zero otherwise. The recursive noise estimator is then defined as

$$\widehat{\sigma}_n^2(t,f) = \begin{cases} \alpha\widehat{\sigma}_n^2(t-1,f) + (1-\alpha)\left|Y(t,f)\right|^2 & \text{if VAD}(t) = 0, \\ \widehat{\sigma}_n^2(t-1,f) & \text{otherwise.} \end{cases} \tag{2.23}$$

Although its implementation is relatively simple, the fact that noise tracking is only possible during speech pauses limits the performance in non-stationary noisy environments. Most of the VAD techniques can be classified into statistical-driven [64] and data-driven [54, 217, 223] approaches.

Minimum statistics (MS) tracking. The MS method was proposed by Martin [139]. It is based on the observation that, although speech can be present in a frame, only a fraction of frequency bins contain speech energy due to the sparsity of the speech signal in frequency domain. Moreover, the speech energy is often distributed in spectral peaks located at specific frequencies, while the surrounding bins, with lower energy levels, are representative of the noise spectrum. The MS approach exploits this fact by computing the noisy speech periodogram as

$$\widehat{\sigma}_y^2(t,f) = \alpha(t,f)\widehat{\sigma}_y^2(t-1,f) + (1-\alpha(t,f))\left|Y(t,f)\right|^2, \tag{2.24}$$

where $\alpha(t,f)$ is a smoothing parameter that is computed by minimizing the MSE between the noisy speech variance and the actual noise variance when speech is absent,

$$\alpha(t,f) = \underset{\alpha(t,f)}{\operatorname{argmin}}\, E\left\{ \left(\widehat{\sigma}_y^2(t,f) - \sigma_n^2(t,f)\right)^2 \middle| \widehat{\sigma}_y^2(t-1,f), \sigma_x^2(t,f) = 0 \right\}. \tag{2.25}$$

The algorithm tracks the minimum value of the noisy periodogram within a window of D neighboring frames,

$$\widehat{\sigma}_{y,\min}^2(t,f) = \min\left(\left[\widehat{\sigma}_y^2(t,f), \widehat{\sigma}_y^2(t-1,f), \cdots, \widehat{\sigma}_y^2(t-D+1,f)\right]\right). \tag{2.26}$$

The resulting value is then considered as an underestimate of $\sigma_n^2(t,f)$. A bias compensation factor is introduced and the noise variance is then obtained as

$$\widehat{\sigma}_n^2(t,f) = \frac{\widehat{\sigma}_{y,\min}^2(t,f)}{E\left\{\widehat{\sigma}_{y,\min}^2(t,f)\right\}_{\sigma_n^2(t,f)=1}}. \tag{2.27}$$

Despite its simplicity, its implementation requires some simplifications that make the estimation suboptimal. In addition, fast changes in the noise level are detected with an undesirable high delay, resulting in a large amount of residual noise after applying a noise reduction technique.

Time-recursive averaging algorithms. These methods obtain a noise estimation by time-smoothing of the noisy speech signal in a similar way than that from the VAD approach. As a difference, this recursion is done independently in each frequency, exploiting the aforementioned property of the speech energy distribution at each frame. The noise variance estimate is obtained as

$$\widehat{\sigma}_n^2(t,f) = \alpha(t,f)\widehat{\sigma}_n^2(t-1,f) + (1-\alpha(t,f))\left|Y(t,f)\right|^2, \tag{2.28}$$

where the computation of the smoothing factor $\alpha(t,f)$ depends on the particular time-averaging algorithm used. For example, some approaches as [121] define the smoothing factor as a function of the SNR at each time-frequency bin. This way, $\alpha(t,f)$ is close to one when the SNR is high at that bin, so the previous noise estimate is kept. On the other hand, the noise is updated when the SNR is low at that bin. Another well-known approach relies on updating the noise variance whenever the probability of speech being present is low. This approximation is followed by the minimum-controlled recursive averaging (MCRA) algorithm [30]. Using a detection theory framework, we can define a binary random variable $\mathscr{D}(t,f) = \{\mathscr{H}_x, \mathscr{H}_n\}$ indicating speech presence or absence, respectively, for each bin. Then, two hypotheses are considered:

$$Y(t,f) = X(t,f) + N(t,f) \quad \text{if } \mathscr{D}(t,f) = \mathscr{H}_x, \tag{2.29}$$

$$Y(t,f) = N(t,f) \quad \text{if } \mathscr{D}(t,f) = \mathscr{H}_n. \tag{2.30}$$

The MCRA method updates the noise estimate using the following smoothing factor,

$$\alpha(t,f) = \widetilde{\alpha} + (1-\widetilde{\alpha})\, p_x(t,f), \tag{2.31}$$

where $\widetilde{\alpha}$ is a weighting factor and

$$p_x(t,f) = P\left(\mathscr{D}(t,f) = \mathscr{H}_x | Y(t,f)\right) \tag{2.32}$$

is the a posteriori speech presence probability (SPP). The MCRA algorithm first obtains an a posteriori SNR-like estimate from the noisy speech signal, which is then compared with a threshold to decide if speech is present. Finally, the binary estimate is smoothed over time to obtain the a posteriori SPP. An improved version of the MCRA (IMCRA) was proposed in [27]. This method introduces a bias compensation factor for the noise estimate. In addition, the a posteriori SPP is now computed using a Bayesian approach where the noisy speech and noise STFTs follow a complex Gaussian distribution. Besides, this approach requires knowledge about the a priori speech absence probability (SAP),

$$q_n(t,f) = P\left(\mathscr{D}(t,f) = \mathscr{H}_n\right). \tag{2.33}$$

The a priori SAP depends on estimates for the a priori SNR and the a posteriori SNR. A two-pass iteration is used at each frame to increase the robustness against non-stationary noises. The IMCRA approach provides better robustness and more accurate noise estimates than previous methods. Nevertheless, it still has difficulties with noise abrupt changes, which can be misunderstood as speech presence.

MMSE-based noise estimation. A low-complexity MMSE-based noise estimation criterion was proposed in [74]. The noise variance is estimated as

$$\widehat{\sigma}_n^2(t,f) = E\left\{ |N(t,f)|^2 \middle| Y(t,f) \right\}, \tag{2.34}$$

This approach assumes that both the clean speech and noise STFTs coefficients are statistically independent zero-mean complex random variables and that they can be modeled using complex Gaussian distributions. Thus, the estimator can be written in terms of the a priori SNR and the a posteriori SNR as follows,

$$\widehat{\sigma}_n^2(t,f) = \left(\frac{1}{(1+\xi(t,f))^2} + \frac{\xi(t,f)}{(1+\xi(t,f))\,\gamma(t,f)} \right) |Y(t,f)|^2. \tag{2.35}$$

Despite this estimation is unbiased, the inaccuracies in the a priori SNR estimation makes a bias compensation necessary. Furthermore, a time-smoothing is applied to the noise estimate to decrease the estimation variance. As an advantage, the MMSE approach shows a stable performance at a wide range of SNR values, and it is more robust to quick changes in the noise spectra than the previously described methods. Later, the authors improved the method with the introduction of an SPP-based approach with fixed priors [61] and a speech variance estimator based on temporal cepstrum smoothing [62].

MAP and ML noise estimators. Apart from the MMSE-based estimation, other Bayesian approaches have been explored for noise estimation. For example, a MAP estimator of the noise variance was addressed in [25, 24]. This approach assumes the availability of an initial clean speech variance estimate, so it is meant to be used as a post-processor stage in an enhancement algorithm. On the other hand, a recursive ML-based noise estimator was derived in [196]. This algorithm uses a recursive expectation-maximization (EM) algorithm to jointly estimate the a priori SPP, the a posteriori SPP, and the noise variance in a unified framework. The derived method shows some similarities with the IMCRA approach. The obtained estimates are optimal in the ML sense, the computational load is low, and the performance is similar or better than the previously described approaches.

2.3 Multichannel speech enhancement based on beamforming techniques

Until now, we have considered that only a single-channel noisy speech signal was available. In the case that several microphones are used to capture the speech signal from a target speaker, using a microphone array, for example, the noisy speech signal at each microphone can be written as

$$Y_j(t,f) = X_j(t,f) + N_j(t,f), \tag{2.36}$$

where $j = 1,..,J$ is the microphone index and J the number of microphones. Let us also consider that the clean speech signal at each microphone is different, due to the room acoustics and/or the microphone responses. Besides, let us assume that the reverberation level is low enough, so we can relate the clean speech between each microphone and a reference channel (we take $j = 1$ without loss of generality) using a relative transfer function (RTF) [55], given as,

$$H_{j1}(t,f) = \frac{X_j(t,f)}{X_1(t,f)}, \tag{2.37}$$

This is known as the narrowband model assumption [57], where the effect of the channel acoustics is multiplicative at each time-frequency bin and we can use RTFs. Considering this assumption, the multichannel noisy speech signal can be written in vector form as

$$\mathbf{y}(t,f) = \mathbf{h}(t,f)X_1(t,f) + \mathbf{n}(t,f), \tag{2.38}$$

where

$$\mathbf{y}(t,f) = \begin{bmatrix} Y_1(t,f) & Y_2(t,f) & \cdots & Y_J(t,f) \end{bmatrix}^\top, \tag{2.39}$$

$$\mathbf{n}(t,f) = \begin{bmatrix} N_1(t,f) & N_2(t,f) & \cdots & N_J(t,f) \end{bmatrix}^\top, \tag{2.40}$$

$$\mathbf{h}(t,f) = \begin{bmatrix} 1 & H_{21}(t,f) & \cdots & H_{J1}(t,f) \end{bmatrix}^\top, \tag{2.41}$$

and $\{\cdot\}^\top$ is the transpose operator. The multichannel noise signal can be described using the noise spatial covariance matrix (SCM) $\Sigma_N(t,f) = E\left\{\mathbf{n}(t,f)\mathbf{n}^H(t,f)\right\}$, where $\{\cdot\}^H$ is the Hermitian transpose operator.

The objective is the estimation of the clean speech signal at the reference microphone, $X_1(t,f)$, from the multichannel noisy speech signal $\mathbf{y}(t,f)$. The most common multichannel speech enhancement techniques are the beamforming algorithms [11, 112], which consist of applying spatial filtering on the multichannel noisy speech signal as

$$\widehat{X}_1(t,f) = \mathbf{d}^H(t,f)\mathbf{y}(t,f), \tag{2.42}$$

where $\mathbf{d}(t,f)$ is the vector of beamforming weights. These weights are in general complex-valued, so the multichannel noisy speech coefficients at each bin are modified both in amplitude and phase.

There exist different design criteria for beamforming. For example, they can be designed in view of the geometry of the problem: the position of the microphones and the target speaker, beampatterns at different frequencies, etc. Examples of these beamformers are the delay-and-sum beamformer [216], the superdirective beamformer [36], and the differential beamformers [12], which have gained popularity in recent years. On the other hand, the beamformer can be designed taking into account the statistics of the noisy speech signal and an optimization criterion, so they are known as data-dependent beamformers [57]. We focused on this kind of beamformers in this Thesis. In the next subsections, we will describe the most common data-dependent beamformers and how to estimate the needed statistics.

2.3.1 Minimum variance distortionless response

Minimum variance distortionless response (MVDR) beamforming [16] is a common data-dependent beamformer that minimizes the noise power while ensuring that the clean speech signal is not distorted. The MVDR beamformer weights are calculated by solving the following minimization problem,

$$\mathbf{d}_{\mathrm{MVDR}}(t,f) = \underset{\mathbf{d}(t,f)}{\operatorname{argmin}}\ \mathbf{d}^H(t,f)\Sigma_N(t,f)\mathbf{d}(t,f) \quad \text{s.t. } \mathbf{d}^H(t,f)\mathbf{h}(t,f) = 1, \tag{2.43}$$

where the restriction $\mathbf{d}^H(t,f)\mathbf{h}(t,f) = 1$ implies a distortionless response on the clean speech signal at the reference channel. This optimization problem can be solved by the Lagrange multipliers method, yielding the following beamformer weights

$$\mathbf{d}_{\mathrm{MVDR}}(t,f) = \frac{\Sigma_N^{-1}(t,f)\mathbf{h}(t,f)}{\mathbf{h}^H(t,f)\Sigma_N^{-1}(t,f)\mathbf{h}(t,f)}. \tag{2.44}$$

At the beamformer output we have the clean speech signal at the reference channel plus a residual noise with a variance given as,

$$\sigma_o^2(t,f) = \left(\mathbf{h}^H(t,f)\Sigma_N^{-1}(t,f)\mathbf{h}(t,f)\right)^{-1}. \tag{2.45}$$

Alternatively, the MVDR beamformer can be written using the formulation derived in [193],

$$\mathbf{d}_{\mathrm{MVDR}}(t,f) = \frac{\Sigma_N^{-1}(t,f)\Sigma_X(t,f)}{\operatorname{tr}\left\{\Sigma_N^{-1}(t,f)\Sigma_X(t,f)\right\}}\mathbf{u}_1, \tag{2.46}$$

where $\operatorname{tr}\{\cdot\}$ is the trace operator, $\mathbf{u}_1 = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}^\top$ is the unitary column vector of dimension J, and $\Sigma_X(t,f) = \mathbf{h}(t,f)\mathbf{h}^H(t,f)\sigma_{x_1}^2(t,f)$ is the rank-1 SCM matrix for the clean speech signal, with $\sigma_{x_1}^2(t,f) = E\left\{|X_1(t,f)|^2\right\}$.

2.3.2 Multi-channel Wiener filter and postfiltering

Similarly to its single-channel version, the multichannel Wiener filter (MWF) [35, 37] can be derived as a multichannel MMSE linear filter,

$$\mathbf{d}_{\mathrm{MWF}}(t,f) = \underset{\mathbf{d}(t,f)}{\operatorname{argmin}}\ E\left\{\left|X_1(t,f) - \mathbf{d}^H(t,f)\mathbf{y}(t,f)\right|^2\right\}. \tag{2.47}$$

The solution to this minimization problem gives the following beamformer weights,

$$\mathbf{d}_{\mathrm{MWF}}(t,f) = \left(\Sigma_X(t,f) + \Sigma_N(t,f)\right)^{-1}\Sigma_X(t,f)\mathbf{u}_1. \tag{2.48}$$

This expression is similar to the single-channel Wiener filter one, but spatial covariance matrices are used instead of the single-channel variances.

It must be noted that MWF can be decomposed into an MVDR beamformer followed by a single-channel Wiener filter,

$$\mathbf{d}_{\text{MWF}}(t,f) = \mathbf{d}_{\text{MVDR}}(t,f)\frac{\sigma_{x_1}^2(t,f)}{\sigma_{x_1}^2(t,f)+\sigma_o^2(t,f)}, \tag{2.49}$$

The single-channel WF filter further eliminates the residual noise at the beamformer output, but distortion is introduced in the clean speech signal. Moreover, both MVDR and MWF belong to a broad class of generalized beamformers that steers to the same spatial direction and only differs in a spectral gain [232],

$$\mathbf{d}_{\text{GBF}}(t,f) = G_{\text{F}}(t,f)\Sigma_N^{-1}(t,f)\Sigma_X(t,f)\mathbf{u}_1, \tag{2.50}$$

where $G_{\text{F}}(t,f)$ represents the real-valued spectral gain.

The previous idea can be generalized with the concept of postfiltering, which consists of the use of a single-channel speech enhancement technique at the beamformer output to improve the noise reduction performance. Several postfilters can be found in the literature, most of them based on the use of an MVDR beamformer. For example, Zelinski [248] uses an MWF approach and it assumes a spatially-white noise field (uncorrelated noises across microphones) to estimate the speech and noise statistics. Marro *et al.* [138] also applies an MWF approach, further improving its performance by considering acoustic echoes and reverberation. The approach in [149] assumes a diffuse noise field, which is a better approximation for real acoustic environments as in an office or inside a car. The postfilter presented in [118] obtains more accurate statistics using the enhanced signal at the beamformer output. That work also explores the concept of non-linear postfilters, combining an MVDR beamformer with an MMSE or a log-MMSE estimator. Other approaches use a generalized eigenvalue (GEV) beamformer, which maximizes the SNR at the beamformer output, along with a Blind Analytic Normalization (BAN) postfilter [234] to reduce the speech distortion. This approach obtains a comparable or superior performance than the MVDR beamformer. Other beamformer architecture that has been widely explored in the literature is the Generalized Sidelobe Canceller (GSC) [65], which uses a beamformer, as MVDR, followed by an adaptive filter to eliminate the residual noise. Gannot *et al.* [56] proposes the use of a GSC architecture followed by an OMLSA estimator as a postfilter, and it also includes an intermediate step for the SPP estimation. To sum up, the use of beamforming plus postfiltering is a widely used approach for multichannel speech enhancement. This method achieves a good trade-off between the spatial filtering with low speech distortion and the noise reduction capabilities of the single-channel enhancement techniques.

2.3.3 Noise spatial covariance and relative transfer function estimation

The performance of the aforementioned beamforming techniques depends on accurate estimates for the noise spatial covariance matrix (SCM) $\Sigma_N(t,f)$ and the relative transfer function $\mathbf{h}(t,f)$. For example. MVDR needs to be steered correctly to the target speaker to avoid speech distortion at the beamformer output. Moreover, its noise reduction performance is conditioned to a good knowledge of the noise field. This subsection is intended to review relevant works in the literature devoted to the estimation of these acoustic parameters.

Even though the noise estimation methods have been extensively studied in the literature for the single-channel case, its application to the multichannel scenario is more recent and poses additional challenges. The main difference with respect to classical single-channel noise estimators is that we need to estimate not only the noise variance at each microphone, $\sigma^2_{n_j}(t,f)$, but also the cross-correlation noise terms $\sigma^2_{n_{jk}}(t,f) = E\left\{N_j(t,f)N_k^*(t,f)\right\}$. The estimation of these terms is not straightforward as they are generally complex-valued, so classical noise estimators are not directly applicable. Moreover, in order to achieve robustness against non-stationary noises, we have to accurately track the spectral and spatial characteristics of the noise field. Despite these difficulties, different research lines have been followed to extend the noise estimation algorithms to the multichannel scenario. Also, new methods have been proposed exploiting the spatial information or assuming certain properties of the noise field. The multichannel noise estimation approaches can be grouped into the following general categories [162]:

1. The direct extension of the minimum statistics tracking approach to update the cross-correlation terms. Different alternatives have been proposed, as the additional use of a VAD to detect speech pauses [48] or a soft-decision based scheme [99].

2. The assumption of certain spatial properties for the noise field. This approach simplifies the estimation of the noise statistics involved in the SCM matrix. The methods in this category usually work with the coherence function, which is a normalized version of the SCM matrix. The terms of the coherence function are obtained as follows,

$$\Gamma_{N_{jk}}(t,f) = \frac{\sigma^2_{n_{jk}}(t,f)}{\sqrt{\sigma^2_{n_j}(t,f)\sigma^2_{n_k}(t,f)}}. \tag{2.51}$$

   A common assumption is that of diffuse noise field, which has a coherence function defined as

$$\Gamma_{\text{diff}_{jk}}(f) = \text{sinc}\left(\frac{2\pi f F_s d_{jk}}{F c_v}\right), \tag{2.52}$$

where F_s is the sampling frequency, c_v is the speed of sound and d_{jk} is the distance between each pair of microphones j and k. This model assumes a high correlation at low frequencies between the channels, while the noises at the different microphones are independent at high frequencies. This diffuse noise field assumption has been exploited in different works [175, 94]. The works in [158, 97] improve this approach using a single-channel SPP estimator with fixed priors to update the noise statistics.

3. The availability of knowledge about the relative transfer function [73, 110]. The basic idea is to use the RTF function to cancel the speech component between the microphones and, then, estimate the noise cross-terms from the resulting signals.

4. Blind methods that use the speech presence probability to update the noise estimation. An extension of the IMCRA method for multichannel scenarios was proposed in [195] and further improved in [194]. These approaches use multivariate Gaussian models along with SNR-based multichannel SAP estimators to predict the a posteriori SPP. The work in [209] uses a coherent-to-diffuse ratio SAP approach, which exploits the spatial characteristics of the speech and noise signals. That work also proposed an ML scheme using an EM algorithm to jointly estimate the a posteriori SPP and the noise SCM in an online fashion. The EM algorithm has been also explored for noise estimation in [80] using a complex Gaussian mixture model (cGMM) approach. That work was extended in [79] by using spatial priors to regularize the noise estimation. In addition, other spatial models as complex angular Gaussian distributions [93] has been explored. Finally, the available information about the speaker position was exploited in [210, 211] for the SPP estimation.

The estimation of the relative transfer function (RTF) is also a challenging task as it depends on the speaker position, the room acoustics, and the microphone responses. The most common RTF estimators are based on sub-space searching using estimates for the noisy speech and noise spatial covariance matrices. Assuming statistical independence between the clean speech and the noise, and the narrowband model assumption, the covariance subtraction (CS) method [28] obtains an RTF estimate as

$$\mathbf{h}_{\mathrm{CS}}(t,f) = \frac{\left(\widehat{\Sigma}_Y(t,f) - \widehat{\Sigma}_N(t,f)\right)\mathbf{u}_1}{\mathbf{u}_1^\top \left(\widehat{\Sigma}_Y(t,f) - \widehat{\Sigma}_N(t,f)\right)\mathbf{u}_1}, \tag{2.53}$$

where $\widehat{\Sigma}_Y(t,f)$ and $\widehat{\Sigma}_N(t,f)$ are estimates of the noisy speech and the noise spatial covariance matrices, respectively. As observed, the CS method obtains an RTF estimate as a column vector, conveniently normalized, from the estimated clean speech SCM, computed as the difference between $\widehat{\Sigma}_Y(t,f)$ and $\widehat{\Sigma}_N(t,f)$. Alternatively, the eigenvalue decomposition (EVD) method [188, 218] estimates the RTF function as the principal eigenvector of the clean speech SCM,

$$\mathbf{h}_{\mathrm{EVD}}(t,f) = \mathscr{P}\left(\widehat{\Sigma}_Y(t,f) - \widehat{\Sigma}_N(t,f)\right), \tag{2.54}$$

where $\mathscr{P}(\cdot)$ stands as the principal component of a matrix. As advantage, EVD is more robust than CS against inaccuracies in the rank-1 model for the clean speech signal. Another alternative is the covariance whitening (CW) approach [136, 137], which estimates the RTF as

$$\mathbf{h}_{\mathrm{CW}}(t,f) = \frac{\widehat{\Sigma}_N^{H/2}(t,f)\widetilde{\mathbf{h}}(t,f)}{\mathbf{u}_1^\top \widehat{\Sigma}_N^{H/2}(t,f)\widetilde{\mathbf{h}}(t,f)}, \tag{2.55}$$

where $\{\cdot\}^{1/2}$ is the square-root decomposition of a regular matrix and

$$\widetilde{\mathbf{h}}(t,f) = \mathscr{P}\left(\widehat{\Sigma}_N^{-H/2}(t,f)\widehat{\Sigma}_Y(t,f)\widehat{\Sigma}_N^{-1/2}(t,f)\right) \tag{2.56}$$

is the principal eigenvector of the whitened noisy speech SCM. This method is based on a generalized eigenvalue decomposition (GEVD) problem followed by a de-whitening. Although the previous sub-space methods are the most common ones, other approaches are described in the literature to deal with the RTF estimation. For example, the approach in [55] formulates a least-squares problem that uses the speech sparsity to jointly obtain the RTF and noise statistics. A weighted least-squares approach was also proposed in [41] as well as a recursive least squares method to solve the joint estimation problem in an online fashion. In [28], SPPs were incorporated into the weighted least-squares problem, yielding more accurate solutions.

Previous methods independently estimate the model parameters and the clean speech signal. In contrast, other approaches jointly estimate the different statistics and acoustic parameters using an ML or MAP framework. The resulting equations have not a direct solution, so the EM algorithm is used to find a suboptimal solution. The EM framework has been applied to different offline multichannel speech enhancement, dereverberation, and source separation problems [40, 213, 186, 66, 183]. On the other hand, to deal with online scenarios, like those we are considering in this Thesis, a recursive EM algorithm (REM) [17] can be used instead. The REM framework has already been explored in speech processing tasks as multichannel speech enhancement [185] and speech dereverberation [184], and it is also used for the proposals described in Chapter 5.

2.4 Deep neural networks for speech enhancement

The different speech enhancement algorithms presented in the previous sections have in common that they are mainly based on statistical signal processing. This means that these techniques have been derived from assumptions about the clean speech signal and the environmental noise statistics, such as their correlations, stationarity, probabilistic distributions, and so on. These assumptions are needed to derive a mathematical formulation to solve the estimation of the clean speech signal, giving closed-form optimal estimators or sub-optimal approximations that provide good enough results. Although these methods are relatively simple to be implemented in practical applications, their performance is highly dependent on the accuracy of both the assumptions made and the estimation of the underlying statistics. Therefore, classical approaches are limited in many real-world noisy environments (low SNRs, highly non-stationary noises, reverberant acoustic environments, etc). These limitations are commonly translated to the introduction of distortions in the enhanced speech signal that degrades the speech quality and/or intelligibility perceived by a human listener.

Another paradigm for speech enhancement is the use of data-driven approaches. These methods do not make assumptions about the characteristics of the signals, but they learn these properties from observations of speech examples to solve a certain problem. Therefore, they are widely known as machine learning approaches [13]. These algorithms are mainly classified into two categories: supervised learning, where pairs of observation and target data are available (for example, noisy speech and clean speech data in a speech enhancement task), or unsupervised learning if only the observations are available. Different methods have been used in the past years for speech enhancement [203, 104, 69], consisting mainly of large parametrized mathematical models that are optimized through a training procedure using databases. Some examples are the Gaussian mixture models (GMM) [150], the hidden Markov models (HMM) [174], the support vector machines (SVM) [31], or the swallow artificial neural networks (ANN) [88]. Nevertheless, the performance of these approaches was limited in practical applications when compared to classical statistical signal processing methods. Interestingly, the use of HMMs with GMMs was a state-of-the-art technique in speech recognition in the last decade [117].

It was not until a decade ago that the research interest in machine learning approaches reborn, mainly due to the revolution of the deep learning paradigm [114]. Deep learning employs more complicated ANN architectures with several layers and millions of parameters, which are capable of learning difficult non-linear relationships in the data. These models are known as deep neural networks (DNN) [182] as they often comprise three to a hundred layers of non-linear transformations between the observable data and the target to be estimated. We can define the DNN model as a general non-linear function that, given a set of features $\mathscr{Y}$

obtained from the observable data, computes an estimate of the target data $\mathscr{X}$ as,

$$\widehat{\mathscr{X}} = f(\mathscr{Y};\theta), \tag{2.57}$$

where $f(\cdot)$ stands for the non-linear function and θ are the model parameters. These model parameters are trained using a dataset of pairs $(\mathscr{X},\mathscr{Y})$ by solving the next optimization problem,

$$\theta^* = \operatorname*{argmin}_{\theta} \mathscr{L}(f(\mathscr{Y};\theta),\mathscr{X}), \tag{2.58}$$

where $\mathscr{L}(\cdot,\cdot)$ is the cost function that measures the error between the DNN estimates and the actual targets. DNNs gained popularity thanks to the back-propagation algorithm proposed in [179] and the use of stochastic gradient descent [116]. These techniques allowed the efficient training of DNNs using large datasets. Moreover, in [83] Hinton *et al.* proposed an unsupervised pre-training of the DNN using Restricted Boltzmann Machines (RBM) that prevents the DNN to stuck in local minima during the supervised training. Currently, the more advanced regularization techniques [198, 92] along with the use of more powerful models alleviates the need for pre-training, making the training of DNNs quicker and more efficient. Moreover, deep learning has been also favored thanks to the availability of large amounts of data and high-performance computational resources as graphical processing units (GPUs). These advances have boosted the use of DNNs, which have shown state-of-the-art performance both in research and industry.

Deep learning for speech applications emerged from the use of DNN-HMM models in speech recognition [82], and then they extended to other areas like speech synthesis [122], speaker recognition [176], and speech enhancement and source separation [224]. In the next subsections, we will review the most common DNN networks for speech processing and the use of DNNs for single-channel and multichannel speech enhancement.

2.4.1 An overview of deep learning architectures

In this subsection, we review the main architectures currently used for DNNs in speech enhancement. We will focus on the three most common networks: feed-forward or fully-connected neural networks (FNN), recurrent neural networks (RNN), and convolutional neural networks (CNN). These architectures can be also used together to define more complex models, taking advantage of the particularities of each network.

Feedforward neural networks. Also known as fully-connected networks, these architectures are based on the classic multilayer perceptron (MLP) [88], but they usually comprise three or more layers of non-linear transformations. A feedforward layer can be mathematically expressed as a transformation between the vectors $\mathbf{z}^l$ and $\mathbf{z}^{l-1}$, where superscript l is the layer index. Let us assume that these vectors have dimensions M and N, respectively. Then, we can obtain $\mathbf{z}^l$ as

$$\mathbf{z}^l = f^l\left(\mathbf{W}^l\mathbf{z}^{l-1} + \mathbf{b}^l\right), \tag{2.59}$$

where $\mathbf{W}^l$ is the layer matrix of weights with dimensions $M \times N$, $\mathbf{b}^l$ is the layer bias vector of dimension M, and $f^l(\cdot)$ is the layer non-linearity, known as the activation function. The parameters of the layer are $\theta^l \in \left\{\mathbf{W}^l, \mathbf{b}^l\right\}$. As activation functions, different alternatives can be used, but the most common ones are the sigmoid, the hyperbolic tangent, or the rectified linear unit (ReLU) [247]. These non-linearities are element-wise applied to the vector components of each layer. Multiple feed-forward layers can be concatenated, achieving more powerful complex representations of the data. From an input layer $\mathbf{z}^0$, obtained as a feature representation of the data, the network transforms this vector at each layer until an output vector $\mathbf{z}^L$ is obtained, being L the total number of layers. These architectures are used when the input data can be transformed to feature vectors that are independent among them, obtaining a single output for each input vector.

Recurrent neural networks. These networks are preferred for processing sequences of vectors, as in the case of the speech signal (i.e. time-frequency transformations). This is due to the network capability for modeling temporal correlations in the data. The simplest way to obtain an RNN is by introducing a recurrent connection in a feedforward layer [178] as

$$\mathbf{z}_t^l = f^l\left(\mathbf{W}^l\mathbf{z}_t^{l-1} + \mathbf{V}^l\mathbf{z}_{t-1}^l + \mathbf{b}^l\right), \tag{2.60}$$

where t is the time sequence index and $\mathbf{V}^l$ is the weight matrix for the recurrence. We can establish an analogy between RNNs in deep learning and IIR filters in digital signal processing, as both employ recurrent relationships at the output of the model. The training of RNNs is performed by using the backpropagation through time (BPTT) algorithm proposed in [153].

Deep RNN architectures suffer from the vanishing gradient problem [86] due to its deepness both in the number of layers and the time. Thus, the gradient could become close to zero at lower layers and the correlations in long sequences become difficult to model, reducing the performance of these architectures. A proposed solution was the use of gated RNNs, as the well-known long-short term memory (LSTM) [86] or the gated recurrent unit (GRU) [26]. These RNNs are characterized by the use of internal gates, implemented as non-linear functions, that control the flow of the information between the layers and across time instants. Moreover, the layers have an internal memory that is also controlled. Therefore, the networks can learn long-term temporal dependencies in the sequential data. Currently, RNNs are one of the preferred architectures for speech processing [235].

It is noteworthy that the recurrences are not restricted to be forward (from previous time steps), but also backward (from future time steps). Recurrent networks that use both kinds of recurrences are known as bidirectional RNNs. Nevertheless, in this Thesis, we are interested in online speech processing algorithms. Therefore, we will focus on forward RNNs, which only can use present and past information.

Convolutional neural networks. CNNs differ from the previous fully-connected networks in that the weights at each layer are shared for multiple input components. This allows for computational efficiency, while these networks can exploit spatial structures in the data [115]. To show this, let's assume that the input data at each layer is three-dimensional, so it can be described as $z^l_{i,j,k}$, where the indices i, j, and k specifies each dimension. Moreover, the layer is comprised by a set of kernels parameters $\theta^{l,q} \in \left\{ w^{l,q}_{m,n,k}, b^{l,q}_{k} \right\}$, where q is the kernel index, and $w^{l,q}_{m,n,k}$ and $b^{l,q}_{k}$ are the weights and bias at each kernel. The variables m and n index the length and width of the kernel weights, which have dimensions $M \times N \times K$. Besides, the input data has a larger length and width than the kernel weights, but both the input and the kernels have the same depth K. The bias vector has a dimension of K, which matches the depth of the data. The convolutional layer performs the following operation with the input data at each layer,

$$z^l_{i,j,q} = f^l \left(\sum_{k=0}^{K-1} \sum_{m=\frac{-(M-1)}{2}}^{\frac{M-1}{2}} \sum_{n=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} w^{l,q}_{i+m,j+n,k} z^l_{i,j,k} + b^{l,q}_{k} \right), \tag{2.61}$$

which shares similarities with a classical convolutional operation. The data at the output of the layer has a depth of Q, which corresponds to the number of kernels at the layer. It can be observed that the weights at each kernel are reused for the different components of the input data. Also, learning the data spatial structures is possible thanks to the activation for a given component depends on the neighboring values, which are known as the receptive field. The number of parameters, that depends on the size and number of kernels at each layer, is usually lower than classical fully-connected DNNs and RNNs. The different outputs of each kernel, stacked as the depth of the layer output, are also known as the feature maps. This name refers to each kernel learns to extract some kind of complex features from the input data at the layer, exploiting spatial and shift-invariant properties.

The convolutional layers often include some kind of padding for the input data to operate in the edges. The stride in the convolution can also control the movement of the kernels in steps bigger than one. Moreover, a max-pooling operation can be applied at the output to decimate the signal [115]. On the other hand, there also exist transposed convolutional layers [87], which perform an upsampling operation to increase the size of the data. These layers are particularly useful if we desire to restore the dimensions of our original input data after several convolutional layers.

CNNs have gained popularity in the image and video processing area, but they can also be used in different speech processing tasks [6]. For example, they have been evaluated for time-domain speech signals, as in the case of the Wavenet [161] and Tacotron [230] architectures for speech synthesis, or for speech enhancement and source separation [53]. In this case, we can establish a relation between these convolutional layers and FIR filters, with the convolutional layer exploiting short temporal correlations. Similarly, the convolutional layer can operate in a time-frequency domain [163], exploiting the temporal and frequency correlations of neighboring bins, similarly to the case of images. We can also ensure online processing by avoiding the use of future frames when processing the current information, or working with convolutions in the frequency domain only.

2.4.2 DNN-based spectral mapping approaches

The spectral mapping is one of the two main approaches for speech enhancement using DNNs. The idea is that the DNN directly estimates the magnitude spectrogram of the clean speech signal, using features extracted from the noisy speech signal [238, 132]. Spectral mapping is a supervised regression problem, where an output target is predicted from the input features. Thus, the DNN directly models the complex non-linear relationships between clean and noisy speech signals.

A popular approach for spectral mapping is the use of the log-power spectrum (LPS) domain. The log-operation compresses the dynamic range of the speech signal, thus facilitating the training of the DNN. The approaches in [242, 243] proposed the use of normalized LPS features for the noisy and clean speech signals, that are obtained respectively as,

$$z_y(t,f) = \frac{\log|Y(t,f)|^2 - \widehat{\mu}_y(f)}{\widehat{\sigma}_y(f)}, \tag{2.62}$$

$$z_x(t,f) = \frac{\log|X(t,f)|^2 - \widehat{\mu}_y(f)}{\widehat{\sigma}_y(f)}, \tag{2.63}$$

where $\widehat{\mu}_y(f)$ and $\widehat{\sigma}_y(f)$ are, respectively, estimates for the mean and standard deviation for the LPS noisy speech training data. These estimates are computed, at each frequency, using the frames for all the data samples available in the training set. The features are normalized to zero mean and unit variance, benefiting the DNN training procedure. The cost function is the MSE loss between the target and estimated clean speech LPS,

$$\mathscr{L}_{\text{log-MSE}} = \frac{1}{T}\frac{1}{F}\sum_{t}\sum_{f}\left(z_x(t,f) - \widehat{z}_x(t,f)\right)^2 \tag{2.64}$$

where $\widehat{z}_x(t,f)$ is the estimated clean LPS. During the test phase, the estimated LPS vectors are de-normalized and expanded to the magnitude domain. Finally, the noisy speech phase is used to obtain the estimated clean STFT signal. The works in [242, 243] use a feed-forward DNN for vector-to-vector regression, where each noisy speech frame is mapped to its corresponding clean speech frame. To improve the performance, a temporal context of surrounding frames is stacked into an input vector to the DNN. Moreover, a noise-aware training (NAT) is performed by providing the DNN with an estimation of the noise from the initial frames of the speech sequence.

The LPS approach has been also followed by other works using more advanced architectures, as RNNs [200] or CNNs [52]. Also, we evaluated this approach for DNN spectral mapping on a dual-channel smartphone application [140], where the LPS of the noisy speech signal at both microphones is used at the network input. Recently, a maximum-likelihood training approach was proposed in [18] for the DNN training, modeling the estimation error using generalized Gaussian distributions. Both the DNN and the probabilistic model parameters are optimized using an iterative procedure.

The use of more advanced network models, as convolutional recurrent neural networks (CRN) with skipped connections [205], or gated residual networks with dilated convolutions [204], has favored the spectral mapping of the magnitude spectrogram. These works were extended in [206, 207] for the prediction of the real and imaginary parts of the clean speech STFT, showing improvements at low SNRs. Generative adversarial networks have also been evaluated for spectral mapping [152, 164]. Finally, other works directly deal with the mapping of the speech signals in the time-domain, using generally convolutional networks to process the noisy speech samples [164, 109].

2.4.3 DNN-based spectral masking approaches

The spectral masking approach differs from the spectral mapping in that it does not try to directly predict the clean speech features, but it uses the DNN to estimate a mask function. This mask function is applied to the noisy speech magnitude to obtain the clean speech magnitude as,

$$\left|\widehat{X}(t,f)\right| = \widehat{M}_x(t,f)\left|Y(t,f)\right|, \tag{2.65}$$

Table 2.1 DNN target mask function for different spectral masking approaches using the mask approximation. The subscripts r and i represents real and imaginary part of a complex variable, respectively.

Masking approach	**Expression**
Ideal binary mask (IBM) [231]	$M_x(t,f) = \begin{cases} 1 & \text{if } \frac{\lvert X(t,f)\rvert^2}{\lvert N(t,f)\rvert^2} > \text{thr}(f), \\ 0 & \text{otherwise} \end{cases}$
Ideal ratio mask (IRM) [157]	$M_x(t,f) = \left(\frac{\lvert X(t,f)\rvert^2}{\lvert X(t,f)\rvert^2 + \lvert N(t,f)\rvert^2}\right)^{\beta}$
Spectral magnitude mask (SMM) [229]	$M_x(t,f) = \frac{\lvert X(t,f)\rvert}{\lvert Y(t,f)\rvert}$
Phase sensitive mask (PSM) [44]	$M_x(t,f) = \frac{\lvert X(t,f)\rvert}{\lvert Y(t,f)\rvert} \cos\left(\theta_x(t,f) - \theta_y(t,f)\right)$
Complex IRM (cIRM) [236]	$M_x^r(t,f) = \frac{Y^r(t,f)X^r(t,f) + Y^i(t,f)X^i(t,f)}{Y^{2,r}(t,f) + Y^{2,i}(t,f)}$ $M_x^i(t,f) = \frac{Y^r(t,f)X^i(t,f) - Y^i(t,f)X^r(t,f)}{Y^{2,r}(t,f) + Y^{2,i}(t,f)}$

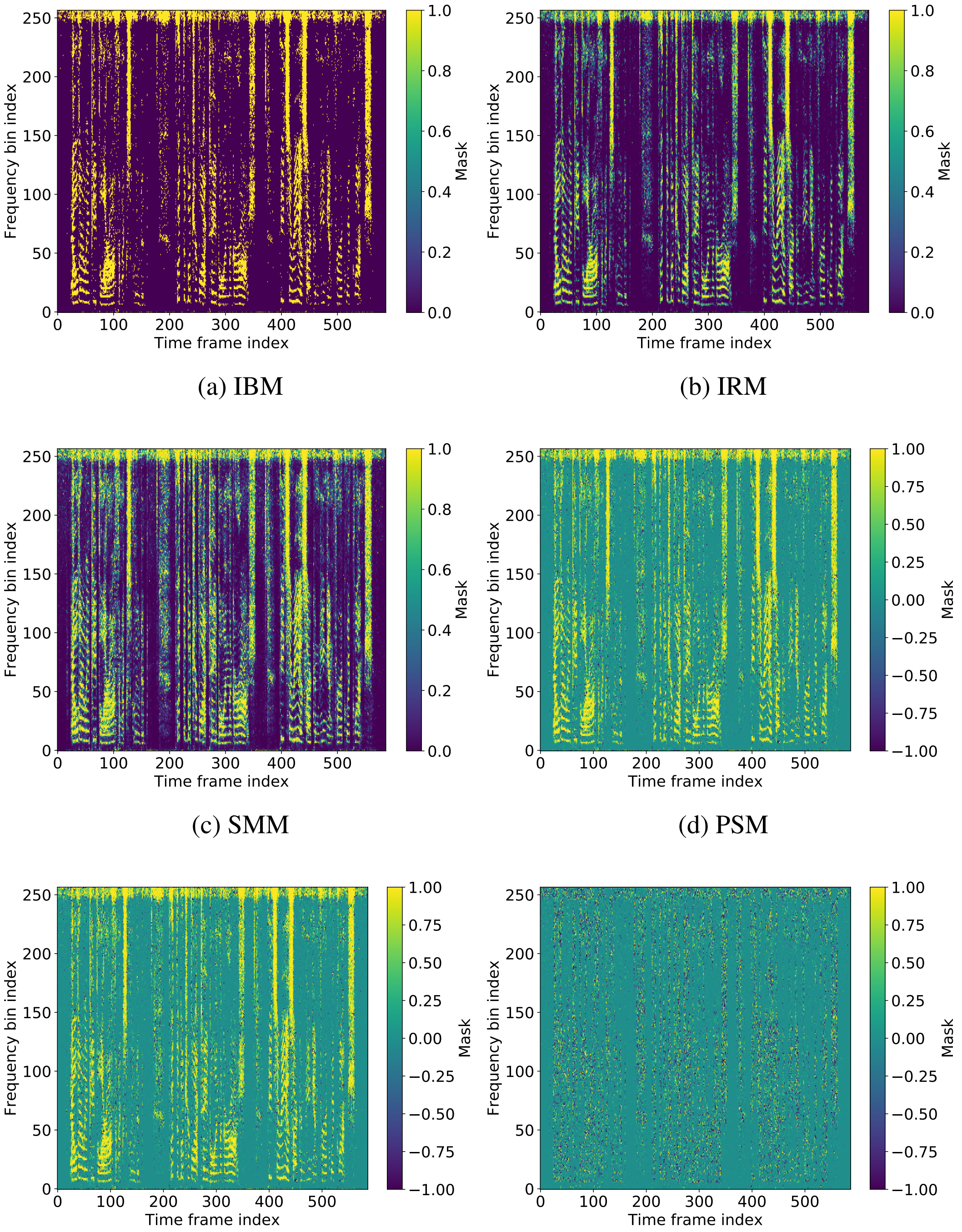
where $\widehat{M}_x(t,f)$ is the mask function predicted by the DNN. As can be observed, these methods are similar to classical speech enhancement approaches previously reviewed. Instead of using heuristics or statistical models, the mask is predicted from a trained DNN with a given loss function. Two approximations are commonly used for training the DNN mask estimators, the mask approximation and the signal approximation, which differ in the used target.

In the mask approximation, the DNN target is the mask function $M_x(t,f)$, which must be defined in advance. Table 2.1 summarizes the most common mask functions used in DNNs, while Fig. 2.2 depicts examples of these masks for a given noisy STFT spectrum. The loss function used also depends on the mask function. For binary masks (target takes values 0 or 1), the problem is addressed as a binary classification and the binary cross-entropy criterion is chosen,

$$\mathscr{L}_{\text{M-BCE}} = \frac{1}{T}\frac{1}{F}\sum_t\sum_f M_x(t,f)\log\widehat{M}_x(t,f) + (1 - M_x(t,f))\log\left(1 - \widehat{M}_x(t,f)\right). \tag{2.66}$$

On the other hand, the MSE loss function is commonly used for soft-valued masks,

$$\mathscr{L}_{\text{M-MSE}} = \frac{1}{T}\frac{1}{F}\sum_t\sum_f \left(M_x(t,f) - \widehat{M}_x(t,f)\right)^2. \tag{2.67}$$



(a) IBM

(b) IRM

(c) SMM

(d) PSM

(e) cIRM (real)

(f) cIRM (imag.)

Fig. 2.2 Examples of the most common target masks used in the mask approximation approach for DNN spectral masking speech enhancement. The clean speech signal is mixed with a pedestrian street noise at 0 dB SNR.

The concept can also be extended to complex mask functions [236], where the real and imaginary components of the mask are estimated. The complex mask is then applied to the noisy speech STFT, thus enhancing both the magnitude and the phase of the speech signal. The MSE loss function can also be used to train both components of the mask. The component separation avoids dealing with complex operations within the DNN.

In the signal approximation, the loss is directly obtained using the estimated clean speech magnitude after applying the mask function [91]. For example, if the MSE loss function is used the loss for magnitude estimation is obtained as,

$$\mathscr{L}_{\text{MSE}} = \frac{1}{T}\frac{1}{F}\sum_{t}\sum_{f}\left(\widehat{M}_x(t,f)\,|Y(t,f)| - |X(t,f)|\right)^2. \tag{2.68}$$

The advantage of this approach is that the target mask is not defined using an oracle expression, but the DNN is trained to estimate masks that optimize a given loss function. Apart from improving the speech enhancement performance, this also allows for the use of more complex loss functions. In the phase-sensitive approximation [44], the phase information is also included to improve the training performance. Moreover, the signal approximation has also been extended to complex spectral masking [201], commonly using the MSE loss for the real and imaginary spectrum components.

Finally, some recent works have extended the idea of spectral masking to other time-frequency domains than the STFT. An example is the TasNet proposed in [134, 135], which uses an encoder-decoder network that transforms from the time-domain to a time-feature domain and vice versa. A separation network then estimates the mask function that enhances the speech signal in this new domain, while the loss is measured in the time-domain after reconstruction. The encoder, decoder, and separation networks are jointly trained. Other works use predefined transformations as Gammatone filterbanks [34] instead of the encoder-decoder architecture. These approaches have been mainly explored for source separation, but they could be also applied to speech enhancement.

2.4.4 Integration of DNNs in multichannel beamforming algorithms

This subsection is devoted to the use of deep learning approaches along with multichannel beamforming techniques. In this Thesis, we will focus on these particular approaches for multichannel speech enhancement. There also exist other multichannel enhancement approaches that use DNN models, as the use of spatial features for enhancement or the use of DNNs for postfiltering purposes, which share similarities with the aforementioned single-channel approaches. Moreover, DNN-based multichannel approaches can be also applied to source separation. In [224, 68], the interested reader can find general reviews for the use of DNN-based multichannel speech enhancement and separation.

A general approach followed in recent works is the use of DNN mask estimators for the accurate estimation of the speech and noise spatial covariance matrices, which are then used to obtain the beamformer weights. This idea was first proposed by Heymann *et al.* [76, 78]. The DNN was trained to estimate speech and noise dominance masks at each microphone, which were then combined using a median operation to obtain the final speech and noise masks. These masks were used to obtain offline spatial covariance matrices for beamforming. While that work used IBM target masks, Zhang *et al.* [250] proposed the use of IRM masks to train the DNN estimators. In [45], the estimated IRM masks are also employed to further enhance the speech signal at the beamformer output. The approach in [241] proposed the use of a speech recognition loss in the mask estimator training. Regarding the input features for the DNN estimator, the cosine distances between the principal components of consecutive frames have been evaluated in [167, 169]. The DNN mask estimators have been also integrated with spatial statistical models to improve the estimation of the speech and noise masks [156]. On the other hand, recent works have extended the mask estimation to online speech enhancement systems [81, 20]. In [148], DNN models and spatial statistical models are combined for online processing. In this Thesis, we will mainly focus on the integration of DNN mask estimators in multichannel speech enhancement approaches.

A different approach is proposed in [159], which performs multichannel speech enhancement using an EM algorithm. In that work, the clean speech signals are estimated using MWF during the E-step, while the acoustic parameters are obtained during the M-step. The particularity about this approach is that it integrates DNNs to model the source spectra, using different DNNs for initialization and each one of the EM iterations. This approach applies DNN-based spectral mapping and the speech presence probability is not considered.

Finally, several works try to directly estimate the beamformer weights using DNNs. In [240], a deep beamforming network is used to estimate the weights to apply to the multichannel noisy speech signal. The work in [120] proposed the use of LSTM networks to obtain the time-domain spatial filter weights to be applied at each channel. This time-domain beamforming approach is also evaluated in the DeepBeam work [173]. This idea was extended in [181, 180] using trainable frequency-domain spatial filters. Meng *et al.* [151] proposed adaptive beamforming using LSTMs, where the weights are estimated at each time-frequency bin. This approach was also evaluated in [160] for multichannel end-to-end automatic speech recognition (ASR). On the other hand, a recent work [168] proposed the use of deep complex-valued neural beamformers that works directly with complex-valued back-propagation, thus avoiding the separation between real and imaginary components.

2.5 Summary

In this chapter, we have introduced the fundamentals of speech enhancement in the STFT domain. These fundamentals serve as the theoretical basis for the different contributions that will be later presented in this Thesis. To this end, we have first revisited the STFT technique for processing time-domain speech signals. We have also explained how to reconstruct the enhanced speech signal back to the time-domain using the ISTFT.

Then, the main single-channel classical speech enhancement algorithms were presented, remarking its implementation as time-frequency real gain functions. We have covered the spectral subtractive algorithms, the Wiener filtering approach, and the model-based Bayesian estimators. In this last category, we have focused on the MMSE-based estimators of the speech amplitude and the estimation of the a priori SNR. These classical methods require an estimation of the noise statistics, so we have also introduced classical single-channel noise estimators: from simple VAD to MS and MCRA tracking, and more advanced statistical estimators including MMSE, ML, and MAP estimators.

Next, the multichannel speech enhancement algorithms have been presented. We have focused on beamforming techniques, especially on data-dependent beamformers, which are formulated from the statistics of the underlying signals. The MVDR beamformer has been first introduced, and then we have extended it to the MWF approach and the use of postfiltering techniques. The estimation of the important acoustic parameters for the beamformer (the noise SCM and the RTF function) has been also covered, including its joint estimation using the EM algorithm.

Finally, the DNN-based approaches for speech enhancement have been introduced. We have first presented the main architectures used for speech processing, including the fully-connected, recurrent, and convolutional architectures. Then, we have explained the two main approaches for single-channel DNN-based speech enhancement in the STFT domain. These are spectral mapping, which directly tries to estimate the amplitude spectrum, and spectral masking, which estimates a mask function. Regarding the last approach, the differences between the mask and signal approximation have been covered. The last part of the section presented the integration of DNNs along with multichannel beamforming techniques. The main approaches described has been the use of DNN mask estimators for the estimation of the acoustic parameters, the integration of DNNs in an EM algorithm, and the direct implementation of the beamformer using network architectures.

Chapter 3

Experimental framework

This chapter presents the experimental framework used for the evaluation of the different speech enhancement algorithms proposed in this Thesis. Section 3.1 first describes the noisy speech databases used for the experimental evaluation of these contributions. Moreover, these databases were also employed for the training and validation of the DNN models that integrate the different techniques. Section 3.2 presents the objective evaluation metrics used to measure the performance of the different approaches. Finally, Section 3.3 shows the details for the training and validation of the DNN architectures, as well as the techniques used for its optimization and regularization.

3.1 Databases

This section describes the noisy speech databases used for the different evaluations and the training of the DNN models. Three databases (Aurora-2, VCTK-Noisy, and TIMIT-1C) contains single-channel noisy recordings in different environments. The TIMIT-2C-CT/FT and the CHiME-4 are multichannel databases which consider, respectively, a dual-microphone smartphone and a six-microphone tablet both used in noisy environments. Finally, the SMS-WSJ database is intended for the evaluation of scenarios with two overlapped speakers, who are recorded using an array of microphones in a room with reverberation.

3.1.1 Aurora-2

The Aurora-2 [165] is a simulated single-channel database of noisy speech utterances sampled at 8 kHz. The clean speech signals come from the TIDigits corpus [119], a database of connected digits spoken by American English speakers. Four different noisy signals are used to generate noisy speech samples: bus, babble, car, and pedestrian street. The tool FaNT

(Filtering and Noise adding tool) [85] is used to mix clean speech and noise signals at a given SNR level, which is selected among six possible SNRs in the range from -5 a 20 dB, with a 5 dB increase step.

The original training data is split into a training set of 8280 clean samples and a validation set of 160 clean samples. The training set is also divided into 24 subsets, one for each possible combination of SNR and noise type. Therefore, the clean speech samples are not repeated in the training set. On the other hand, the clean samples of the validation set are divided into 4 subsets, one for each noise type, but the same utterance is used at the different SNR levels. This gives a total of 960 utterances for the validation set. Finally, each of the original 1001 utterances of the test set is submitted to each of the 24 possible combinations of noise type and SNR. This results in a total of 24024 noisy speech utterances for evaluation.

3.1.2 VCTK-Noisy

The VCTK-Noisy is a simulated database of single-channel noisy speech utterances derived from the VCTK corpus [245] downsampled at 8 kHz. This clean speech database contains speech data from 108 native English speakers with various accents, with each speaker uttering about 400 sentences. The division between sets is done as follows: 72 speakers for training, 18 for validation, and the remaining 18 speakers are saved for testing. The noise signals are artificially added at six different SNRs from -5 to 20 dB (5 dB increase step). To this end, we recorded five-minute length noises at eight different locations. Four noises (babble, car, street, and mall) are used for training and validation, while the remaining noises (bus, cafe, pedestrian street, and bus station) are used for testing. Therefore, the test set contains *unseen* noises different from the ones in the training and validation set.

3.1.3 TIMIT-1C

The TIMIT-1C is a single-channel simulated noisy speech database. The clean speech signals are obtained from the TIMIT corpus [59, 113], downsampled at 16 kHz. Different utterances from the same speaker are concatenated to obtain samples with a duration between six and ten seconds, as long utterances are more appropriated for the evaluation algorithms. We finally have a total of 200 clean speech utterances for the training set and 50 utterances for each of the validation and test set. The number of speakers in each set is 195, 49, and 47, respectively, and speakers are not shared across sets. Moreover, the number of male and female speakers at each set is kept balanced.

The clean speech signals are mixed with different noise types using SNR ratios from -5 dB to 20 dB (with 5 dB increase step). Four noises are used for the training and validation sets: car, bus station, restaurant, and street. For the testing set, we have two different subsets: one subset using the same four noises as in training and validation (*seen* noises), and another subset with other four noises (*unseen* noises): bus, train station, cafeteria, and pedestrian street. Different noise recordings are used in each set, with a duration between one and five minutes. Each clean sample is used for each of the possible combinations of SNRs and noise types, which gives us a total of 4800, 1200, and 2400 noisy speech samples for training, validation, and testing, respectively.

3.1.4 TIMIT-2C-CT/FT

The TIMIT-2C-CT/FT are two different databases of simulated dual-channel noisy speech recordings from a dual-microphone smartphone, with a sample rate of 16 kHz. Two different databases were developed, depending on how the device is used. The close-talk (CT) database simulates the case where the loudspeaker of the smartphone is placed at the ear of the user, and the far-talk (FT) database, on the other hand, simulates the case where the user holds the device at a distance from her/his face. The number of clean speech signals and noisy speech utterances for the different sets (training, validation, and test) are the same as in the TIMIT-1C database, as well as the same SNRs are evaluated. The number of speakers in each set is 54, 14, and 14, respectively, with speakers not shared across sets, and the gender of speakers kept balanced. The dual-channel noisy recordings are simulated using these clean speech samples, dual-channel noise recordings, and simulated dual-channel acoustic impulse responses (AIR). The procedure to obtain the dual-channel noisy samples and the different elements needed is described below. This procedure is valid for both CT and FT databases.

To simulate the recordings, a methodology similar to the one considered in [126, 129] is followed. The clean speech signals are first filtered using dual-channel AIRs and then real dual-channel noises are added simulating the different SNRs. Thus, a certain utterance is generated as,

$$y_1(m) = h_1(m) * x(m) + Gn_1^{'}(m), \tag{3.1}$$

$$y_2(m) = h_2(m) * x(m) + Gn_2^{'}(m), \tag{3.2}$$

where $h_1(m)$ and $h_2(m)$ are the AIRs for the primary and secondary channel of the smartphone, respectively, and G is the gain to apply to the noise segments $n_1^{'}(m)$ and $n_2^{'}(m)$ to obtain the desired SNR. Four reverberation environments, with different reverberation times, and eight noises are evaluated. Each type of noise is assigned to a reverberation environment, yielding a total of eight different acoustic environments (including both reverberation and noise). Four acoustic environments are used in all the sets (*seen* conditions), while the other four are used only for testing (*unseen* conditions). Table 3.1 show the matching applied

Table 3.1 Predefined acoustic environments in TIMIT-2C-CT/FT: each environment combines a reverberation environment with a given noise. The noises are classified in that seen in the training set (S) and the ones that are used only on test and are unseen in training (U).

Reverberation	**Noise**
No reverb.	Car (S), Street (U), Pedestrian street (U)
Low reverb.	Bus (S), Cafe (U)
Medium reverb.	Babble (S), Bus station (U)
High reverb.	Mall (S)

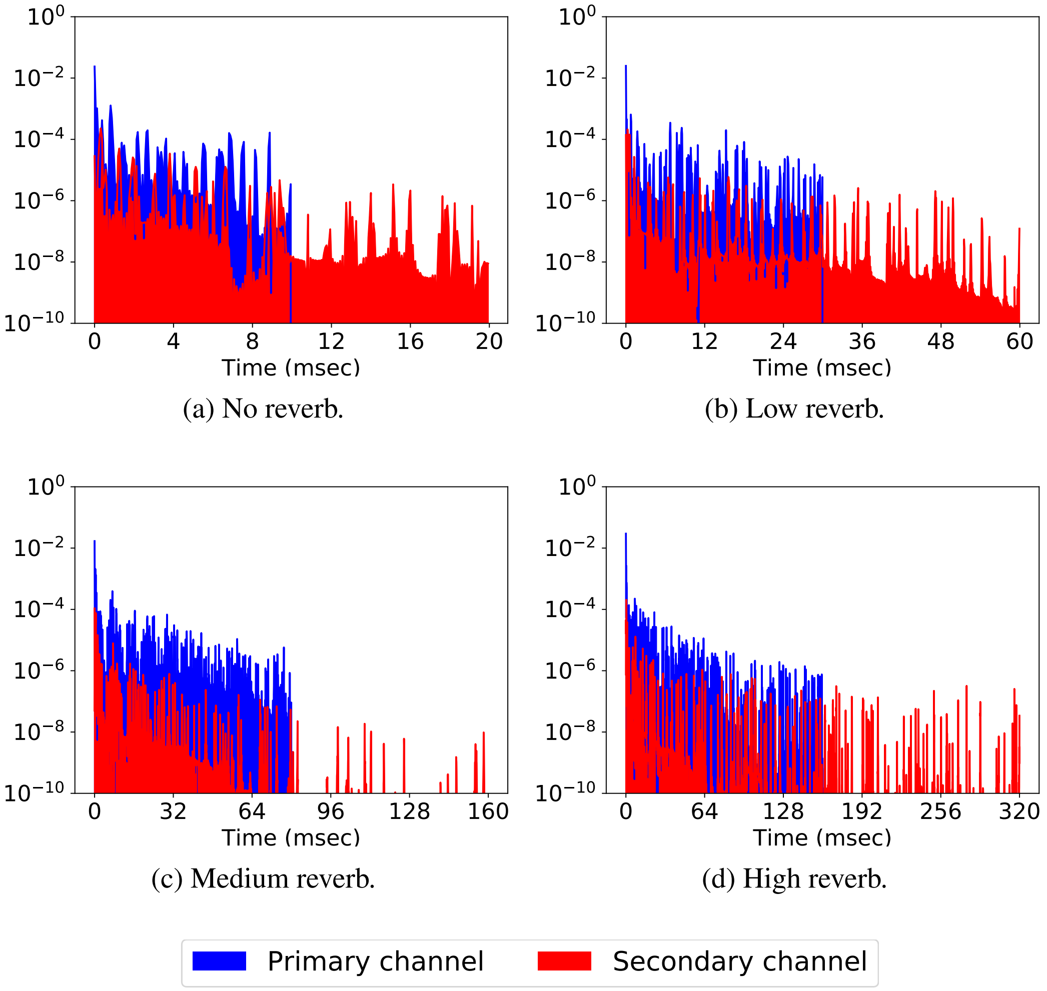
between reverberation environments and noises, and also the distribution of the noises in *seen* and *unseen* conditions. The different combinations between noises and SNRs give a total of 4800, 1200, and 2400 noisy speech samples for training, validation, and test, respectively, as in the case of the TIMIT-1C database.

A Motorola Moto G smartphone was employed to capture the noises and simulate the AIRs. This smartphone has a primary microphone at its bottom and a secondary one at its top, with a distance of 13 cm between them. First, the dual-channel noise recordings were done at the eight noisy environments in both CT and FT modes, obtaining five-minute length recordings. For *seen* noises, the files are split into 3 minutes for training, 1 minute for validation and 1 minute for test. To obtain the AIRs, we acquired paired clean speech signals using a close-talk high-quality cardioid microphone, in one case, and a dual-channel microphone in the other. These recordings were synchronized later. A sampling frequency of 48 kHz was selected to have a good temporal resolution. The AIRs $h_j(m)$ were estimated by assuming the close-talk microphone as the ground-truth clean speech signal $s(m)$, while the smartphone recordings $x_j(m)$ $(j = 1,2)$ are approximated as filtered versions of $s(m)$ using FIR filters $h_j(m)$. These AIRs model both the environment and the microphone responses. In order to obtain realistic sparse AIRs $h_j(m)$, their estimation is formulated as a least-square (LS) problem with sparse coefficients enforced by using $\mathscr{L}_1$-norm. First, the LS-based cost function is defined as

$$J(\mathbf{h}_j) = \mathbf{h}_j^\top \mathbf{R}_s \mathbf{h}_j - \mathbf{h}_j^\top \mathbf{r}_{x_j s} - \mathbf{r}_{x_j s}^\top \mathbf{h}_j, \tag{3.3}$$

where $\mathbf{h}_j$ is an $N \times 1$ vector with the AIR coefficients, $\mathbf{R}_s$ is the $N \times N$ autocorrelation matrix of $s(m)$ and $\mathbf{r}_{x_j s}$ is the $N \times 1$ cross-correlation vector between $x_j(m)$ and $s(m)$. We define $\mathbf{h}_j^*$ as the value of the AIR that minimizes $J(\mathbf{h}_j)$. Finally, $\mathbf{h}_j$ is obtained as

$$\mathbf{h}_j = \underset{\mathbf{h}_j}{\operatorname{argmin}} \left\{ (1-\lambda) \frac{J(\mathbf{h}_j) - J(\mathbf{h}_j^*)}{\left|J(\mathbf{h}_j^*)\right|} + \lambda \frac{\|\mathbf{h}_j\|_1}{\|\mathbf{h}_j^*\|_1} \right\}, \tag{3.4}$$



(a) No reverb.

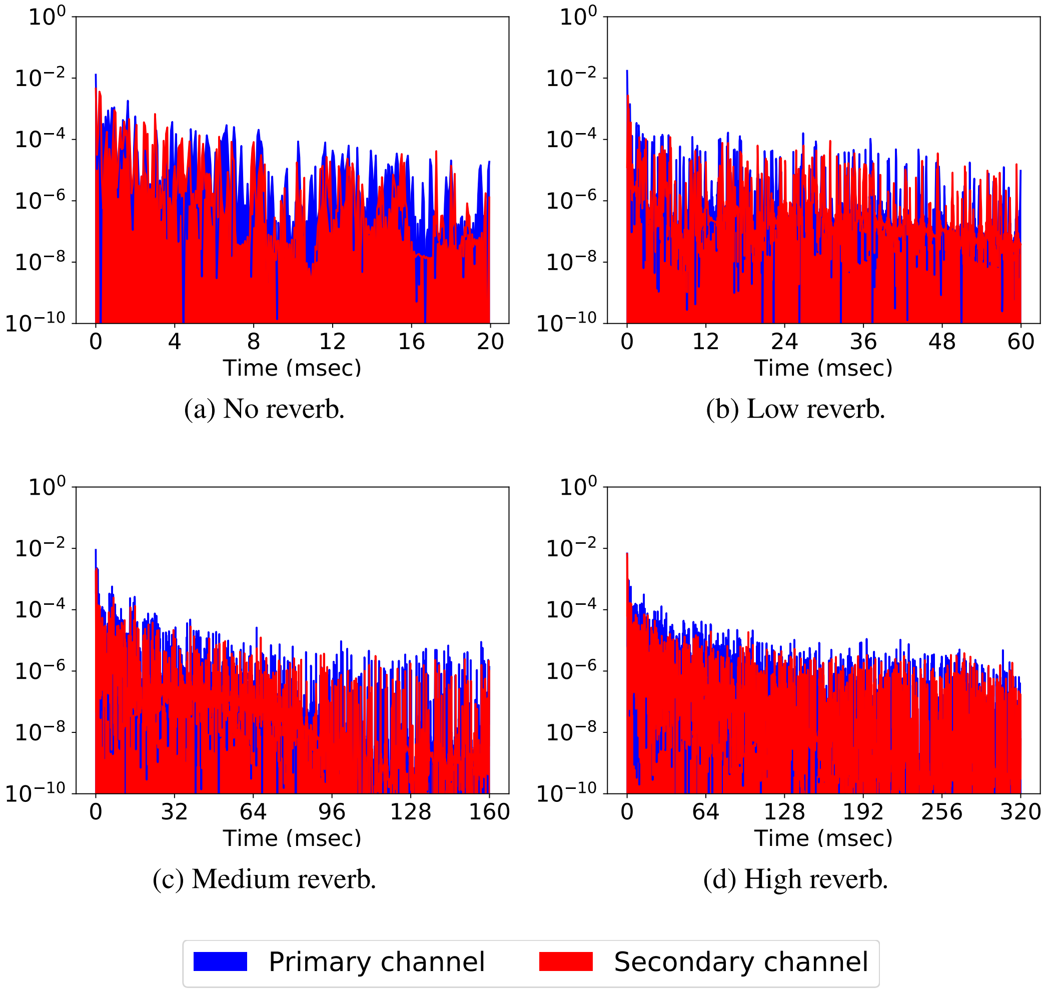
(b) Low reverb.

(c) Medium reverb.

(d) High reverb.

Fig. 3.1 Semi-logarithmic plot of AIR examples for the different reverberant environments in close-talk (CT) conditions.

where $\|\cdot\|_1$ stands for $\mathscr{L}_1$-norm and $\lambda = 0.15$ is a trade-off factor between LS minimization and filter sparseness. The minimization problem in (3.4) has not a closed-form solution, but it is a convex equation, so it can be solved using either convex optimization or gradient-based methods. A total of 30 AIR pairs are obtained for each reverberant environment and smartphone use mode. They are split into 16 AIRs for the training set, 4 AIRs for validation, and the remaining 10 AIRs for testing. Fig. 3.1 and 3.2 show examples of AIRs for the different reverberant environments in CT and FT conditions, respectively. It can be observed that the environments differ in the length of the acoustic responses and their power decay with time. More reverberant environments present longer AIRs and a slow power decay. As can be observed, the main difference between CT and FT is that the power difference



(a) No reverb.

(b) Low reverb.

(c) Medium reverb.

(d) High reverb.

Fig. 3.2 Semi-logarithmic plot of AIR examples for the different reverberant environments in far-talk (FT) conditions.

between the primary and secondary channel responses is clearer in the CT mode, while both channels have similar power response in the FT mode.

3.1.5 CHiME-4

CHiME-4 [221] is a multichannel noisy speech database, part of the 4th CHiME Speech Separation and Recognition challenge. The database comprises six-channel tablet recordings in four noisy environments from different speakers. An scheme of the microphone positions in the tablet is depicted in Fig. 3.3. Both real and simulated data are provided for training and evaluation purposes. The real data was recorded from 12 US English speakers reading phrases prompted in the tablet. The recordings were done in several public areas: bus, cafeteria, street, and pedestrian area. On the other hand, the simulated data was obtained by

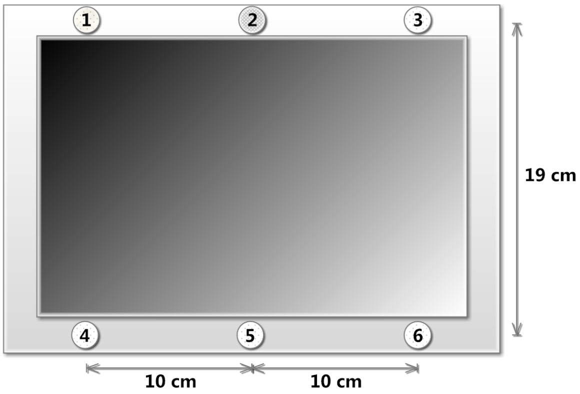


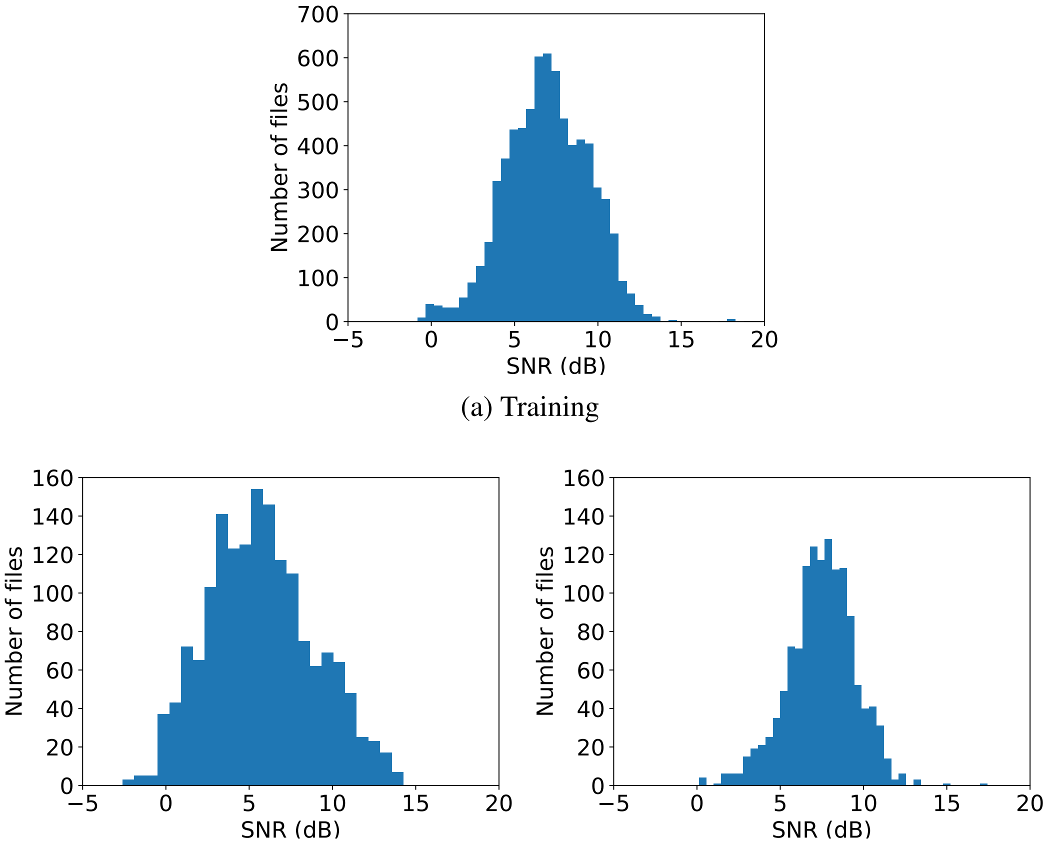
Fig. 3.3 Scheme of the 6-microphone tablet used to develop the CHiME-4 database. All microphones are faced forward, except microphone number 2 that faces backwards.

artificially mixing clean speech data, convolved with estimated impulses responses for each microphone, with background noises recorded in the different noisy environments previously indicated. The audio data is provided as 16-bit stereo WAV files sampled at 16 kHz.

The database consists of training, development, and evaluation sets. The training set is formed by 1600 real utterances from four speakers. Moreover, this set also includes 7138 simulated utterances from 83 speakers of the Wall Street Journal Corpus (WSJ0) [58] SI-84 training set. The development and evaluation sets consist of 3280 and 2640 utterances, respectively, with four different speakers in each set. Half of the utterances in the development and evaluation sets are simulated data and the other half are real data. In addition, the different noisy environments are represented equally both in the real and simulated data. The SNRs of the noisy simulated data were selected to be similar to the ones measured in the real data. These SNRs are approximately in the range between 0 and 15 dB. Fig. 3.4 represent the histograms of the SNRs measured in the simulated data utterances of each set. The 5th-microphone was used in the measurements, as it is usually selected as the reference (main) microphone.

3.1.6 SMS-WSJ

The Spatialized Multi-speaker Wall Street Journal (SMS-WSJ) [38] is a multichannel database of overlapping speakers for the evaluation of source separation algorithms. This database contains utterances of two speakers that are recorded using a 10 cm radius circular array of six microphones in a room with reverberation. The training, development, and test set contains, respectively, 30000, 500, and 1500 multichannel mixture utterances with a sample rate of 8 kHz. The clean speech signals are obtained from three non-overlapping WSJ sets: si294, dev93, and eval92. The speakers do not overlap between different sets.



(a) Training

(b) Development

(c) Evaluation

Fig. 3.4 Histogram of the SNRs measured on the noisy speech files in the different sets of the CHiME-4 simulated corpus.

Each utterance is created by randomly choosing two clean speech utterances from different speakers. These utterances are convolved with the room impulses responses (RIR) of the microphone array. The RIRs are generated using the Image method [8]. The RIR generator randomly samples the room dimensions, the microphone array and speaker positions, and the reverberation time (between 200 and 500 ms). The final noisy utterance thus consists of the sum of both reverberated utterances. The shorter of the two samples is padded with zeros to match the duration of the other one. This padding is done randomly at the start and end of the utterance. In addition, white Gaussian noise, with an SNR between 20 and 30 dB, is added to the multichannel mixture to simulate the sensor noise at each microphone. Therefore, the main distortion source is the interference between speakers.

3.2 Evaluation objective metrics

The different objective evaluation metrics used in this Thesis are described in this section. These objective metrics show a high correlation with subjective tests using human listeners. Besides, they are easy to evaluate and do not require extensive and costly tests with real listeners. Different objective metrics are used, each of them focusing on the evaluation of a specific aspect of the enhanced signal. These aspects are perceptual speech quality, speech intelligibility, signal distortion, and, additionally, the recognition accuracy of automatic speech recognition (ASR) systems using enhanced speech signals.

3.2.1 Perceptual Evaluation of the Speech Quality

The Perceptual Evaluation of the Speech Quality (PESQ) [3] is one of the most extended metrics to evaluate the perceptual speech quality. This metric was recommended in 2000 by the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T) to evaluate the speech quality of the speech signal that goes through a telecommunication network. Thus, this measure assess a wide variety of distortions introduced by the narrowband speech coders. Despite its original purpose, the PESQ algorithm has been widely adopted for the evaluation and comparison of speech enhancement algorithms in terms of speech distortion, noise reduction, and overall speech quality.

The PESQ measure computes an objective quality score by comparing the degraded speech signal (distorted or enhanced speech signal) to the reference speech signal. The level of both time signals is first equalized to a standard listening level, and after that, they are filtered using a filter with a similar response to the telephone headset. Then, the signals are time-aligned and processed using an auditory transform to obtain the loudness spectra. Finally, two disturbances terms, namely the symmetric d_s and asymmetric d_a disturbances, are computed using the original and degraded loudness spectra. The PESQ score is then obtained as

$$d_{\text{PESQ}} = 4.5 - 0.1 d_s - 0.0309 d_a \tag{3.5}$$

which yields scores between -0.5 and 4.5. The higher the score the best the speech quality. A more detailed explanation of how to obtain these disturbance terms will be given in Chapter 7.

The PESQ algorithm was extended to wideband speech signals [4], with a sampling rate of 16 kHz. This new version adapts the algorithm from 8 kHz to 16 kHz extending the auditory transform to higher frequencies. Moreover, the result represents a mean opinion score (MOS) [1] between 1 and 5, with high scores indicating better speech quality perceived by human listeners.

3.2.2 Short-Time Objective Intelligibility

The Short-Time Objective Intelligibility (STOI) [202] is a speech intelligibility measure commonly used for the evaluation of speech enhancement algorithms. The reason is that it correlates well with intelligibility subjective tests in a wide variety of acoustic scenarios, including classical time-frequency weighting speech processing algorithms. The STOI score is computed as follows. First, the original and degraded speech signals are resampled at 10 kHz. Then, a VAD detector is used to remove the silent frames, and the STFT is computed for both speech signals. The one-third octave band spectra are obtained from the STFTs as

$$A(t,j) = \sqrt{\sum_{f=f_1(j)}^{f_2(j)} |X(t,f)|^2} \tag{3.6}$$

where $f_1(j)$ and $f_2(j)$ are the first and last frequency indices corresponding to the j band. The same procedure is done to obtain $\widehat{A}(t,j)$ for the degraded clean speech. Short-time temporal envelope vectors are defined at each band, for example in the case of the clean speech signal as,

$$\mathbf{a}(t,j) = \begin{bmatrix} A(t-N+1,j) & A(t-N+2,j) & \cdots & A(t,j) \end{bmatrix}^\top \tag{3.7}$$

where $N = 30$ corresponds to a temporal window of 384 ms. Similarly, $\widehat{\mathbf{a}}(t,j)$ can be defined for the degraded speech signal. The vector $\widehat{\mathbf{a}}(t,j)$ is normalized and clipped using the following expression,

$$\widehat{\mathbf{a}}'(t,j) = \min\left(\frac{\|\mathbf{a}(t,j)\|_2}{\|\widehat{\mathbf{a}}(t,j)\|_2} \cdot \widehat{\mathbf{a}}(t,j), \left(1 + 10^{-0.75}\right) \mathbf{a}(t,j) \right) \tag{3.8}$$

where $\|\cdot\|_2$ means $\mathscr{L}_2$-norm and $\min(\cdot,\cdot)$ is the element-wise minimum operator between vectors. The intermediate intelligibility measure for a pair of normalized envelope vectors is then defined as the linear correlation between them, computed as

$$d(t,j) = \frac{\left(\mathbf{a}(t,j) - \mu_a(t,j)\right)^\top \left(\widehat{\mathbf{a}}'(t,j) - \mu_{\widehat{a}'}(t,j)\right)}{\|\mathbf{a}(t,j) - \mu_a(t,j)\|_2 \cdot \|\widehat{\mathbf{a}}'(t,j) - \mu_{\widehat{a}'}(t,j)\|_2} \tag{3.9}$$

where $\mu_a(t,j)$ and $\mu_{\widehat{a}'}(t,j)$ are the sample mean of the vectors $\mathbf{a}(t,j)$ and $\widehat{\mathbf{a}}'(t,j)$, respectively. Finally, the STOI score d_{STOI} is obtained by averaging the different values $d(t,j)$ obtained per frame and band. The STOI score is a value between 0 and 1, the higher the score the better the speech intelligibility of the degraded clean speech signal.

An extended version of STOI, named ESTOI, was proposed in [95] to improve the intelligibility predictions under highly fluctuating noises, as in the case of competitive talkers. ESTOI has shown to outperform STOI and other intelligibility predictors, in terms of correlation with subjective tests, under these noise conditions with high fluctuations. Moreover, its performance is similar to STOI with other noise types. The ESTOI measure computation is similar to the one in STOI. First, the one-third octave bands coefficients are obtained. Then, the short-time spectrogram matrix for the clean speech signal is defined as

$$\mathbf{A}(t) = \begin{bmatrix} A(t-N+1,1) & A(t-N+2,1) & \cdots & A(t,1) \\ \vdots & & & \vdots \\ A(t-N+1,J) & A(t-N+2,J) & \cdots & A(t,J) \end{bmatrix}, \tag{3.10}$$

with J the number of one-third octave bands. The matrix $\widehat{\mathbf{A}}(t)$ for the degraded speech signal is defined in a similar way. These matrices are processed as follows. First, the rows of the matrices are normalized by zero-mean and unit-variance. Then, the columns of the resulting matrices are also normalized by zero-mean and unit-variance, which gives the processed matrices $\mathbf{A}_n(t)$ and $\widehat{\mathbf{A}}_n(t)$. Let us denote $\mathbf{a}_n(t,n)$ and $\widehat{\mathbf{a}}_n(t,n)$ to the column vectors of the matrices $\mathbf{A}_n(t)$ and $\widehat{\mathbf{A}}_n(t)$, respectively, with $n = 1,...,N$. The ESTOI measure is then computed as

$$d_{\text{ESTOI}} = \frac{1}{NT} \sum_t \sum_n \mathbf{a}_n^\top(t,n)\widehat{\mathbf{a}}_n(t,n). \tag{3.11}$$

The ESTOI score is also a value between 0 and 1, with higher scores indicating better speech intelligibility.

3.2.3 Signal-to-Distortion Ratio

The signal-to-distortion ratio (SDR) was a metric proposed in the toolkit *BSS_eval* [220] for the evaluation of source separation algorithms. This toolkit proposed different versions for the metric which can factor out some effects that are not important to evaluate distortion, as different gains in the signal or time-invariant filtering. The simplest version directly measures an SNR between the clean $x(m)$ and enhanced $\widehat{x}(m)$ speech signals in the time domain as

$$d_{\text{SDR}} = 10\log_{10} \frac{||x(m)||_2^2}{||x(m) - \widehat{x}(m)||_2^2}. \tag{3.12}$$

The recent work in [177] showed some problems when using this simple version of the SDR, which was the most extended in the source separation community. Instead of the standard SDR, they proposed the use of a scale-invariant SDR (SI-SDR), that re-scale the clean speech signal to compensate for different gains in the speech signals. The SI-SDR is then computed as

$$d_{\text{SI-SDR}} = 10\log_{10}\frac{\|\alpha x(m)\|_2^2}{\|\alpha x(m) - \widehat{x}(m)\|_2^2} \tag{3.13}$$

where

$$\alpha = \frac{\sum_m x(m)\widehat{x}(m)}{\|x(m)\|_2^2} \tag{3.14}$$

is the scale factor obtained by minimizing the $\mathscr{L}_2$-norm of the residual component.

3.2.4 Speech Distortion index

The speech distortion (SD) index [11] is a metric that allows us to directly measure the distortion effect of a speech processing algorithm, as beamforming, on the clean speech signal. Let us first define $\widetilde{x}(m)$ as the time-domain signal resulting from directly processing the clean speech signal $x(m)$ using the speech enhancement algorithm. As can be observed, this signal is different from the enhanced speech signal $\widehat{x}(m)$, which results from processing the noisy speech signal $y(m)$. The SD index is measured segmentally across the speech signal, in such a way that the SD value at the i-th segment is obtained as

$$d_{\text{SD}}(i) = \frac{\sum\limits_{m=(i-1)N}^{iN-1} |x(m) - \widetilde{x}(m)|^2}{\sum\limits_{m=(i-1)N}^{iN-1} |x(m)|^2}, \tag{3.15}$$

where N is the number of samples per segment. The segmental values are averaged to obtain the final SD index. The lower the SD value, the lower the speech distortion. The authors of [209] also proposed to remove the silence frames from the evaluation by calculating the median of the segment-wise signal power and removing those segments with a power 15 dB lower than that median. We used that version in our evaluations.

3.2.5 Word Error Rate

Speech enhancement algorithms can also be used as the front-end of an automatic speech recognition (ASR) system to improve recognition accuracy. A common measure to evaluate the performance of ASR systems is the word error rate (WER) [166]. To compute this metric, let us suppose a speech signal that contains N_T words. The speech signal is transcribed into text by an ASR system, which has the following errors during that transcription: N_S words are misrecognized (substituted), N_D words are deleted, and N_I words are inserted. The WER is then obtained as

$$d_{\mathrm{WER}} = \frac{N_S + N_D + N_I}{N_T}. \tag{3.16}$$

This measure is commonly expressed as a percentage, with smaller values indicating better recognition performance. The ASR performance also depends on the back-end used, which involves the acoustic and language modeling [166, 117]. Nevertheless, the use of a good performance ASR back-end can allow us to compare different speech enhancement front-ends in terms of the recognition accuracy.

3.3 Training of DNN architectures

The speech enhancement algorithms proposed in this Thesis integrate deep neural networks, which must be trained in advance using speech databases. In this section, we explain the general methodology followed during the training of DNNs in this Thesis.

The different algorithms proposed in this Thesis are designed using Python. This programming environment also allows the use of deep learning libraries and frameworks. We have worked with two deep learning frameworks: Tensorflow [5] and Pytorch [199]. Although each of them has its pros and cons, these frameworks are characterized by the possibility of designing different network architectures for data processing. Gradient computation, required for backpropagation, is already embedded in those frameworks, thus easing the design of new deep learning procedures. Also, they allow us to work at different levels, from basic operations and structures to high-level pre-designed networks. Moreover, they include a variety of algorithms and procedures for the optimization and regularization of the training setup. Finally, these frameworks allow us to run the implementation on either CPU or GPU, the latter reducing the computational time, especially during the DNN training.

The DNN training using these frameworks requires to previously define the following blocks: the data pre-processing to obtain the input and target features, the network architecture, and the loss function. The training set of the corresponding database is used to train the network parameters. The utterances of the training set are randomly grouped into mini-batches, and a single mini-batch is fed into the network at each step. Then, the average loss on the mini-batch is obtained and the gradients are computed by using the backpropagation algorithm. These gradients allow us to adjust the parameters of the network by using a specific optimizer. In this Thesis, we choose the ADAM optimizer [105]. This approach improves the classic stochastic gradient descent (SGD) with the use of an adaptive updating coefficient based on an approximation of the second-order gradients and on information of previous steps. The default parameters proposed in [105] are used. Once all the utterances (also known as samples) of the training set have been used, an epoch has concluded, and the training set is reorganized in new mini-batches for the next epoch. The training can be finished after a maximum number of epochs or using a stopping criterion.

The training performance can be improved using regularization techniques, which reduce the needed time for training and yield better generalization capabilities of the DNN. In this Thesis, we use two different regularization techniques:

- **Dropout** [84, 198]: Dropout is a technique that consists of randomly putting to zero a certain percentage of input values for a DNN layer. The dropout can be used in all the hidden layers or only in some of them. The general idea of this procedure is to prevent complex co-adaptations between the data and the architecture. Moreover, this technique is a way to introduce noise in the processed data. Thus, Dropout can act as a data augmentation technique, improving generalization. Another point of view is that Dropout allows for a statistical averaging of different network architectures, as we are de-activating units randomly. This procedure gives different combinations, thus improving the robustness of the final network architecture. The Dropout is only applied during training, using a scaling factor to avoid the issues derived from the unit dropping.

- **Early-stopping** [172]: Early-stopping is an easy procedure for training regularization. It uses a validation set (or development set) different from those used during training and evaluation. Also, a patience counter is defined, which is initially set to zero. After each epoch, the DNN is evaluated in the validation set using the defined loss, but backpropagation is not performed. The averaged loss in the set is obtained and compared with the one obtained in the previous epoch. If the new loss is lower than the previous one, we save the network parameters as the best network, set the patience counter to zero, and continue the training. Otherwise, we increase the patience counter, and the training continues only if the patience is lower than a maximum value predefined. This maximum patience acts as the stopping criterion along with the maximum number of epochs. Once training is finished, we keep the best parameters obtained (the ones yielding the lowest loss in the validation set). The early-stopping criterion allows training stopping in a configuration that can generalize well to data not seen during the training. Besides, this reduces training time. In general, we use a maximum number of 200 epochs and a patience of 20 epochs in the different algorithms proposed in this Thesis.

3.4 Summary

In this chapter, we have introduced the experimental framework used in this Thesis. We have first described the features of the noisy speech databases used for training and testing. These databases include single-channel and multichannel noisy recordings using different devices in reverberant and noisy environments. The case of interfering speakers is also considered. Then, we have presented the objective metrics used to evaluate and compare the different methods. These metrics include PESQ for speech quality, STOI and ESTOI for speech intelligibility, and SDR and SD index for signal distortion evaluation. Moreover, WER was also used to assess the accuracy of ASR systems, where the speech enhancement algorithm is used as part of the front-end processing.

To conclude this chapter, we have also given the details of the setup used for training the different DNN architectures in our proposals. We have first pointed out the programming framework used to work with these deep learning models, including deep learning libraries. Then, we have explained the training procedure and the optimization techniques used. Finally, some regularization techniques intended to improve the training performance have been presented: the dropout technique and the early-stopping criterion. Although other regularizations are also possible, the employed methods have provided a suitable convergence behavior.

Chapter 4

Dual-channel speech enhancement algorithm based on extended Kalman filter RTF estimation

Smartphones are the most widely used mobile devices across the world, enabling a lot of user services such as voice communications or Internet access. This has extended the use of speech-related services such as mobile phone calls, voice assistants or navigators. These devices are frequently used in reverberant and noisy environments. Therefore, the speech signal quality and intelligibility can be degraded, reducing the performance offered by these speech services. Moreover, these services commonly need low-latency and online processing of the speech signal to run in real-time. Current smartphone devices embed several microphones, being dual-channel microphones a widely used configuration, comprising a primary microphone in the bottom and a secondary microphone on the top or back of the device. This allows for the use of multichannel processing techniques to increase the noise reduction performance. Although beamforming techniques are often used in multi-microphone devices, their performance can be quite limited on dual-microphone smartphones due to the reduced number of microphones, their particular position on the device, and its close separation [212]. A possible solution is the use of postfiltering techniques on dual-microphone scenarios [67, 253], increasing the performance obtained by a standalone beamformer.

Other alternatives have been proposed for speech enhancement in dual-microphone smartphones which exploit the properties of the dual-channel signals. The approach of some works consists of the use of single-channel filters for the primary channel using dual-channel statistical information. For example, the power level difference (PLD) algorithm proposed in [96] exploits the level difference between the speech signals at each channel when the smartphone is used in close-talk (CT) conditions (i.e., the smartphone is placed at the ear of the user). A Wiener filter is calculated from an estimate of the single-channel noise statistics using both microphones. In [214], the PLD approach was extended with the integration of the speech presence probabilities in the noise estimation and filter design. For far-talk (FT) conditions (i.e., the user holds the device at a distance from her/his face), the proposal in [158] exploits the spatial properties of the noisy speech and noise signals. This information is used along with a single-channel speech presence detector to estimate the noise statistics and the SNR needed. That work was extended in [97] to multichannel devices. The dual-channel information has been also exploited for feature enhancement in ASR systems [126, 129] intended for smartphones. The noise estimation using both microphones was improved by the use of unscented Kalman filters in [128], showing better performance than other single- and dual-channel estimators. Finally, the use of DNNs has been also explored in dual-microphone smartphones. For example, the noise-robust speech recognition on smartphones with DNNs was investigated using missing-data mask estimation [127] and vector Taylor series noise estimation [130] for the case of feature enhancement. Moreover, in our preliminary work [140] we proposed a dual-channel DNN speech enhancement algorithm based on spectral mapping for smartphones. Recently, this idea was evaluated in [208] for spectral masking using phase-sensitive masks and dual-channel features with a convolutional-recurrent neural network.

In this chapter, we detail our proposal for a dual-microphone speech enhancement framework specially intended for smartphone devices. The algorithm is based on an MVDR beamformer plus a postfilter approach which increases the noise reduction performance with low speech distortion. The proposed framework includes three main contributions to the state-of-the-art. The first contribution is the development of a novel RTF estimator based on an extended Kalman filter (eKF) algorithm suitable for RTF tracking in noisy and reverberant environments. The eKF estimator uses a priori knowledge about the RTF and noise statistics, and it has the advantage that no assumptions about the clean speech are needed, as in many of commonly used sub-space methods. The second contribution is the use of dual-channel information for the estimation of the single-channel speech statistics used by the postfilter. Finally, the last contribution is the development of a noise estimator based on speech presence probability (SPP) which exploits the dual-channel properties of the noisy signals in CT and FT conditions. We propose two different approaches for the SPP estimator, using either statistical spatial models or DNN mask estimators. Nevertheless, both alternatives take advantage of the dual-microphone information on the smartphone, including the power level differences and the spatial coherence properties.

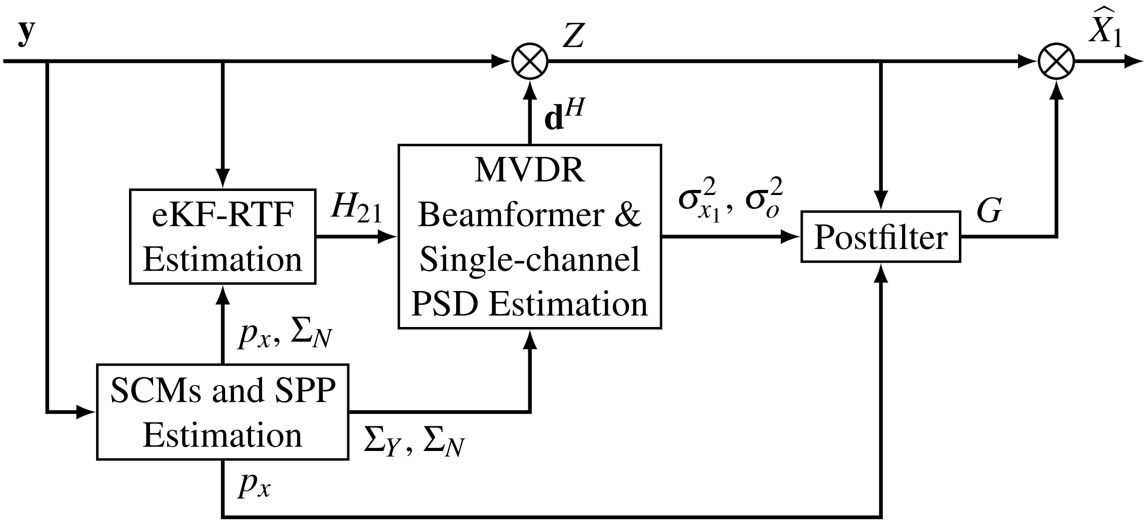


Fig. 4.1 Overview of the dual-channel enhancement algorithm for dual-microphone smartphones.

The remainder of this chapter is structured as follows. Section 4.1 describes the general dual-channel enhancement framework based on beamforming and postfiltering. The different postfilters and the proposals for the single-channel clean speech PSD estimation are also presented. The eKF framework is developed in Section 4.2, introducing the state-space model for the RTF estimation and the linearization of the model equations. The estimation of the a priori RTF statistics is also addressed. The SPP-based noise estimation is introduced in Section 4.2. Two different methods are presented either using spatial models with a priori dual-channel speech presence information or DNN mask estimator exploiting dual-channel features. Finally, in Section 4.4 the proposed approach is evaluated and the obtained results are analyzed.

4.1 Dual-channel speech enhancement framework

The proposed speech enhancement algorithm for dual-microphone smartphones is depicted in Fig. 4.1. The microphones capture the time-domain noisy speech signals $y_j(m)$, where j indicates the microphone index ($j = 1$ for the reference microphone and $j = 2$ for the secondary microphone). Then, the STFT is computed for the noisy speech signals, and they are stacked in the multichannel noisy speech vector $\mathbf{y}(t, f)$ as in (2.39). In the following, we will consider that each frequency component can be processed independently from the others, which is commonly referred to as the narrowband approximation [57].

The dual-channel noisy speech vector $\mathbf{y}(t,f)$ is first processed using an MVDR beamformer, which yields an output signal $Z(t,f)$ defined as

$$Z(t,f) = \mathbf{d}^H(t,f)\mathbf{y}(t,f), \tag{4.1}$$

where $\mathbf{d}(t,f)$ are the beamformer weights computed using Eq. (2.44). The MVDR beamformer requires knowledge of the noise spatial covariance matrix (SCM) $\Sigma_N(t,f)$ and the relative transfer function (RTF) between the secondary and reference microphones, $H_{21}(t,f)$. The noise estimation is performed using a speech presence probability (SPP)-based algorithm, which is based on the estimation of the a posteriori SPP $p_x(t,f)$ (2.32). This estimation needs for the computation of the noisy speech SCM, $\Sigma_Y(t,f)$, which is estimated using the following time-recursive averaging,

$$\widehat{\Sigma}_Y(t,f) = \tilde{\alpha}\widehat{\Sigma}_Y(t-1,f) + (1-\tilde{\alpha})\,\mathbf{y}(t,f)\mathbf{y}^H(t,f), \tag{4.2}$$

where $\tilde{\alpha}$ is an updating factor. On the other hand, an extended Kalman filter (eKF)-based estimator is used to obtain the RTF between microphones. The proposed noise and RTF estimators are described in detail in the next sections.

Finally, the speech signal at the beamformer output is enhanced by a single-channel postfilter for additional noise reduction. The estimated clean speech signal at the reference microphone is thus obtained as,

$$\widehat{X}_1(t,f) = G(t,f)Z(t,f), \tag{4.3}$$

where $G(t,f)$ is the postfiltering spectral gain. This spectral gain is obtained using the a posteriori SPP, the speech variance at the reference microphone $\sigma^2_{x_1}(t,f)$, and the residual noise variance $\sigma^2_o(t,f)$, which is obtained as in (2.45). These single-channel variances are also known as power spectral densities (PSD). The gain function yielded by the postfilter is further processed by a musical noise reduction algorithm [46].

In the next subsections, we focus on the postfiltering procedure and our proposals for the estimation of the clean speech PSD $\sigma^2_{x_1}(t,f)$ required by the former one.

4.1.1 Postfiltering approaches for dual-microphone smartphones

The performance achieved by beamforming algorithms is limited in our dual-microphone smartphone scenario due to the reduced number of microphones and its particular placement. We evaluate two different alternatives for single-channel postfiltering at the beamformer output that take advantage of the information about the SPP: the parametric Wiener filter (pWF) and the optimally-modified log spectral amplitude (OMLSA) estimator.

- **Parametric Wiener filtering** [37]: The pWF postfilter is based on the single-channel Wiener filter used in Eq. (2.49). The noise reduction performance of the Wiener filter can be improved if we consider the a posteriori SPP $p_x(t,f)$ in the postfiltering design. The idea is to further decrease the gain factor for time-frequency bins where speech is absent, which is achieved through the following gain function,

$$G_{\text{pWF}}(t,f) = \frac{\xi(t,f)}{\beta(t,f) + \xi(t,f)}, \tag{4.4}$$

where $\xi(t,f) = \sigma^2_{x_1}(t,f)/\sigma^2_o(t,f)$ is the a priori SNR of the speech signal $Z(t,f)$, and $\beta(t,f)$ is an SPP-driven trade-off parameter. To obtain this trade-off parameter, we use the same mapping function as proposed in [209] for multichannel Wiener filtering,

$$\beta(t,f) = \beta_{\min} + (\beta_{\max} - \beta_{\min}) \frac{10^{\frac{c\rho}{10}}}{10^{\frac{c\rho}{10}} + \left(\frac{p_x(t,f)}{1-p_x(t,f)}\right)^{\rho}}. \tag{4.5}$$

where c is an offset parameter and ρ controls the steepness of the transition region between the minimum and maximum values, $\beta_{\min}$ and $\beta_{\max}$ respectively. It can be observed that the lower the a posteriori SPP, the higher $\beta(t,f)$, thus improving the noise reduction performance.

- **Optimally-modified log-spectral amplitude estimator** [29]: The OMLSA estimator is based on the log-MMSE amplitude estimator in Eq. (2.19). The log spectral amplitude (LSA) estimator is only valid under the assumption of speech presence. To improve the performance, the OMLSA estimator takes into account the a posteriori SPP and different gain functions for speech presence and absence. The final gain function is then obtained as

$$G_{\text{OMLSA}}(t,f) = G_{\text{LSA}}(t,f)^{p_x(t,f)} G_{H_n}(t,f)^{1-p_x(t,f)}, \tag{4.6}$$

where G_{H_n} is a constant attenuation applied when speech is absent, which is usually set as -25 dB [29]. The LSA gain function G_{LSA} is computed using the a priori and the a posteriori SNRs of the speech signal $Z(t,f)$, the latter obtained as $\gamma(t,f) = |Z(t,f)|^2/\sigma^2_o(t,f)$.

4.1.2 Single-channel clean speech PSD estimators

The computation of the postfiltering gain functions previously presented relies on the estimation of the single-channel clean speech PSD, which is difficult to obtain because of its higher variability. We propose two different estimators that make use of the noisy speech and noise statistics and the RTF between microphones: the power level difference (PLD)-based estimator and the MVDR-based estimator.

- **PLD-based estimator**: This estimator is derived from the method in [96], which exploits the PLD between the microphones of a smartphone used in close-talk (CT) conditions; that is, a more attenuated clean speech signal is expected at the secondary microphone with respect to the reference one. This PLD is defined in terms of the noisy speech variances at the microphone inputs,

  $$\Delta\widehat{\sigma}^2_{\mathrm{PLD}} = \max\left(\widehat{\sigma}^2_{Y_1} - \widehat{\sigma}^2_{Y_2}, 0\right), \tag{4.7}$$

  where the noisy speech variances are obtained from the diagonal of $\widehat{\Sigma}_Y(t,f)$. Assuming that the power at the reference microphone is always higher than the one at the secondary microphone and that the noise PSD difference between microphones can be neglected, the clean speech PSD can be estimated as in [96],

  $$\widehat{\sigma}^2_{x_1} = \frac{\Delta\widehat{\sigma}^2_{\mathrm{PLD}}}{1 - |\widehat{H}_{21}|^2}. \tag{4.8}$$

  Although this estimator offers good performance in CT conditions, the previous assumptions are no longer valid in far-talk (FT) conditions.

- **MVDR-based estimator**: This estimator calculates, by means of spectral subtraction, the clean speech PSD directly at the beamformer output, taking into account the distortionless property of the MVDR beamformer. The estimator is defined as

  $$\widehat{\sigma}^2_{x_1} = \mathbf{d}^H\left(\widehat{\Sigma}_Y - \widehat{\Sigma}_N\right)\mathbf{d}, \tag{4.9}$$

  which can be seen as a maximum-likelihood estimator of the clean speech PSD [110]. This method fully exploits the spatial information of noisy speech and noise signals. The combination of both channels through the beamformer weights allows for a more robust estimation than directly taking the first element of the matrix subtraction.

The previous estimator can yield negative values due to the estimator variances. Thus, the resulting PSD is bounded by 0.

4.2 Relative transfer function estimation based on extended Kalman filter

As aforementioned, the proposed algorithm requires knowledge of the RTF between the two microphones, $H_{21}(t,f)$, for both the beamforming stage and the correct estimation of the clean speech PSD. In our dual-microphone scenario, the RTF can be considered as a random variable which changes across time frames, independently for each frequency bin. This variability can lie on the likely temporal variations of the acoustic channel due to environment changes, speaker head or smartphone movements, etc. Moreover, this variability helps to overcome the inaccuracy of the narrowband approximation in the multichannel distortion model [57]. The narrowband approximation assumes that the convolution between the acoustic impulse response and the clean speech signal translates to a multiplicative model in the STFT domain. Nevertheless, this is not valid when the analysis window of the STFT is shorter than the impulse response, which is especially true under reverberant environments with long acoustic responses. In that case, the acoustic channel acts as a convolutive transfer function both in time and frequency, which makes nearby frames and frequencies to be correlated. The estimation of convolutive transfer functions is a challenging task to address. Instead, by assuming that the RTF between channels is time-variant and that we can model the statistics of these variations, the narrowband approximation can be used as a possible solution.

In order to track the variability of the RTF across time frames, we propose an estimator based on Kalman filtering [102]. The Kalman filter (KF) is a linear MMSE estimator, similar to the Wiener filter. The KF considers both a prediction model for the variable to be tracked and an observation model that includes this latent variable, the observable variable, and distortion noise. This approximation has the advantage that it does not make any assumptions about the statistics of the clean speech signal, which can be inaccurate in the related scenario. Before describing the prediction and observation models for our scenario, we first define our complex variables as vectors including both real and imaginary components to avoid dealing with complex values in our Kalman filter [33, 39]. The vectors for the noisy speech, noise, and RTF are defined, respectively, as

$$\mathbf{Y}_j(t) = \begin{bmatrix} Y_j^r(t) & Y_j^i(t) \end{bmatrix}^\top, \tag{4.10}$$

$$\mathbf{N}_j(t) = \begin{bmatrix} N_j^r(t) & N_j^i(t) \end{bmatrix}^\top, \tag{4.11}$$

$$\mathbf{H}_{21}(t) = \begin{bmatrix} H_{21}^r(t) & H_{21}^i(t) \end{bmatrix}^\top, \tag{4.12}$$

where r and i superscripts refer to real and imaginary parts, respectively, and the frequency index is omitted for simplicity and brevity. Using vector-based definitions for these variables, we can define the dynamic and the observation models needed for the Kalman filter.

First, the RTF vector variable $\mathbf{H}_{21}(t)$ is assumed to be a multivariate Gaussian variable $\mathbf{H}_{21}(t) \sim \mathcal{N}\left(\mu_{H_{21}}, \Phi_{H_{21}}\right)$, where $\mu_{H_{21}}$ and $\Phi_{H_{21}}$ are the overall mean and covariance of the RTF vector, respectively. We will also assume that this vector is a random walk stochastic process that can be expressed as

$$\mathbf{H}_{21}(t) = \mathbf{H}_{21}(t-1) + \phi(t), \tag{4.13}$$

where $\phi(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ is a zero-mean multivariate white Gaussian noise that models the variability of the RTF across time frames. We choose this model as a simple approximation of the variabilities of the RTF previously exposed, which makes the definition of the Kalman filter easier.

In addition, we define an observation model for the noisy speech signal at the secondary microphone given the noisy speech at the reference microphone. First, we re-write the distortion model for the secondary microphone in the STFT domain as

$$Y_2(t) = H_{21}(t)\left(Y_1(t) - N_1(t)\right) + N_2(t), \tag{4.14}$$

which is a model independent of the clean speech signal. Then, this model is re-formulated to be expressed in terms of vectors with the real and imaginary components of the variables. This gives the final observation model,

$$\begin{aligned}\mathbf{Y}_2(t) &= \mathbf{f}(\mathbf{H}_{21}(t), \mathbf{N}_1(t); \mathbf{Y}_1(t)) + \mathbf{N}_2(t) \\ &= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\left(Y_1^r(t) - N_1^r(t)\right) + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\left(Y_1^i(t) - N_1^i(t)\right)\right)\mathbf{H}_{21}(t) + \mathbf{N}_2(t),\end{aligned} \tag{4.15}$$

where $\mathbf{f}$ is a non linear function (due to the product between the variables $\mathbf{H}_{21}(t)$ and $\mathbf{N}_1(t)$), which depends on the observation $\mathbf{Y}_1(t)$. The noises are assumed to be zero-mean multivariate Gaussian variables,

$$\begin{bmatrix} \mathbf{N}_1(t) \\ \mathbf{N}_2(t) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Phi_{N_{11}}(t) & \Phi_{N_{12}}(t) \\ \Phi_{N_{12}}^{\top}(t) & \Phi_{N_{22}}(t) \end{bmatrix}\right), \tag{4.16}$$

where $\Phi_{N_{ij}}(t) = E\left\{\mathbf{N}_i(t)\mathbf{N}_j^{\top}(t)\right\}$. Additionally, we assume that there is no correlation between the RTF vector and the noise vectors.

Once the previous models have been defined, the Kalman filter framework is applied to obtain a recursive MMSE estimate of $\mathbf{H}_{21}(t)$. The general equations for the Kalman filter model are described in [98]. The framework consists of a two-step procedure, which is applied frame-by-frame for each one of the frequency bins:

1. The **prediction step** gives a first estimation for the RTF using the model in (4.13),

$$\widehat{\mathbf{H}}_{21}(t|t-1) = \widehat{\mathbf{H}}_{21}(t-1), \tag{4.17}$$

$$\mathbf{P}(t|t-1) = \mathbf{P}(t-1) + \mathbf{Q}, \tag{4.18}$$

where

$$\mathbf{P}(t) = E\left\{\left(\mathbf{H}_{21}(t) - \widehat{\mathbf{H}}_{21}(t)\right)\left(\mathbf{H}_{21}(t) - \widehat{\mathbf{H}}_{21}(t)\right)^\top\right\}, \tag{4.19}$$

$$\mathbf{P}(t|t-1) = E\left\{\left(\mathbf{H}_{21}(t) - \widehat{\mathbf{H}}_{21}(t|t-1)\right)\left(\mathbf{H}_{21}(t) - \widehat{\mathbf{H}}_{21}(t|t-1)\right)^\top\right\} \tag{4.20}$$

are the error covariance matrices.

2. The **updating step** is then applied to correct the previous estimation using the observations $\mathbf{Y}_1(t)$ and $\mathbf{Y}_2(t)$ along with the observation model in (4.15),

$$\widehat{\mathbf{H}}_{21}(t) = \widehat{\mathbf{H}}_{21}(t|t-1) + \mathbf{K}(t)\left(\mathbf{Y}_2(t) - \mu_Y(t)\right), \tag{4.21}$$

$$\mathbf{P}(t) = \mathbf{P}(t|t-1) - \mathbf{K}(t)\Phi_Y(t)\mathbf{K}^\top(t), \tag{4.22}$$

where

$$\mathbf{K}(t) = \Phi_{HY}(t)\Phi_Y^{-1}(t) \tag{4.23}$$

is the Kalman gain, and

$$\mu_Y(t) = E\left\{\mathbf{Y}_2(t)\right\}, \tag{4.24}$$

$$\Phi_Y(t) = E\left\{\left(\mathbf{Y}_2(t) - \mu_Y(t)\right)\left(\mathbf{Y}_2(t) - \mu_Y(t)\right)^\top\right\}, \tag{4.25}$$

$$\Phi_{HY}(t) = E\left\{\left(\mathbf{H}_{21}(t) - \widehat{\mathbf{H}}_{21}(t|t-1)\right)\left(\mathbf{Y}_2(t) - \mu_Y(t)\right)^\top\right\}. \tag{4.26}$$

The previous model is defined in terms of the statistics $\mu_Y(t)$, $\Phi_Y(t)$ and $\Phi_{HY}(t)$. Because of the non-linearity of **f** in the observation model, the estimation of these statistics is non-trivial. To deal with this, we follow the extended Kalman filter (eKF) approximation [98], which is presented in the next subsection.

4.2.1 Extended Kalman filter approximation using vector Taylor series

The eKF approach is based on applying a linearization over the possible non-linear prediction and/or observation models in order to obtain tractable Kalman filter equations. This linearization is carried out by using first-order vector Taylor series (VTS), which compute the first derivatives for the non-linear equations given the latent (non-observable) variables. In our case, the prediction model is linear, so the linearization is only applied to the function **f** from the observation model. The first-order VTS approach approximates the model defined in (4.15) as

$$\begin{aligned}\mathbf{Y}_2(t) \simeq \mathbf{f}\left(\widehat{\mathbf{H}}_{21}(t|t-1), \mathbf{0}; \mathbf{Y}_1(t)\right) + \mathbf{J}_H(t)\left(\mathbf{H}_{21}(t) - \widehat{\mathbf{H}}_{21}(t|t-1)\right)\\ + \mathbf{J}_{N_1}(t)\mathbf{N}_1(t) + \mathbf{N}_2(t),\end{aligned} \tag{4.27}$$

where

$$\mathbf{J}_H(t) = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{H}_{21}(t)}\right|_{\mathbf{N}_1(t)=\mathbf{0}} = \begin{bmatrix} Y_1^r(t) & -Y_1^i(t) \\ Y_1^i(t) & Y_1^r(t) \end{bmatrix}, \tag{4.28}$$

$$\mathbf{J}_{N_1}(t) = \left.\frac{\partial \mathbf{f}}{\partial \mathbf{N}_1(t)}\right|_{\mathbf{H}_{21}(t)=\widehat{\mathbf{H}}_{21}(t|t-1)} = -\begin{bmatrix} \widehat{H}_{21}^r(t|t-1) & -\widehat{H}_{21}^i(t|t-1) \\ \widehat{H}_{21}^i(t|t-1) & \widehat{H}_{21}^r(t|t-1) \end{bmatrix} \tag{4.29}$$

are the Jacobian matrices required for the VTS approach.

Finally, using (4.27), the noisy speech statistics can be estimated as [98]

$$\mu_Y(t) \simeq \mathbf{f}\left(\widehat{\mathbf{H}}_{21}(t|t-1), \mathbf{0}; \mathbf{Y}_1(t)\right), \tag{4.30}$$

$$\begin{aligned}\Phi_Y(t) \simeq \mathbf{J}_H(t)\mathbf{P}(t|t-1)\mathbf{J}_H^\top(t) + \mathbf{J}_{N_1}(t)\Phi_{N_{11}}(t)\mathbf{J}_{N_1}^\top(t)\\ + \mathbf{J}_{N_1}(t)\Phi_{N_{12}}(t) + \Phi_{N_{12}}^\top(t)\mathbf{J}_{N_1}^\top(t) + \Phi_{N_{22}}(t),\end{aligned} \tag{4.31}$$

$$\Phi_{HY}(t) \simeq \mathbf{P}(t|t-1)\mathbf{J}_H^\top(t). \tag{4.32}$$

The advantage of the VTS approach is that it provides a good approximation to the solution in many cases, and the resulting equations are easy to implement.

4.2.2 A priori RTF statistics

The proposed eKF-based RTF estimator requires knowledge about the RTF statistics. These statistics are its overall mean vector and covariance matrix, respectively,

$$\mu_{H_{21}} = E\left\{\mathbf{H}_{21}(t)\right\}, \tag{4.33}$$

$$\Phi_{H_{21}} = E\left\{ \left(\mathbf{H}_{21}(t) - \mu_{H_{21}}\right) \left(\mathbf{H}_{21}(t) - \mu_{H_{21}}\right)^{\top} \right\}, \tag{4.34}$$

and also the covariance of the RTF variability across frames,

$$\mathbf{Q} = E\left\{ \phi(t)\phi^{\top}(t) \right\} = E\left\{ \left(\mathbf{H}_{21}(t) - \mathbf{H}_{21}(t-1)\right) \left(\mathbf{H}_{21}(t) - \mathbf{H}_{21}(t-1)\right)^{\top} \right\}. \tag{4.35}$$

The first two statistics are used during the initialization of the Kalman filter variables ($\widehat{\mathbf{H}}_{21}(0) = \mu_{H_{21}}$ and $\mathbf{P}(0) = \Phi_{H_{21}}$), while $\mathbf{Q}$ is used for all the time frames. We can also define the overall correlation matrix of the RTF vector as

$$\mathbf{R}_{H_{21}} = E\left\{ \mathbf{H}_{21}(t)\mathbf{H}_{21}^{\top}(t) \right\} = \Phi_{H_{21}} + \mu_{H_{21}}\mu_{H_{21}}^{\top}. \tag{4.36}$$

The different a priori statistics can be estimated in advance using a set of dual-channel clean speech utterances recorded in different acoustic environments and device positions (both in CT and FT conditions). To avoid outliers due to speech absent time-frequency bins, which might yield inaccurate estimates, we select, for every utterance, only those time-frequency bins where the speech power at the reference channel is large enough (i.e. higher than the maximum power at the same frequency bin in the utterance minus 3 dB). For those selected bins, we estimate the RTF by means of Eq. (2.37). This procedure yields a set of RTFs $H_{21}^{(u)}(t,f)$ for the u-th utterance, that are transformed into RTF vectors $\mathbf{H}_{21}^{(u)}(t)$. Then, the sample mean vector $\widehat{\mu}_{H_{21}}^{(u)}$ and the sample correlation matrix $\widehat{\mathbf{R}}_{H_{21}}^{(u)}$ are estimated for each utterance using those RTF vectors, while the sample covariance matrix $\widehat{\mathbf{Q}}^{(u)}$ is computed from consecutive RTF vectors in the utterance. The overall sample statistics $\widehat{\mu}_{H_{21}}$, $\widehat{\mathbf{R}}_{H_{21}}$ and $\widehat{\mathbf{Q}}$ are obtained by averaging the utterance-dependent statistics. Finally, the sample covariance matrix $\widehat{\Phi}_{H_{21}}$ can be obtained from (4.36) using the previous sample statistics.

4.3 Speech presence probability-based noise estimation

The remaining statistic needed for the proposed dual-channel speech enhancement algorithm is the noise spatial correlation matrix (SCM) $\Sigma_N(t,f)$. In a blind scenario where there is no additional knowledge about the clean speech source or the RTF, the noise SCM can be estimated across those time-frequency bins where speech is absent. We follow the multichannel SPP approach described in [194], which is based on the recursive updating of the SCM as follows,

$$\widehat{\Sigma}_N(t,f) = \alpha(t,f)\widehat{\Sigma}_N(t-1,f) + (1-\alpha(t,f))\,\mathbf{y}(t,f)\mathbf{y}^H(t,f), \tag{4.37}$$

where $\alpha(t,f)$ is an updating parameter that depends on the a posteriori SPP $p_x(t,f)$ as indicated in (2.31). The noise statistics used in the eKF-RTF estimator are directly derived from $\Sigma_N(t,f)$. Assuming a zero-mean, symmetric circular complex Gaussian distribution for $\mathbf{n}(t,f)$, and $\sigma^2_{n_{jk}}(t,f) = E\left\{N_j(t,f)N_k^*(t,f)\right\}$, the following relations can be demonstrated [39],

$$\Phi_{N_{11}}(t,f) = \frac{1}{2}\sigma^2_{n_{11}}(t,f)\mathbf{I}_2, \tag{4.38}$$

$$\Phi_{N_{22}}(t,f) = \frac{1}{2}\sigma^2_{n_{22}}(t,f)\mathbf{I}_2, \tag{4.39}$$

$$\Phi_{N_{12}}(t,f) = \frac{1}{2}\begin{bmatrix}\sigma^{2,r}_{n_{12}}(t,f) & -\sigma^{2,i}_{n_{12}}(t,f)\\ \sigma^{2,i}_{n_{12}}(t,f) & \sigma^{2,r}_{n_{12}}(t,f)\end{bmatrix}, \tag{4.40}$$

where $\mathbf{I}_2$ is the 2-dimensional identity matrix.

The a posteriori SPP, $p_x(t,f)$ (2.32), is a crucial parameter for the noise estimation procedure. It controls the recursive updating of the noise SCM and a precise estimate is needed to ensure that speech is not canceled during the beamforming stage. Moreover, $p_x(t,f)$ is also exploited in two additional steps of our framework:

- The RTF estimate provided by the eKF-RTF approach is only accurate in time-frequency bins where speech is present. Therefore, in the implementation of the eKF-RTF estimator, the RTF is only updated in those bins where $p_x(t,f) > p_{\text{thr}}$, being p_{thr} a predefined probability threshold. Otherwise, the value of the previous frame is kept.

- The proposed postfilters will take advantage of the information about the a posteriori SPP to improve the noise reduction performance.

In the next subsections, we will focus on different procedures to estimate the a posteriori SPP. First, we will analyze its estimation by using spatial statistical models. Thus, we will define possible estimators for the a priori speech absence probability (SAP) $q_n(t,f)$ (2.33) that exploit the dual-channel information. Finally, we will propose a DNN-based mask estimator that directly estimates the a posteriori SPP by using spectral and spatial features.

4.3.1 A posteriori SPP estimation based on statistical spatial models

The estimation of the a posteriori SPP can be addressed assuming that the noisy speech signal $\mathbf{y}(t,f)$ follows complex multivariate Gaussian distributions [194, 209], which depend on the presence or absence of the clean speech signal. Using the Bayes' rule, the a posteriori SPP at each time-frequency bin can be re-written as

$$p_x(t,f) = \frac{(1-q_n(t,f))\, f\left(\mathbf{y}(t,f)|H_x\right)}{(1-q_n(t,f))\, f\left(\mathbf{y}(t,f)|H_x\right) + q_n f\left(\mathbf{y}(t,f)|H_n\right)}, \tag{4.41}$$

where

$$f\left(\mathbf{y}(t,f)|H_x\right) = \frac{e^{-\mathbf{y}(t,f)\Sigma_Y^{-1}(t,f)\mathbf{y}^H(t,f)}}{\pi^J \det\left[\Sigma_Y(t,f)\right]}, \tag{4.42}$$

$$f\left(\mathbf{y}(t,f)|H_n\right) = \frac{e^{-\mathbf{y}(t,f)\Sigma_N^{-1}(t,f)\mathbf{y}^H(t,f)}}{\pi^J \det\left[\Sigma_N(t,f)\right]}, \tag{4.43}$$

are the likelihoods of observing the noisy speech signal under the H_x and H_n hypotheses, respectively, with $\det[\cdot]$ being the matrix determinant operator and J the number of microphones. Using these likelihoods, the expression in (4.41) can be redefined as

$$p_x(t,f) = \left(1 + \frac{q_n(t,f)}{1-q_n(t,f)} \frac{\det\left[\Sigma_Y(t,f)\right]}{\det\left[\Sigma_N(t,f)\right]} \frac{e^{-\mathbf{y}(t,f)\Sigma_N^{-1}(t,f)\mathbf{y}^H(t,f)}}{e^{-\mathbf{y}(t,f)\Sigma_Y^{-1}(t,f)\mathbf{y}^H(t,f)}}\right)^{-1}. \tag{4.44}$$

Finally, the following two-iteration algorithm can be used to estimate the a posteriori SPP at each frame for all frequencies:

- *Initialization*: Estimate the noisy SCM using (4.2) and the a priori SAP $q_n(t,f)$ (see the following subsection).
- *1st iteration*: Estimate $p_x(t,f)$ using $\widehat{\Sigma}_N(t-1,f)$ in (4.44). Then, estimate $\widehat{\Sigma}_N(t,f)$ using $p_x(t,f)$ in (4.37).
- *2nd iteration*: Re-estimate $p_x(t,f)$ using now $\widehat{\Sigma}_N(t,f)$ in (4.44). Finally, re-estimate $\widehat{\Sigma}_N(t,f)$ using $p_x(t,f)$ in (4.37).

This procedure allow to use the current observation during the second iteration, thus improving the estimation.

4.3.2 A priori SAP estimation for dual-microphone smartphones

The a priori SAP is a key parameter for the estimation of the a posteriori SPP described in the previous subsection. An accurate estimate of the a priori SAP allows for the robust tracking of the noise statistics. The a priori SAP can be predicted in terms of the a priori SNR [60, 27, 194]. The main problem of these approaches is their lack of robustness in the case of a time-varying SNR, which can make noise changes to be detected as speech presence. Other approaches use spatial information to distinguish between coherent sources, as clean speech signal, and diffuse sources, as environmental noise. This is the case of [209] which make use of the coherent-to-diffuse ratio (CDR). Alternatively, the power level difference (PLD) between the noisy speech signals in a dual-microphone smartphone is exploited in [96] for the estimation of the noise statistics in close-talk (CT) conditions.

In this subsection, we propose an SAP estimator for dual-microphone smartphones that combines spatial and PLD information to improve the accuracy in the estimation. Our proposal first computes an estimate of the noisy speech SCM $\widetilde{\Sigma}_Y(t,f)$ using a rectangular window of eight past frames (typically 128 ms), as in [209]. Then, it calculates: 1) The PSD ratio between microphones as

$$\widehat{\eta}_{y_{21}}(t,f) = \frac{\widetilde{\sigma}^2_{y_{22}}(t,f)}{\widetilde{\sigma}^2_{y_{11}}(t,f)}, \tag{4.45}$$

where $\widetilde{\sigma}^2_{y_{ij}}$ is an estimate of $E\left\{Y_iY_j^*\right\}$ obtained from $\widetilde{\Sigma}_Y(t,f)$, and 2) The short-term complex coherence between microphones,

$$\widehat{\Gamma}_{Y_{12}}(t,f) = \frac{\widetilde{\sigma}^2_{y_{12}}(t,f)}{\sqrt{\widetilde{\sigma}^2_{y_{11}}(t,f)\widetilde{\sigma}^2_{y_{22}}(t,f)}}. \tag{4.46}$$

The previous PLD ratio and coherence terms are used for the SAP estimation, which is composed of two methods: an PLD-based and an CDR-based estimator.

The PLD-based a priori SAP estimation is based on the PLD algorithm described in [96], where a normalized difference of the noisy speech PSDs was defined as

$$\Delta\widehat{\sigma}^2_{\mathrm{nPLD}}(t,f) = \frac{\widetilde{\sigma}^2_{y_{11}}(t,f) - \widetilde{\sigma}^2_{y_{22}}(t,f)}{\widetilde{\sigma}^2_{y_{11}}(t,f) + \widetilde{\sigma}^2_{y_{22}}(t,f)} = \frac{1-\widehat{\eta}_{y_{21}}(t,f)}{1+\widehat{\eta}_{y_{21}}(t,f)}. \tag{4.47}$$

The noise statistics are updated by using this parameter, as it provides information about the speech presence in each time-frequency bin. Assuming that speech is more attenuated at the secondary microphone (with respect to the reference one), while similar noise PSDs are present in both channels, $\Delta\widehat{\phi}_{\mathrm{nPLD}}(t,f)$ is close to one when speech is present and tends to zero otherwise. Thus, a PLD-based a priori SAP estimator can be obtained as

$$q_{\mathrm{PLD}}(t,f) = 1 - \Delta\widehat{\sigma}^2_{\mathrm{nPLD}}(t,f) = \frac{2\widehat{\eta}_{y_{21}}(t,f)}{1+\widehat{\eta}_{y_{21}}(t,f)}, \tag{4.48}$$

where $q_{\text{PLD}}(t,f)$ is upper-bounded by 1. Although this estimator is intended for close-talk (CT) conditions, it can also be useful in far-talk (FT), especially at those frequencies where speech at the secondary microphone is more attenuated.

On the other hand, the CDR is another indicator of speech presence [187] in multichannel conditions. The CDR between two microphones is defined as

$$\Psi_{Y_{12}}(t,f) = \frac{\Gamma_{\text{diff}_{12}}(f) - \Gamma_{Y_{12}}(t,f)}{\Gamma_{Y_{12}}(t,f) - \Gamma_{X_{12}}(t,f)}, \tag{4.49}$$

where $\Gamma_{X_{12}}(t,f)$ is the clean speech short-term complex coherence (defined as in (4.46)) and $\Gamma_{\text{diff}_{12}}(f)$ is the diffuse noise field complex coherence defined in (2.52). Higher values of the CDR indicate the presence of a strong coherent component, often a clean speech signal. On the contrary, lower values are common when a diffuse component is dominant, as in the case of noise signals. In practice, the CDR is computed using the estimator proposed in [187],

$$\widehat{\Psi}_{Y_{12}}(t,f) = \Re\left(\frac{\Gamma_{\text{diff}_{12}}(f) - \widehat{\Gamma}_{Y_{12}}(t,f)}{\widehat{\Gamma}_{Y_{12}}(t,f) - e^{j\angle\widetilde{\sigma}^2_{y_{12}}(t,f)}}\right), \tag{4.50}$$

where $\angle\widetilde{\sigma}^2_{y_{12}}(t,f)$ is the phase of $\widetilde{\sigma}^2_{y_{12}}(t,f)$.The real-part $\Re(\cdot)$ is taken in (4.50) to prevent complex values due to estimation errors, as the CDR must be positive and real-valued. In addition, a frequency-averaged CDR is computed using a normalized Hamming window w_N as

$$\overline{\Psi}_{Y_{12}}(t,f) = \sum_{i=-n}^{n} w_N(i)\widehat{\Psi}_{Y_{12}}(t,f-i), \tag{4.51}$$

with $n = 10$ (the window length is $2n+1$). Then, the local a priori SAP estimate is computed as in [209],

$$q_{\text{local}}(t,f) = q_{min} + (q_{max} - q_{min})\frac{10^{\frac{c\rho}{10}}}{10^{\frac{c\rho}{10}} + \widehat{\Psi}^{\rho}_{Y_{12}}(t,f)}, \tag{4.52}$$

which is a step function similar to that in (4.5). In a similar way, the global a priori SAP estimate $q_{\text{global}}(t,f)$ is computed using $\overline{\Psi}_{Y_{12}}(t,f)$ instead of $\widehat{\Psi}_{Y_{12}}(t,f)$ in (4.52). The CDR-based a priori SAP estimate is finally obtained as [209],

$$q_{\text{CDR}}(t,f) = 1 - (1 - q_{\text{local}}(t,f))(1 - q_{\text{global}}(t,f)). \tag{4.53}$$

This estimator is similar to that proposed in [209], but neglecting the frame a priori SAP term, as it did not lead to a performance improvement in preliminary experiments.

The a priori SAP estimates obtained by the PLD and CDR approaches can be combined to yield a more robust joint decision. Assuming statistical independence between both estimators, the combined a priori SAP estimate can be obtained as the joint probability of speech absence at both estimators,

$$q_n(t,f) = q_{\mathrm{CDR}}(t,f)q_{\mathrm{PLD}}(t,f). \tag{4.54}$$

The above estimator can be used in both CT and FT conditions, and it is expected to be more robust as the speech absence decision, $q_n(t,f) = 1$, is only obtained when both estimators agree on speech absence.

4.3.3 A posteriori SPP estimation based on DNN mask estimator

In the previous subsections, we have described the estimation of the a posteriori SPP using classical statistical signal processing. As seen, this approach relies on the properties of the spatial statistics for the clean speech and noise signals, which need to be estimated in advance, often by using iterative procedures. Moreover, the performance highly depends on an accurate estimation of the a priori SAP, which is based on assumptions about the spectral and spatial properties of the signals that could be unrealistic. On the other hand, deep neural networks have shown a great performance when estimating speech and noise presence masks [224]. The advantage of these deep learning models is that they do not rely on prior assumptions, as they can learn them from the data. Thus, in this subsection we explore the use of DNNs to directly estimate the a posteriori SPP, integrating this model into our proposed dual-channel speech enhancement framework. This yields an integrated algorithm that exploits the power of deep learning architectures.

In particular, we consider a convolutional recurrent network (CRN) architecture [205, 208] for the estimation of the a posteriori SPP $p_x(t,f)$. This model successfully combines the use of CNN layers to exploit the local correlations between nearby frequencies with the use of RNN ones to model the temporal characteristics of the speech signal. In addition, this architecture allows for the use of several input features without the need of input layers with a high number of parameters, as in the case of the fully-connected architectures. A diagram of the used CRN architecture is depicted in Fig. 4.2. As can be observed, the model comprises an encoder with five convolutional layers, a decoder with five deconvolutional layers, and an intermediate LSTM recurrent neural networks. We use exponential linear units (ELUs) as non-linear function in all convolutional and deconvolutional layers except for the output layer, which uses the sigmoid function. A dropout layer is placed before the input of the LSTM layer. The convolutional and deconvolutional layers only operate in the frequency dimension, while the LSTM layer exploits the temporal dimension. We choose LSTM instead of bidirectional RNNs to keep the causality of the network, which allows

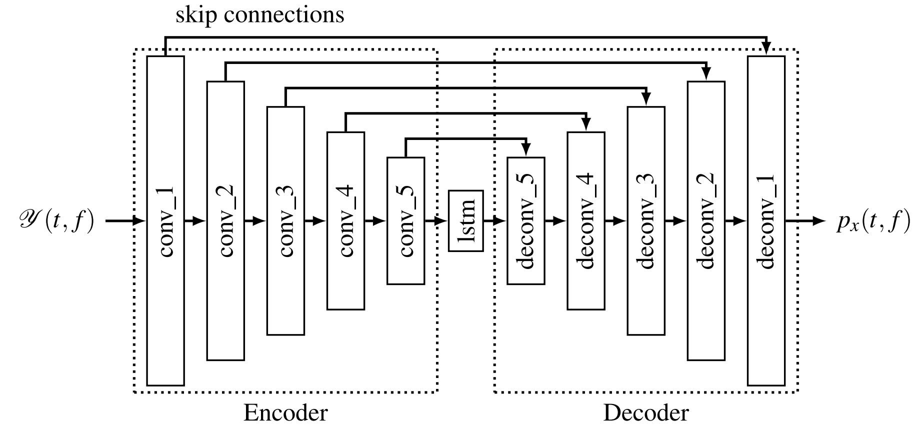


Fig. 4.2 Diagram for the CRN architecture used for the estimation of speech presence probability masks.

for online implementation. Furthermore, we use skip connections which concatenate the output of each encoder layer to the input of each decoder layer. These skip connections can help in the backpropagation procedure during the DNN training. The CRN is trained using ideal binary masks (IBM) from the reference channel as target features using the binary cross-entropy (BCE) as loss function.

The network's input features $\mathscr{Y}(t,f)$ are a set of different feature maps exploiting spectral or spatial properties. The main set of features is the log-magnitude spectrum of the primary channel,

$$\mathscr{Y}_{\mathrm{LMS}}(t,f) = \log|Y_1(t,f)|. \tag{4.55}$$

An online normalization is applied to them using a time-recursive mean computation and subtraction at each frequency bin, which has shown to yield good performance for online processing [72]. These LMS features only exploit the spectral properties of the primary channel, but not the spectral and spatial relationship between the microphones. Therefore, we included additional features that make use of the inter-channel properties of the signals. These features are inspired by those we used for the estimation of the a priori SAP. First, we consider the spectral relation between the channels by using instantaneous PLD features, which are defined as

$$\mathscr{Y}_{\mathrm{PLD}}(t,f) = \frac{|Y_1(t,f)|^2 - |Y_2(t,f)|^2}{|Y_1(t,f)|^2 + |Y_2(t,f)|^2}. \tag{4.56}$$

These PLD features can be useful when the speech component on the primary channel is stronger than on the secondary one, as in close-talk (CT) condition. Also, the spatial properties of the signals can be exploited using the interchannel phase difference (IPD)

Table 4.1 Results provided by the different objective metrics for the noisy speech signals of the test set in the TIMIT-2C-CT/FT database. Results are broken down by SNR and device user mode (CT or FT).

Metric	**Condition**	**SNR (dB)**					
		-5	**0**	**5**	**10**	**15**	**20**
PESQ	CT	1.09	1.11	1.23	1.45	1.81	2.27
	FT	1.07	1.11	1.25	1.50	1.88	2.38
STOI	CT	0.51	0.63	0.74	0.84	0.91	0.95
	FT	0.50	0.61	0.73	0.83	0.90	0.95
SDR	CT	-5.80	-0.81	4.19	9.15	14.02	18.70
	FT	-5.79	-0.80	4.19	9.15	14.03	18.70

features [233],

$$\mathscr{Y}_{IPD}(t,f) = \Big[\cos\left(\theta_{y_1}(t,f) - \theta_{y_2}(t,f)\right) \quad \sin\left(\theta_{y_1}(t,f) - \theta_{y_2}(t,f)\right)\Big]^{\top}, \tag{4.57}$$

where $\theta_{y_1}(t,f)$ and $\theta_{y_2}(t,f)$ are the phase of the noisy speech signals at the reference and secondary microphones, respectively. These IPD features have a similar function that the CDR-based a priori SAP estimator proposed in the previous subsection, as both make use of the phase difference between channels (using the instantaneous noisy observations for the CRN features and the noisy PSDs for the CDR-based estimator).

4.4 Experimental results

We evaluated the proposed dual-channel algorithm in the TIMIT-2C-CT/FT database. As described in Chapter 3, this database includes simulated dual-channel noisy speech recordings from a dual-microphone smartphone used in both close-talk (CT) and far-talk (FT) conditions. The performance of the different estimators and speech enhancement algorithms discussed along was evaluated in the test set of the database. We used the following objective speech quality and intelligibility metrics: PESQ, STOI, and scale-invariant SDR. Table 4.1 shows the results obtained for these metrics when evaluating the noisy speech signals of the test set in CT and FT conditions. These results will serve as a reference. Moreover, the speech distortion (SD) index is also considered to evaluate the RTF estimation accuracy.

For the STFT computation, a 512-point DFT was applied using a 32 ms square-root Hann window with 50% overlap. This resulted in a total of 257 frequency bins for each time frame. The values of the hyperparameters used for the proposed dual-channel algorithm are given in Table 4.2. These values are based on those recommended in [209]. The estimation of the

Table 4.2 Hyperparameters used in the proposed dual-channel speech enhancement algorithm based on RTF estimation.

Param.	**Value**	**Param.**	**Value**
$\tilde{\alpha}$	0.9	p_{thr}	0.9
β_{min}	1	q_{min}	0.1
β_{max}	4	q_{max}	0.998
c (pWF)	-3	c (SAP)	3
ρ (pWF)	4	ρ (SAP)	2.5

Table 4.3 Architecture of the CRN applied to SPP mask estimation. The feature size is indicated in the form *feature maps* × *frames* × *freq. channels*, with N_{in} being the number of feature maps at the input. The hyperparameters column refers to *kernel size*, *stride* and *output channels*. For the LSTM layer, the number of hidden units is also indicated.

Layer Name	**Input size**	**Hyperparameters**	**Output size**
conv_1	N_{in} × T × 257	1 × 3, (1, 2), 8	8 × T × 128
conv_2	8 × T × 128	1 × 3, (1, 2), 8	8 × T × 64
conv_3	8 × T × 64	1 × 3, (1, 2), 16	16 × T × 32
conv_4	16 × T × 32	1 × 3, (1, 2), 32	32 × T × 16
conv_5	32 × T × 16	1 × 3, (1, 2), 64	64 × T × 8
reshape_1	64 × T × 8	-	T × 512
lstm	T × 512	512	T × 512
reshape_2	T × 512	-	64 × T × 8
deconv_5	128 × T × 8	1 × 3, (1, 2), 32	32 × T × 16
deconv_4	64 × T × 16	1 × 3, (1, 2), 16	16 × T × 32
deconv_3	32 × T × 32	1 × 3, (1, 2), 8	8 × T × 64
deconv_2	16 × T × 64	1 × 3, (1, 2), 8	8 × T × 128
deconv_1	16 × T × 128	1 × 3, (1, 2), 1	1 × T × 257

a priori RTF statistics for the eKF-RTF estimator was performed using reverberated clean speech utterances from the training set. Finally, the CRN architecture was trained using either the training set of the CT or the FT condition database. A condensed description of the CRN network architecture used in our experiments is provided in Table 4.3. We used a batch size of 10 utterances, which were zero-padded to have the same number of frames. The dropout rate was set to 0.5 deactivation probability.

In the next subsections, we show the experimental results obtained from the different evaluations. First, we compared the performance of the different a priori SAP estimators, the RTF estimators, and the single-channel clean speech PSD estimators. Then, we evaluated the different proposed postfilters by using the best estimators found in the previous evaluations. Finally, the performance of the CRN SPP estimator in the dual-channel speech enhancement framework (beamforming plus postfiltering) was analyzed considering different combinations of input features. The results were then compared with those obtained using statistical spatial models for the a posteriori SPP.

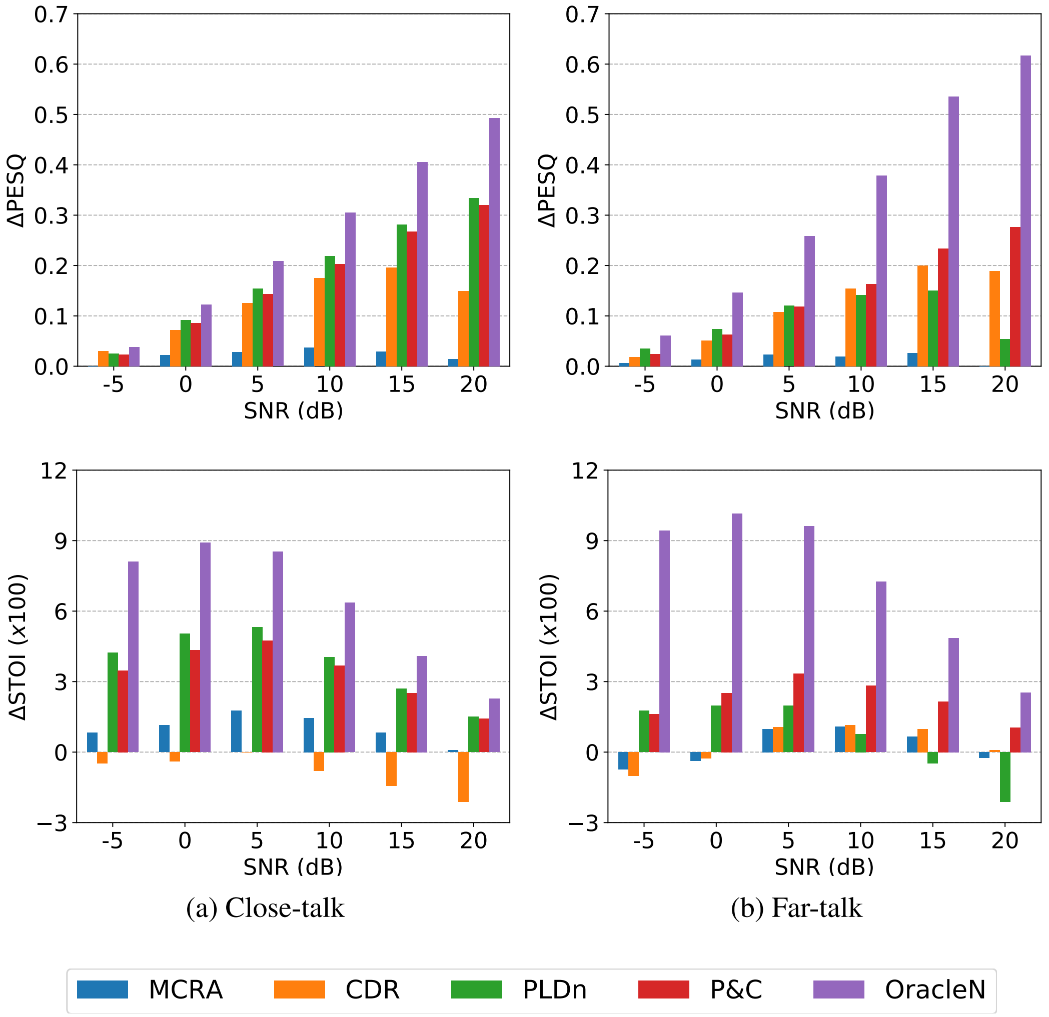
4.4.1 Performance of the a priori SAP estimators

We first compared the different a priori SAP estimators when used for noise statistic estimation in an MVDR beamforming, without postfiltering, along with the eKF-RTF estimator. The obtained PESQ and STOI results are shown in Fig. 4.3. The following SAP estimators were evaluated: the SNR-based approach of the MCRA method (MCRA) [194], the proposed CDR-based SAP estimator (CDR), the proposed PLD-based SAP estimator (PLDn), and the combination of PLDn and CDR estimators (P&C). As a performance upper-bound, we show the results achieved with an oracle estimation of the noise SCM (OracleN). This estimation was obtained using the actual noise that contaminate the noisy speech signals in a recursive procedure similar to that in (4.2). In this case, the a posteriori SPP for the eKF-RTF estimator was directly obtained from the clean speech signal at the reference microphone.

In CT conditions, the best results were obtained for the PLDn approach. In this case, the speech power difference between microphones makes that the PLDn estimator can easily detect speech presence bins. On the other hand, the power difference reduces the CDR ratio when a stronger speech component is present. This leads the CDR approach to underestimate the speech presence prediction, thus decreasing noise tracking performance. Therefore, in comparison with the PLDn method, the P&C combination did not yield improvements.

In FT conditions, on the other hand, speech power is more similar between channels, so the CDR ratio increases under speech presence. This is especially true at higher SNRs, where the CDR approach outperforms the PLDn one. However, the performance of the detector based on the CDR ratio severely degrades at lower SNRs, being the PLD-based detector more robust at them. The P&C combination increases the performance in terms of both speech quality and intelligibility. This approach keeps a noise reduction performance similar to the PLDn method at lower SNRs, while the combined decision outperforms the other single decision approaches at higher SNRs.

To sum up, our proposed SAP estimators improve the noise statistics tracking when used in dual-microphone smartphones. The PLDn proposal shows the best results in CT conditions, while the P&C combination achieves a similar performance. In FT conditions, the P&C joint decision achieves the best results at higher SNRs, while the PLDn method performs slightly better at lower SNRs. Therefore, in the next experiments involving the use of a priori SAP estimation, we selected the PLDn method for CT conditions and the P&C approach for FT conditions.



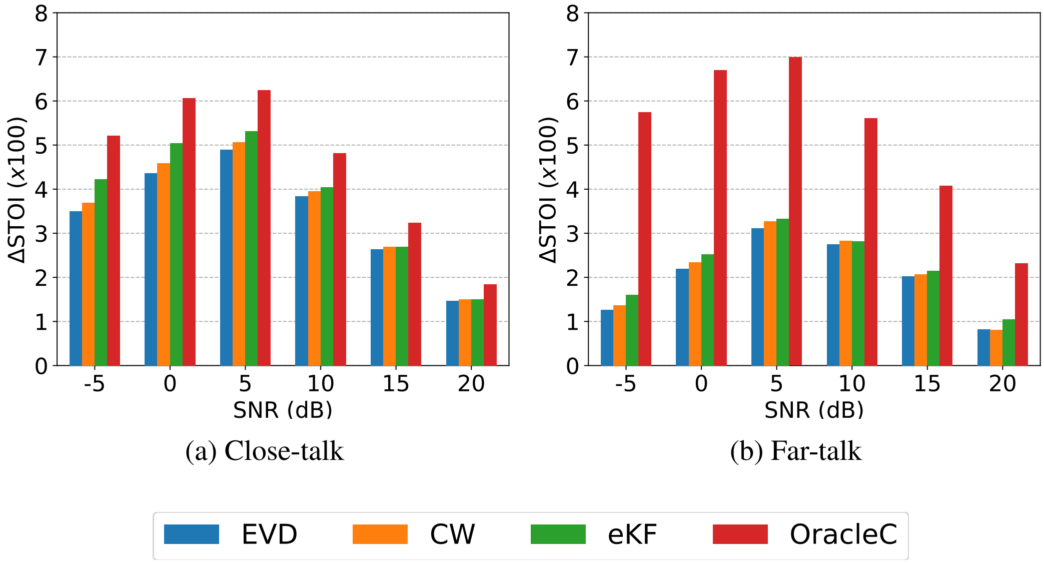
(a) Close-talk

(b) Far-talk

Fig. 4.3 PESQ and STOI results for the different SAP estimators when combined with SPP-based eKF-RTF estimation for MVDR beamforming. The plots only show increments with respect to the results over noisy speech.

4.4.2 Performance of the RTF estimators

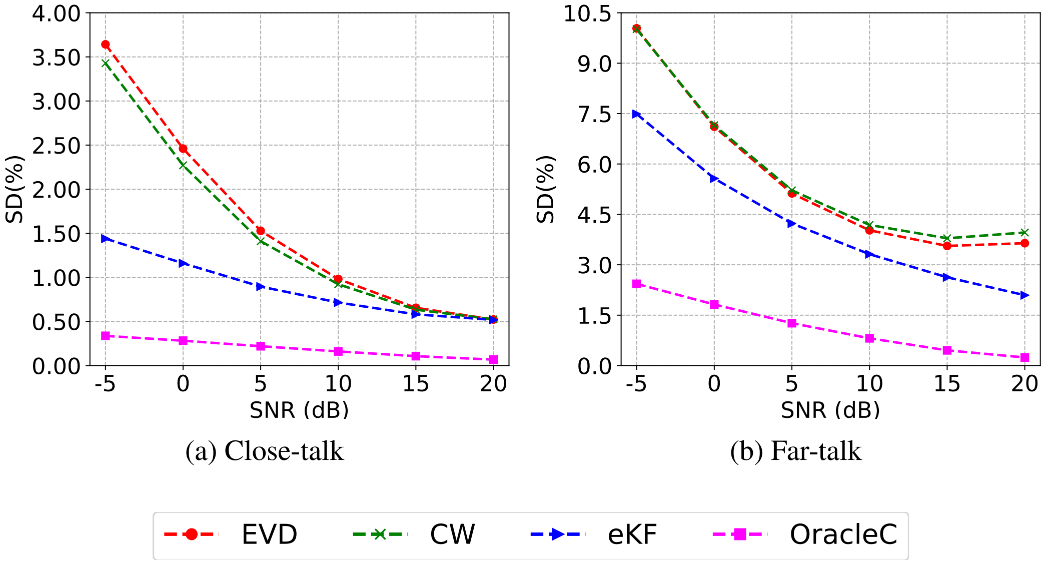
We compared the performance of the proposed eKF-RTF estimator with the EVD and CW sub-space methods for RTF estimation. The STOI results are shown in Fig. 4.4, while Fig. 4.5 shows the results obtained using the SD index. As in this case we were more interested in the distortion introduced by the MVDR beamformer over the clean speech signal due to RTF estimation errors, we focus on speech distortion and intelligibility metrics. In addition, we include upper-bound results obtained with an oracle estimator for the RTF (OracleC). This oracle estimation was obtained by dividing the clean speech STFT components of the secondary channel by those of the primary channel. This procedure was done in that time-frequency bins where speech presence was detected, while we reused the RTF of



(a) Close-talk

(b) Far-talk

Fig. 4.4 STOI results for the different RTF estimators when used for MVDR beamforming. The plots only show increments with respect to the results over noisy speech.



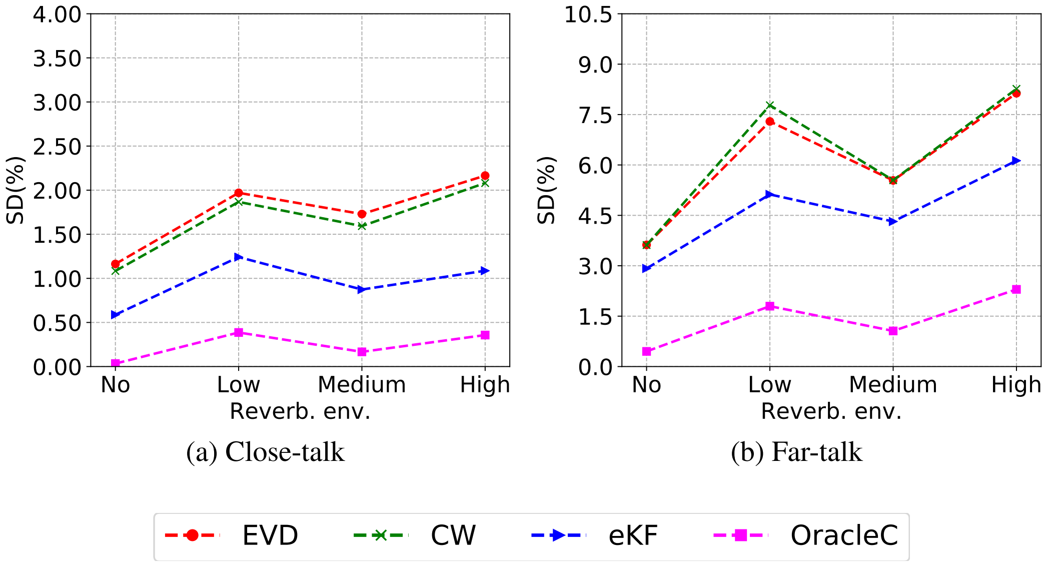
(a) Close-talk

(b) Far-talk

Fig. 4.5 SD results (in terms of the SNR) for the different RTF estimators when used for MVDR beamforming.

the previous frame otherwise. The comparison was performed considering the best SAP estimator obtained for each condition (PLDn for CT and P&C for FT). For a fair comparison, we used the same RTF initialization and the same SPP-based updating scheme for all the methods.

It can be observed that our eKF estimator achieves slightly better results in terms of STOI and much lower speech distortion than the other estimators, especially in FT conditions.



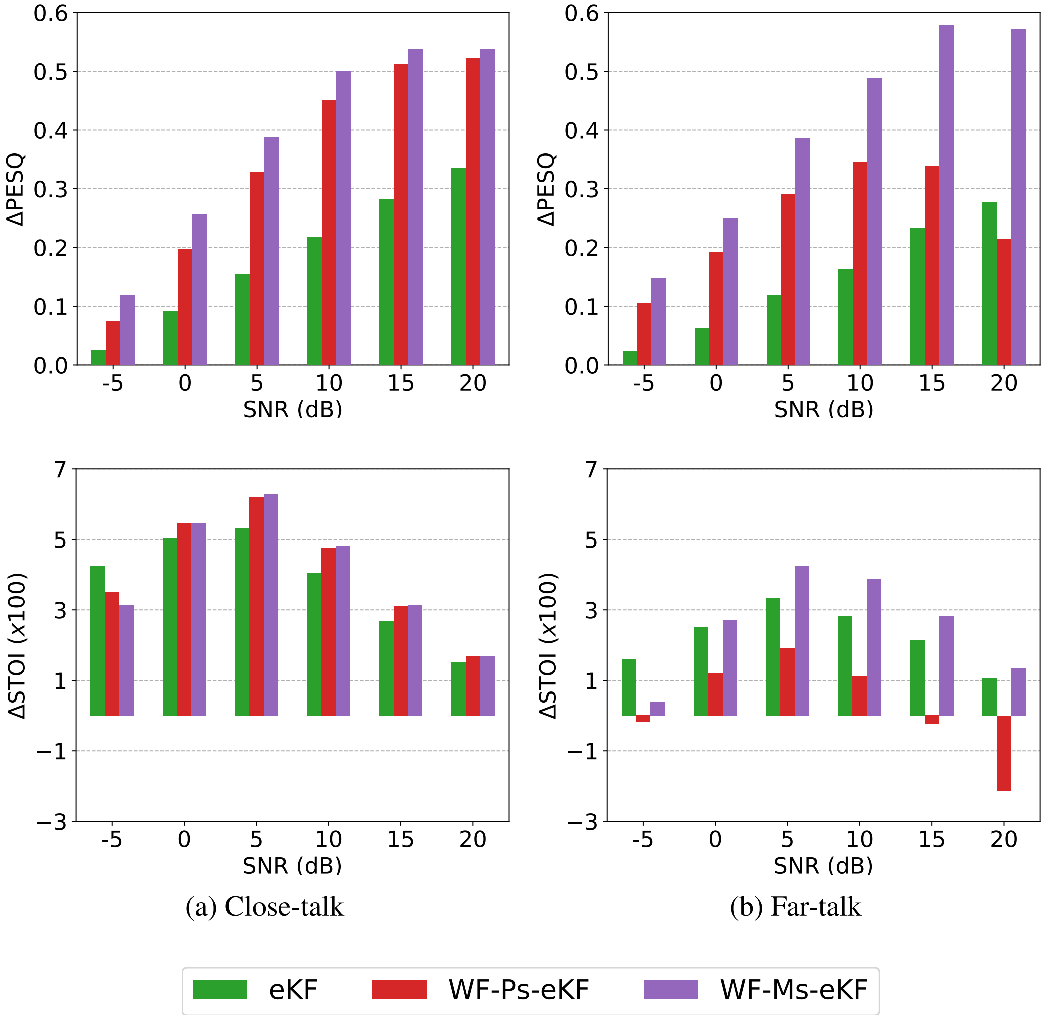
(a) Close-talk

(b) Far-talk

Fig. 4.6 SD results (in terms of the reverberation environment) for the different RTF estimators when used for MVDR beamforming.

It should be noticed that the distortionless property of the MVDR beamformer involves very low speech distortion when an accurate RTF estimate is available. Therefore, we can conclude that our eKF approach can improve the track of the RTF variability across frames in comparison with sub-space methods. This is especially noticeable in FT conditions, where the reverberation level is higher due to the distance between the speaker and both microphones. Thus, the secondary microphone captures similar speech power than the reference one, which increases the RTF variability and makes its tracking more challenging.

Additionally, we show in Fig. 4.6 the SD results for the different approaches broken down by reverberant environment (in this case, the results from common noisy environments and SNR levels are averaged). The oracle results are included for a better comparison, as some results average more challenging noisy environments (e.g., the cafe noise in the low reverberation scenario). As observed, the RTF estimation is generally more difficult in scenarios with higher reverberation levels, where the performance of the different approaches degrades. This is due to the increasing RTF variability, which makes this variable harder to track. Nevertheless, it can be observed that our proposed estimator is more robust against reverberant environments than the tested sub-space approaches, achieving a lower speech distortion at different reverberation levels.



(a) Close-talk

(b) Far-talk

Fig. 4.7 PESQ and STOI results for the different clean speech PSD estimators when combined with a Wiener postfiltering applied at the MVDR beamformer output. The plots only show increments with respect to the results over noisy speech.

4.4.3 Performance of the single-channel clean speech PSD estimators

In this subsection we show the results obtained for both the proposed PLD-based (Ps) and MVDR-based (Ms) clean speech PSD estimators when combined with a Wiener postfiltering (WF) applied at the MVDR output. Fig. 4.7 shows the obtained PESQ and STOI results. In addition, we include the results obtained using a standalone MVDR beamformer without postfiltering (eKF in Fig. 4.7) as a reference. The best SAP configuration previously determined for each condition was used in the different methods.

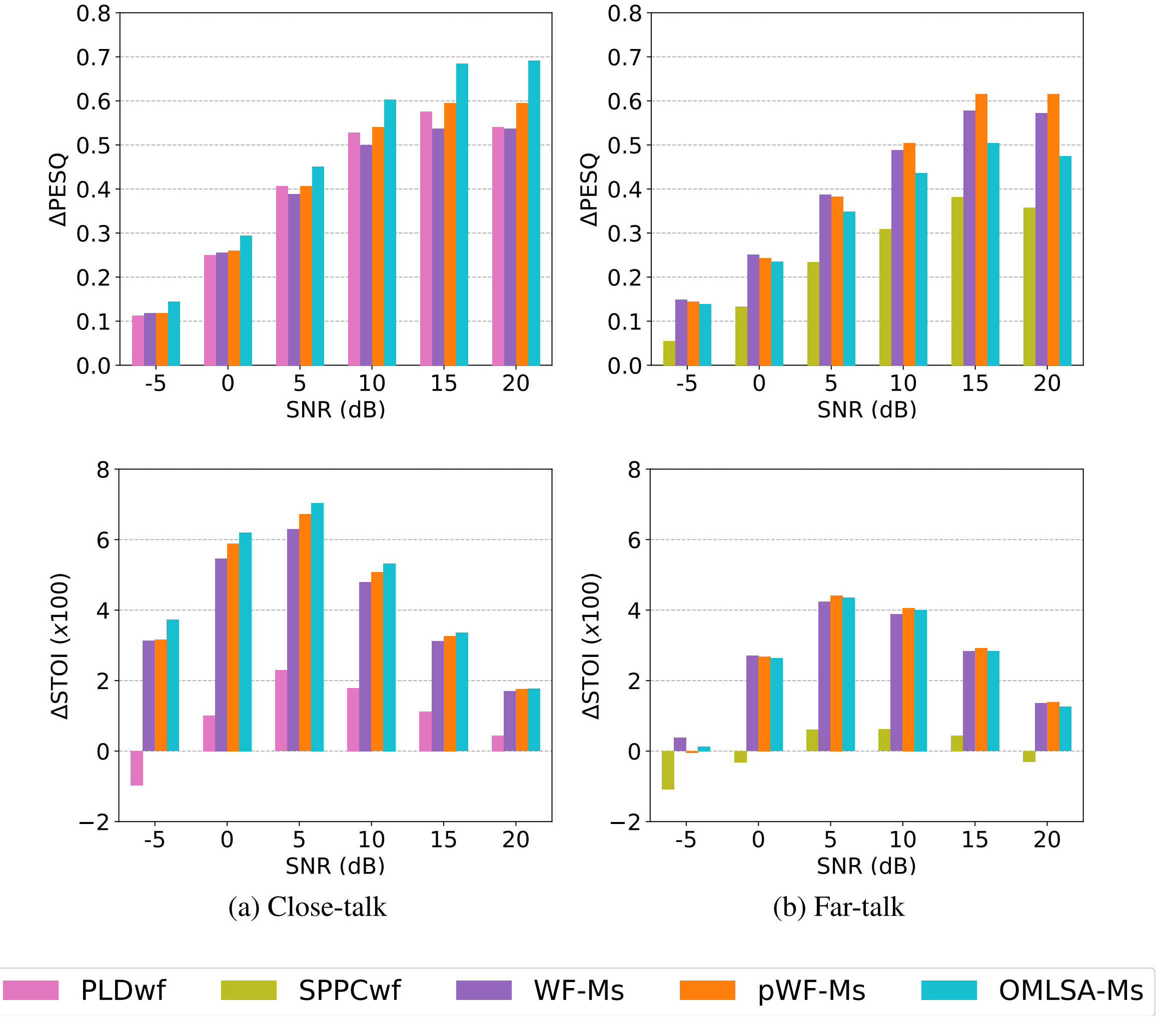
The results show that the MVDR-based PSD estimation performs better than the PLD-based one. The WF-Ms method obtained slightly better results than the WF-Ps approach in CT conditions, while it clearly outperformed the other approaches in FT conditions. The advantage of the MVDR-based estimator is that it does not make any assumptions about similar noise power between microphones, as the PLD-based one. In addition, the MVDR-based estimation also exploits the cross-correlation terms of the noisy speech and noise SCMs. This makes the MVDR-based estimation more robust than its PLD-based counterpart. Moreover, the PLD assumption is no longer valid in FT conditions, leading to a degradation on the performance of the WF-Ps approach, particularly at higher SNRs.

4.4.4 Performance of the postfiltering approaches

Fig. 4.8 compares the results yielded by the pWF and OMLSA proposed postfilters in terms of PESQ and STOI metrics. We also show the results obtained by the WF postfilter and two different related dual-channel enhancement algorithms intended for smartphones: the PLD-based Wiener filter (PLDwf) proposed in [96] for CT conditions, and the SPP- and coherence-based Wiener filter (SPPCwf) proposed in [158] for FT conditions. The MVDR-based PSD estimator was used in the different approaches (WF, pWF, and OMLSA).

In CT conditions, both the pWF and OMLSA postfilters outperform the WF and PLDwf approaches, with the OMLSA postfilter achieving the best results at the different metrics. The PLDwf approach achieves more noise reduction than the WF as it uses an overestimation of the noise, which yields better PESQ scores. The problem is that this overestimation introduces speech distortion, which reduces speech intelligibility. On the other hand, the use of the a posteriori SPP in the pWF and OMLSA postfilters allows for larger noise reduction when speech is absent. Besides, it does not introduce additional speech distortion. The availability of accurate SPPs in the CT scenario allows the OMLSA approach to obtain better results than the other methods. To explain this, it must be taken into account that the LSA estimator performs better than the WF when there is a discriminative speech presence information. The use of the SPP is advantageous for the OMLSA approach.

In FT conditions, on the other hand, the obtained SPPs are less accurate due to the more challenging scenario, thereby degrading the performance of the OMLSA postfilter. Nevertheless, the SPP estimate is still useful for the pWF postfilter, which outperforms, in general, the WF postfilter (especially at higher SNRs). Moreover, both WF and pWF postfilters outperform the SPPCwf approach in all the evaluated metrics. To sum up, the availability of accurate RTF and SPP estimates, as those provided by our proposed framework, clearly improves the performance of the postfiltering approaches in challenging FT conditions.



(a) Close-talk

(b) Far-talk

Fig. 4.8 PESQ and STOI results for the different postfilters applied at the output of the MVDR beamformer as well as other related state-of-the-art approaches. The plots only show increments with respect to the results over noisy speech.

4.4.5 Evaluation of the DNN-based SPP mask estimator

Finally, the CRN-based SPP mask estimator was evaluated along with our proposed dual-channel postfiltering approach with eKF-RTF estimation. The results are shown in Fig. 4.9, where the different methods are compared in terms of PESQ, STOI, and SI-SDR. The OMLSA postfilter with the MVDR-based PSD estimator (OMLSA-Ms) were used for the different approaches in both CT and FT conditions. Different combinations of input feature maps were evaluated, all of them using LMS features: the plain CRN (CRN), the approaches that also integrate either PLD features (PLD) or IPD features (IPD), and the full integration of both features (PLD+IPD). The CRN-based a posteriori SPP estimator was also compared with

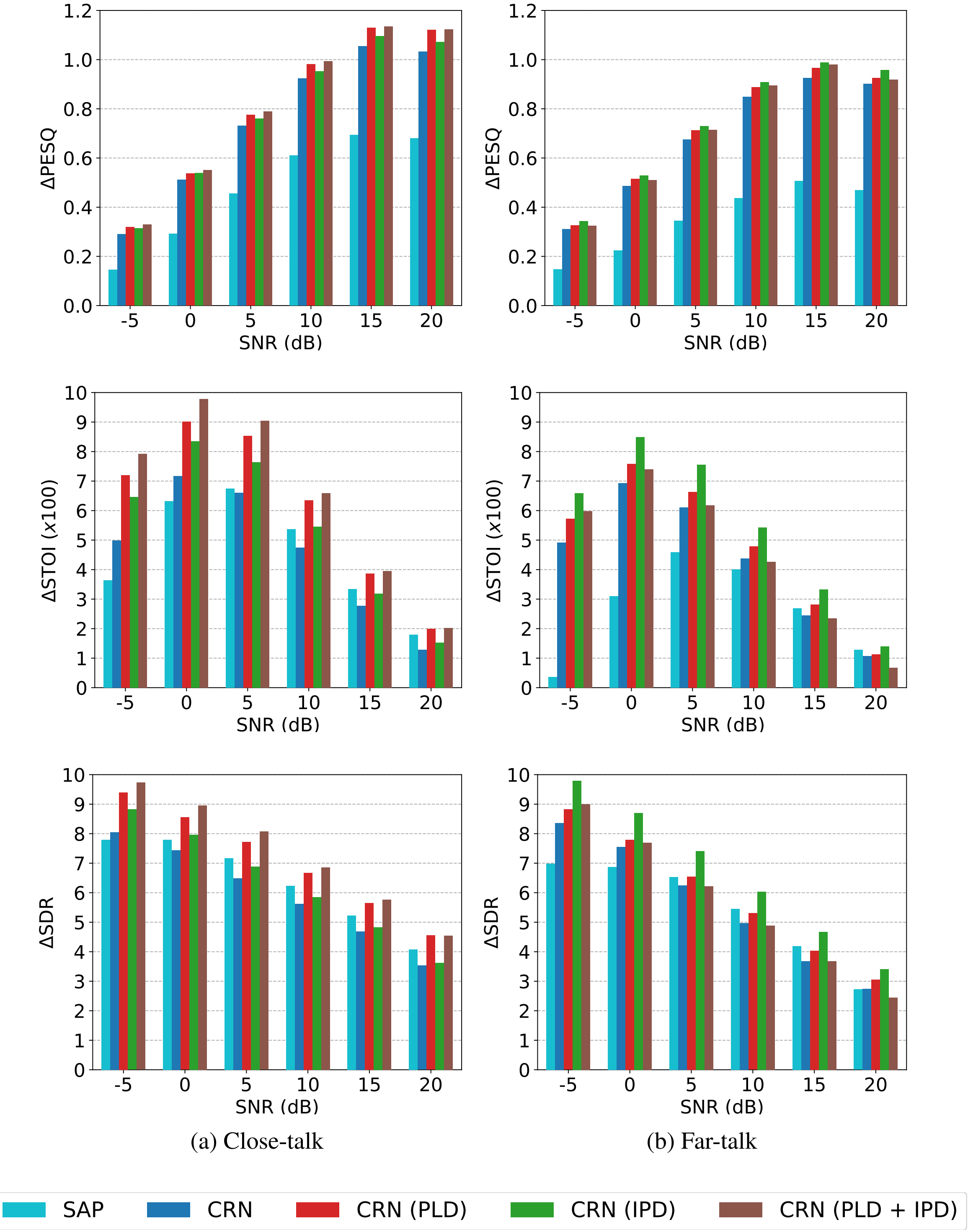


Fig. 4.9 PESQ, STOI and SDR results from the evaluation of the CRN-based SPP mask estimator with the different input features. The OMLSA postfilter with MVDR-based PSD estimation is also showed for comparison purposes. The plots only show increments with respect to the results over noisy speech.

the proposed statistical multichannel SPP estimator using the best a priori SAP estimators for each condition (PLDn for CT and P&C for FT).

The results for CT conditions show that the CRN approach outperforms the SAP-based methods, especially in terms of speech quality. Moreover, the CRN estimator benefits from the use of the dual-channel features. As a result, the PLD+IPD approach obtains the overall best results in the evaluated metrics. The comparison between dual-channel features reveals that the CRN estimator mainly benefits from the PLD features. The CRN estimator using PLD features perform better than the one using IPD features, and it also achieves comparable results with the full integration approach. As observed in the previous experiments, the power difference between microphones is a good indicator of speech presence in CT conditions. Thus, the CRN estimator can exploit this information while predicting SPP masks, achieving more accurate results. Although the CRN with IPD features improve with respect to the plain CRN approach using single-channel information only, the use of both dual-channel features does not show advantages over the standalone PLD features. Therefore, PLD features are a good choice for a trade-off between good performance and network complexity, as this approach requires fewer network parameters.

The CRN approach also achieves better performance than the statistical SAP-based approach in FT conditions. In this case, the use of IPD features stands as the best choice. These features perform better than any other tested combination, especially in terms of ESTOI and SDR scores. On the other hand, the use of PLD features slightly improves the plain CRN approach. Unlike in the CT scenario, the use of power difference features does not provide enough information for the CRN estimator. This is because the power between channels is more similar in this scenario. The phase difference is then the main information source to distinguish between clean speech and noise signals. Finally, the combination of both dual-channel features does not show improvements, but it degrades the performance when comparing with the isolated use of PLD or IPD features. This can be explained by the CRN estimator not being able to deal with multiple input sources in this challenging scenario. Besides, the use of additional PLD features may mislead the network as they do not provide accurate information. As a conclusion, the use of LMS plus IPD features lies as the best alternative in the SPP estimation with CRN models in FT conditions.

4.5 Summary

In this chapter, we have presented the dual-channel speech enhancement framework based on eKF-RTF estimation, which is intended for dual-microphone smartphones used in CT or FT conditions. This framework uses an MVDR beamformer followed by a single-channel postfilter that exploits the dual-channel information. We have first introduced the framework and the different steps involved in the noisy speech processing. In addition, we have described the two postfilters evaluated in our proposal: the pWF and OMLSA filters. These postfilters use single-channel clean speech and noise statistics along with the a posteriori SPP to increase the noise reduction performance. Besides, two single-channel clean speech PSD estimators have been proposed, one of them based on the PLD between the channels and the other using an MVDR-based estimation.

The eKF-RTF estimator has then been introduced for tracking the RTF variability between the microphones of the smartphone in reverberant environments. We have first formulated the state-space models for both the RTF variability across frames and the noisy observations in the secondary channel given the reference microphone. Then, the Kalman filter framework was applied to define the equations for the RTF estimation at each frame given the previous predictions. The issue of dealing with non-linear models has been addressed by using first-order VTS linearization. Moreover, the framework needs a priori information about the RTF statistics, which was obtained in advance by using a training set of dual-channel reverberated speech signals.

The previous steps need noise statistic estimates and the speech presence probability at each time-frequency bin. An SPP-based noise estimation using time-recursive averaging was proposed to update the noise statistics. Two different methods were proposed for the estimation of the a posteriori SPP. The first method uses statistical spatial models based on multivariate Gaussian likelihoods. This spatial model also requires the knowledge of the a priori SAP. Therefore, we proposed dual-channel a priori SAP estimators based on the PLD difference and the CDR ratio between the microphones. Finally, we addressed the direct estimation of the a posteriori SPP by using a DNN mask estimator. A CRN architecture was applied and extended dual-channel features (power level and phase difference) were explored to improve the estimation accuracy.

We concluded the chapter with an experimental evaluation of the different approaches proposed. The proposals were evaluated by using objective quality metrics and a simulated noisy speech database recorded by a dual-microphone smartphone in both CT and FT conditions. We first evaluated the performance of the proposed estimators based on classic signal processing, including the a priori SAP-based noise estimators, the eKF-RTF framework, and the single-channel clean speech PSD estimation. The results showed that our proposed SAP estimators achieve better results than other methods from the state-of-the-art when used in an MVDR beamformer configuration. Our eKF-RTF estimator also showed better accuracy than the classic sub-space search methods, achieving lower signal distortion. Regarding the PSD estimators, the MVDR-based outperformed the PLD-based approach in both CT and

FT configurations. Then, we evaluated the performance of the pWF and OMLSA postfilters using these estimators, showing improvements with respect to other dual-channel speech enhancement methods in terms of speech quality and intelligibility. Finally, a CRN-based a posteriori SPP estimation was also evaluated when used along with our proposed dual-channel postfiltering frameworks. Different combinations of dual-channel features were tested and the results were compared with the approach using a priori SAP-based estimation and single-channel features. The results showed that the DNN mask estimator outperforms the statistical model-based approach and that the DNN can successfully exploit the dual-channel features to improve the accuracy of the estimation. Specifically, the power differences are preferred in CT conditions, while the phase differences are more discriminative in FT conditions.

Chapter 5

Multichannel speech enhancement using a recursive EM algorithm with DNN-based speech presence priors

Multi-microphone devices have spread in recent years thanks to the improvements in technology and the convenience of more advanced tools in our daily life. In the case of speech technologies, it is required that these devices provide good performance, mobility, and low-latency processing. The availability of microphone arrays in these devices allows for the use of multichannel speech enhancement techniques, which makes it possible the use of these mobile devices in environments with different sources of distortion. In contrast to previous dual-microphone smartphones, the presence of several microphones, especially in bigger size devices as tablets, makes it suitable the use of beamforming techniques. The use of a beamforming-plus-postfiltering architecture, as in the case of the multichannel Wiener filter (MWF), can also be employed for boosting the noise reduction capabilities. Recently, a novel multichannel Kalman filter (MKF) was proposed in [244]. It was also demonstrated that MKF can be decomposed into an MVDR beamformer followed by a single-channel modulation-domain Kalman filter (KF) [192]. That approach showed improvements with respect to the classical MWF, as it is able to model the temporal correlations of the clean speech STFT amplitudes. The properties of KF has also been exploited in other speech-related tasks, as speech dereverberation [184, 15] or noise statistics tracking [10].

These multichannel algorithms require knowledge about the involved acoustic parameters, as the statistics of clean speech and noise and the RTF for the beamformer and, in the case of KF, a linear prediction model for the clean speech spectra amplitudes. These statistics have to be estimated in advance, and different methods have been proposed in the literature to obtain each of them, as those reviewed in Chapter 2. Another approach is the adoption of a

Bayesian framework, so that the acoustic parameters are jointly estimated using a maximum likelihood estimation (MLE). Nonetheless, the MLE procedure requires knowledge about the clean speech statistics. The expectation-maximization (EM) iterative algorithm solves this problem, allowing for the joint estimation of the clean speech signal and the acoustic parameters. The EM framework has been extensively studied for offline speech processing tasks [40, 213, 186, 159, 66]. On the other hand, the REM algorithm [17] can be used for online scenarios, as it applies the EM iterative procedure at each time frame using only current and past information. A REM framework for multichannel speech enhancement was proposed in [185], but it did not make use of any speech presence information. Besides, as main disadvantage it requires a priori knowledge of some of the acoustic parameters. Other EM approaches carry out the estimation of the speech presence during the E-step [93, 79, 209], while the noise statistics are obtained in the M-step. The SPP estimation of these approaches can be improved by using DNN mask estimators [156, 148]. Finally, a joint estimation of the clean speech signal, the predominant speaker, and the acoustic parameters was proposed in [183] for offline blind source separation. However, that approach presented several drawbacks, including the distortion introduced by the SPP masking in the estimated speech signals or the potentially high number of iterations needed.

In this chapter, we describe a novel REM framework for multichannel speech enhancement which also incorporates a DNN-based SPP estimator. Our approach uses the beamforming-plus-postfiltering method, employing two different single-channel postfilters, Wiener filter and Kalman filter. Moreover, the framework allows further refinement of the a priori SPP estimates, obtained by the DNN mask estimator, using statistical spatial models defined in terms of the acoustic parameters. This way, these acoustic parameters are re-estimated in each iteration using the obtained clean speech and SPP estimates. The KF-based posfiltering of our proposal is inspired by the Switching Kalman filter framework (SKF) proposed in [154]. We simplify the SKF into two models which either consider speech presence or absence, while the state transition probabilities are replaced by the DNN SPP estimates. As an advantage, our approach allows for the joint estimation, in an online fashion, of the required statistics and acoustic parameters, thus allowing better performance and suitability for real-time applications.

The remainder of this chapter is structured as follows. The statistical framework for the multichannel noisy speech signal is first presented in Section 5.1. Section 5.2 focus on some particularly relevant aspects for the implementation of the framework, as the derivation of the E-step and M-step, the integration of the DNN-based SPP estimation, and the algorithm issues for correct performance. Finally, the experimental results are shown in Section 5.3, where the proposal is evaluated and compared with other state-of-the-art approaches.

5.1 Formulation of the multichannel statistical model

Let us first consider the multichannel noisy speech signal, captured by a microphone array, in the STFT domain under a narrowband assumption, as $\mathbf{y}(t,f)$ in (2.38). The model in (2.38) is defined under a speech presence hypothesis ($\mathscr{H}_x$) and it can be simplified to $\mathbf{y}(t,f) = \mathbf{n}(t,f)$ when speech is absent ($\mathscr{H}_n$). A binary random variable $\mathscr{D}(t,f)$ indicates speech presence/absence for each time-frequency bin, which can be described through the a priori SPP,

$$q_x(t,f) = P\left(\mathscr{D}(t,f) = \mathscr{H}_x\right) = 1 - q_n(t,f). \tag{5.1}$$

Moreover, the clean speech signal at the reference microphone $X_{1,t}$ is a zero-mean circularly symmetric complex random variable, and its variance, under speech presence assumption, can be defined as

$$\sigma_x^2(t,f) = E\left\{ |X_1(t,f)|^2 \middle| \mathscr{H}_x \right\}. \tag{5.2}$$

From now on, with no loss of generality, we will omit the frequency index f for the sake of simplicity.

In addition to the previous multichannel distortion model, we also assume that the clean speech amplitudes from nearby frames can be modeled by the following single-channel temporal linear prediction model [244],

$$|X_1(t)| = \mathbf{a}^\top(t)\mathbf{x}(t-1) + \phi(t), \tag{5.3}$$

where

$$\mathbf{x}(t-1) = \begin{bmatrix} |X_1(t-1)| & |X_1(t-2)| & \cdots & |X_1(t-p)| \end{bmatrix}^\top \tag{5.4}$$

is a vector of clean speech amplitudes from previous frames,

$$\mathbf{a}(t) = \begin{bmatrix} A_1(t) & A_2(t) & \cdots & A_p(t) \end{bmatrix}^\top \tag{5.5}$$

is a vector of linear prediction coefficients (LPC), $\phi(t) \sim \mathscr{N}\left(0, \sigma_v^2(t)\right)$ is the linear prediction error and p is the prediction order.

We can now define the likelihood of the data sequence until frame t as

$$f\left(\mathbf{y}(1:t), X_1(1:t), \mathscr{D}(1:t); \Theta(t)\right), \tag{5.6}$$

where $\mathbf{y}(1:t)$ is the set of observable data until time t, $X_1(1:t)$ and $\mathscr{D}(1:t)$ are the set of latent variables, and $\Theta(t) = \left\{\mathbf{a}(t), \sigma_v^2(t), \mathbf{h}(t), \Sigma_N(t), q_x(t)\right\}$ are the set of model parameters (we will come back to them latter). Using the aforementioned spatial and temporal models, and assuming Markov processes for them, this likelihood can be developed as

$$\begin{aligned} & f\left(\mathbf{y}(1:t), X_1(1:t), \mathscr{D}(1:t); \Theta(t)\right) = \\ & f\left(\mathbf{x}(0)\right) \prod_{\tau=1}^{t} P\left(\mathscr{D}(\tau)\right) \cdot f\left(|X_1(\tau)| | \mathbf{x}(\tau-1), \mathscr{D}(\tau); \mathbf{a}(t), \sigma_v^2(t)\right) \cdot \\ & f\left(X_1(\tau) | |X_1(\tau)|, \mathscr{D}(\tau)\right) \cdot f\left(\mathbf{y}(\tau) | X_1(\tau), \mathscr{D}(\tau); \mathbf{h}(t), \Sigma_N(t)\right). \end{aligned} \tag{5.7}$$

The previous expression distinguish between the prediction model for the speech amplitudes,

$$f\left(|X_1(\tau)| | \mathbf{x}(\tau-1), \mathscr{D}(\tau); \mathbf{a}(t), \sigma_v^2(t)\right), \tag{5.8}$$

and the multichannel distortion model given the clean speech signal,

$$f\left(\mathbf{y}(\tau) | X_1(\tau), \mathscr{D}(\tau); \mathbf{h}(t), \Sigma_N(t)\right). \tag{5.9}$$

The term $f\left(X_1(\tau) | |X_1(\tau)|, \mathscr{D}(\tau)\right)$ comprises the relation between the clean speech amplitude and the complex value of the clean speech, so it can be related to the phase of the signal. We will not deal with the phase estimation, so that this term will not be considered.

We are interested in an online estimation of the clean speech signal at the reference microphone. To this end, we will consider the joint estimation of the latent variables and the model parameters using the likelihood model previously defined. In the next section, we will describe the proposed online algorithm to deal with the joint estimation of the different variables and parameters.

5.2 REM algorithm for multichannel speech enhancement

The joint estimation of the latent variables and the model parameters from the likelihood in (5.7) is a cumbersome task that has no closed-form solution. Moreover, we want these variables to be estimated in an online fashion, which makes the procedure more difficult. Instead of estimating all the variables at once, we can achieve a good approximate of them by using an iterative procedure that, in successive steps, further improve the estimation. We propose the use of the recursive expectation-maximization (REM) algorithm [17], a frame-wise procedure which is repeated for a given frame until a number of iterations is reached. To this end, we define the exponentially-weighted log-likelihood of the data sequence at time t,

$$\mathscr{L}_\lambda(t) = \sum_{\tau=1}^{t} \lambda^{t-\tau} \log f\left(\mathbf{y}(\tau), X_1(\tau), \mathscr{D}(\tau); \Theta(\tau)\right), \tag{5.10}$$

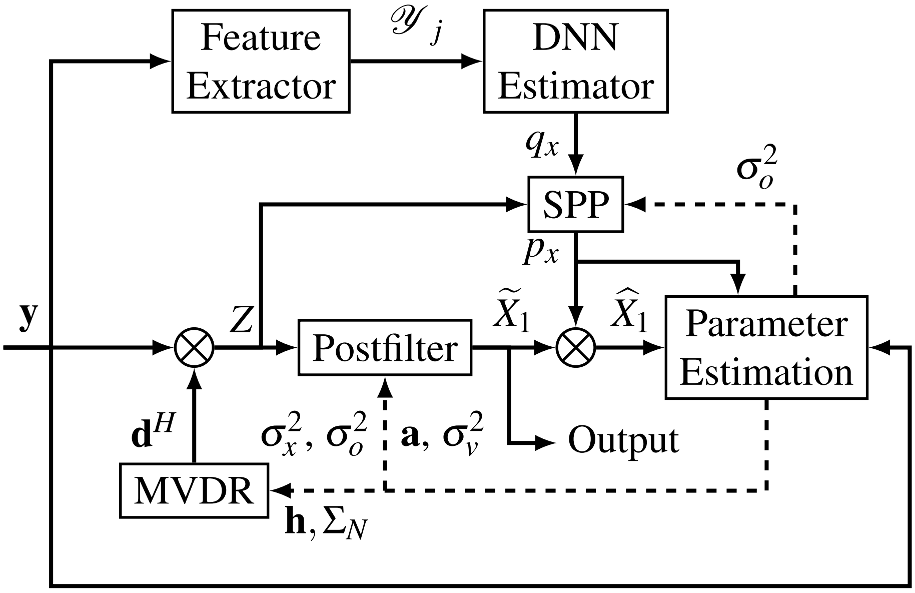


Fig. 5.1 Block diagram of the proposed REM algorithm for multichannel speech enhancement, depicting the most relevant parts. The dashed lines indicate the feedback due to the M-step.

where $\lambda \in (0,1]$ is a forgetting factor. Given a set of model parameters $\Theta^l(t)$ at iteration l, the following two-step procedure is performed in the next iteration:

- **E-step**: An auxiliary function Q is calculated taking the conditioned expectation of the log-likelihood $\mathscr{L}_\lambda(t)$ given the observations and the current parameters,

$$Q\left(\Theta(t)|\Theta^l(t)\right) = E\left\{\mathscr{L}_\lambda(t)|\mathbf{y}(t);\Theta^l(t)\right\}. \tag{5.11}$$

  This results in a function that depends on the conditional expectations over the latent variables $X_1(t)$ and $\mathscr{D}(t)$.

- **M-step**: A new set of parameters is obtained by means of maximum likelihood estimation (MLE) over the auxiliary function Q,

$$\Theta^{l+1}(t) = \underset{\Theta(t)}{\operatorname{argmax}}\, Q\left(\Theta(t)|\Theta^l(t)\right). \tag{5.12}$$

Fig. 5.1 depicts a diagram of the proposed REM algorithm for multichannel speech enhancement, which involves both the E-step and M-step procedures. In the next subsections, we will describe how these steps are performed as well as we will detail each block in Fig. 5.1. For simplicity, we will omit the iteration index l in the following.

5.2.1 E-step: Estimation of the latent variables

The E-step of the REM algorithm considers the computation of the expectation in (5.11), which gives the auxiliary Q function. This function can be re-written considering (5.7) and (5.10), as follows

$$\begin{aligned} &Q\left(\Theta(t)|\Theta^{l}(t)\right) = C_1 + \\ &\sum_{\tau=1}^{t}\lambda^{t-\tau}\sum_{\mathscr{D}(\tau)} p_{\mathscr{D}(\tau)} E_{\Theta,\tau}\left\{\log f\left(|X_1(\tau)|\,|\,\mathbf{x}(\tau-1),\mathscr{D}(\tau);\mathbf{a}(t),\sigma_v^2(t)\right)\right\} + \\ &\sum_{\tau=1}^{t}\lambda^{t-\tau}\sum_{\mathscr{D}(\tau)} p_{\mathscr{D}(\tau)} E_{\Theta,\tau}\left\{\log f\left(\mathbf{y}(\tau)\,|X_1(\tau),\mathscr{D}(\tau);\mathbf{h}(t),\Sigma_N(t)\right)\right\}, \end{aligned} \tag{5.13}$$

where $p_{\mathscr{D}(t)} = P\left(\mathscr{D}(t)|\,\mathbf{y}(t);\Theta(\tau)\right)$ and $E_{\Theta,t}\{\cdot\} = E\{\cdot\,|\mathbf{y}(t);\Theta(t)\}$. The terms C_i in the expressions jointly refer to all those terms that are independent of the model parameters of interest and, therefore, can be neglected. For example, this includes the term related to the phase, as well as those terms depending on the parameter $q_x(t)$, as its estimation will not be performed using the M-step.

The computation of the terms in the Q function requires the conditional expectations for the latent variables given the observations and the model parameters. For latent variable $\mathscr{D}(t)$, this expectation is the a posteriori SPP,

$$p_x(t) = P\left(\mathscr{D}(t) = \mathscr{H}_x|\,\mathbf{y}(t);\Theta(t)\right). \tag{5.14}$$

On the other hand, the clean speech signal $X_{1,t}$ is described using its first- and second-order conditioned expectations. These expectations are obtained by applying the REM framework as indicated in [183],

$$\widehat{X}_1(t) = E_{\Theta,t}\{X_1(t)\} = p_x(t)\widetilde{X}_1(t), \tag{5.15}$$

$$S_x(t) = E_{\Theta,t}\left\{|X_1(t)|^2\right\} = \left|\widehat{X}_1(t)\right|^2 + P(t), \tag{5.16}$$

where

$$\widetilde{X}_1(t) = E\{X_1(t)|\,\mathbf{y}(t),\mathscr{H}_x;\Theta(t)\} \tag{5.17}$$

is the filtered clean speech signal under the speech presence assumption [209], and

$$P(t) = E\left\{\left|X_1(t) - \widehat{X}_1(t)\right|^2\middle|\mathscr{H}_x\right\} \tag{5.18}$$

is the error variance for the estimated clean speech signal when speech presence is assumed (i.e. $\widehat{X}_1(t) = \widetilde{X}_1(t)$). Therefore, the E-step mainly consists of the estimation of the expectations for the clean speech signal and the a posteriori SPP.

First, the expectations in (5.17) and (5.18) are obtained using a multichannel MMSE estimator. As already mentioned, this estimator can be implemented using a beamformer followed by a single-channel linear postfilter. Thus, we first apply an MVDR beamformer (2.44) to the noisy speech signal,

$$Z(t) = \mathbf{d}^H(t)\mathbf{y}(t) = X_1(t) + O(t), \tag{5.19}$$

where $\mathbf{d}(t)$ are the beamformer weights and $O(t) \sim \mathcal{N}\left(0, \sigma_o^2(t)\right)$ is the residual noise at the beamformer output, with a variance given by (2.45). Then, a postfilter is applied to Z_t to obtain $\widetilde{X}_{1,t}$ and the error variance P_t. This postfilter only modifies the amplitude of Z_t, while the phase remains the same. Finally, we use $\widetilde{X}_1(t)$ as the output signal in our framework instead of $\widehat{X}_1(t)$. This is because, in practice, the SPP masking in (5.15) introduces severe speech distortions in the clean speech signal. Nevertheless, the estimation $\widehat{X}_{1,t}$ is still required for the M-step.

Two different linear postfilters are considered in the proposed REM framework, the Wiener and the Kalman filter. These postfilters are implemented as follows:

- **Wiener filter (WF)**: WF only considers the variance of the clean speech and the noise at the current time-frequency bin. The filtered clean speech signal is obtained as

  $$\widetilde{X}_1^{(\mathrm{WF})}(t) = W(t)Z(t) \tag{5.20}$$

  where $W(t)$ is the Wiener gain computed as in (2.13), with the a priori SNR obtained as $\xi(t) = \sigma_x^2(t)/\sigma_o^2(t)$. The error variance is then computed as

  $$P^{(\mathrm{WF})}(t) = (1 - W(t))\,\sigma_x^2(t). \tag{5.21}$$

- **Kalman filter (KF)**: KF takes into account the linear prediction model for the speech amplitudes in (5.3). The proposed KF filter slightly modifies the modulation-domain KF proposed in [244], which follows the standard Kalman filtering [102]. In our case, we only estimate the filtered speech signal at the current frame given the estimated clean speech values from previous frames, instead of the complete vector state for different frames. First, we consider an estimate $\widehat{\mathbf{x}}(t-1)$ of the vector of clean speech amplitudes across the previous frames, as in (5.4). Similarly, we consider a vector $\widetilde{\mathbf{x}}(t-1)$ with the filtered amplitudes from previous frames. The difference between

$\widehat{\mathbf{x}}(t-1)$ and $\widetilde{\mathbf{x}}(t-1)$ is that the latter does not include the SPP masking of (5.15). In addition, we define

$$\mathbf{P}(t-1) = E\left\{(\mathbf{x}(t-1) - \widetilde{\mathbf{x}}(t-1))\,(\mathbf{x}(t-1) - \widetilde{\mathbf{x}}(t-1))^{\top}\right\}, \tag{5.22}$$

as the error covariance matrix of the filtered clean speech amplitudes from previous frames.

KF is then applied to the beamformer output as follows. First, the temporal prediction model in (5.3) is used to obtain an initial prediction as

$$\left|\widetilde{X}_1(t|t-1)\right| = \mathbf{a}^{\top}(t)\widehat{\mathbf{x}}(t-1), \tag{5.23}$$

and its corresponding error variance,

$$P(t|t-1) = \mathbf{a}^{\top}(t)\mathbf{P}(t-1)\mathbf{a}(t) + \sigma_v^2(t). \tag{5.24}$$

Then, this prediction is combined with the observation model in (5.19), which yields the following linear MMSE estimator,

$$\left|\widetilde{X}_1^{(\mathrm{KF})}(t)\right| = \left|\widetilde{X}_1(t|t-1)\right| + K(t)\left(|Z(t)| - \left|\widetilde{X}_1(t|t-1)\right|\right) \tag{5.25}$$

where

$$K(t) = \frac{P(t|t-1)}{P(t|t-1) + \sigma_o^2(t)} \tag{5.26}$$

is the Kalman gain. The error variance is then computed as

$$P^{(\mathrm{KF})}(t) = (1 - K(t))\,P(t|t-1). \tag{5.27}$$

Moreover, the cross-covariance error vector between the current and previous frames can be obtained as

$$\begin{aligned}\mathbf{p}(t,t-1) = E\left\{\left(\,|X_1(t)| - \left|\widetilde{X}_1(t)\right|\right)\left(\mathbf{x}(t-1) - \widetilde{\mathbf{x}}(t-1)\right)^{\top}\right\} = \\ = (1 - K(t))\,\mathbf{a}^{\top}(t)\mathbf{P}(t-1).\end{aligned} \tag{5.28}$$

Finally, the estimation $\widetilde{X}_1^{(\mathrm{KF})}(t)$ is obtained by using the magnitude $\left|\widetilde{X}_1^{(\mathrm{KF})}(t)\right|$ from (5.25) and keeping the phase of the MVDR output $Z(t)$.

The values $\widehat{\mathbf{x}}(t)$ and $\mathbf{P}(t)$ needed for the next frame are obtained as

$$\widehat{\mathbf{x}}(t) = \mathbf{U}\widehat{\mathbf{x}}(t-1) + \mathbf{u}\left|\widetilde{X}_1^{(\mathrm{KF})}(t)\right|, \tag{5.29}$$

$$\mathbf{P}(t) = \mathbf{U}\mathbf{P}(t-1)\mathbf{U}^\top + \mathbf{U}\mathbf{p}(t|t-1)\mathbf{u}^\top + \mathbf{u}\mathbf{p}^\top(t|t-1)\mathbf{U}^\top + \mathbf{u}P^{(\mathrm{KF})}(t)\mathbf{u}^\top, \tag{5.30}$$

where

$$\mathbf{u} = \begin{bmatrix} 1 & \mathbf{0}_{1\times p-1} \end{bmatrix}^\top, \tag{5.31}$$

$$\mathbf{U} = \begin{bmatrix} \multicolumn{2}{c}{\mathbf{0}_{1\times p}} \\ \mathbf{I}_{p-1\times p-1} & \mathbf{0}_{p-1\times 1} \end{bmatrix}, \tag{5.32}$$

are, respectively, a structure vector and matrix, in which $\mathbf{0}$ is a zero vector and $\mathbf{I}$ is the identity matrix (the sizes of them are indicated by a subindex). When no prediction of the clean speech amplitudes is performed (the coefficients $\mathbf{a}(t)$ are zero), and then $P(t|t-1) = \sigma_x^2(t)$, both WF and KF are equivalent. Therefore, the KF can be seen as a generalization of WF.

Finally, the a posteriori SPP in (5.14) is computed. Using the Bayes' rule, the expression in (5.14) can be re-written as

$$p_x(t) = \frac{q_x(t) f\left(\mathbf{y}(t)|\mathscr{H}_x;\Theta(t)\right)}{q_x(t) f\left(\mathbf{y}(t)|\mathscr{H}_x;\Theta(t)\right) + \left(1-q_x(t)\right) f\left(\mathbf{y}(t)|\mathscr{H}_n;\Theta(t)\right)}. \tag{5.33}$$

In our proposed model, the likelihoods in (5.33) are multivariate Gaussian distributions as those in (4.41). Additionally, they can be expressed directly from the MVDR output. Using the MVDR definition in (2.44), the multivariate likelihoods can be simplified into the following single Gaussian likelihoods,

$$f\left(\mathbf{y}(t)|\mathscr{H}_x;\Theta(t)\right) = \mathscr{N}\left(Z(t);0,\sigma_z^2(t)\right), \tag{5.34}$$

$$f\left(\mathbf{y}(t)|\mathscr{H}_n;\Theta(t)\right) = \mathscr{N}\left(Z(t);0,\sigma_o^2(t)\right), \tag{5.35}$$

where

$$\sigma_z^2(t) = p_x(t)S_x(t) + \sigma_o^2(t). \tag{5.36}$$

is the estimated variance of Z_t, which is the sum of the clean speech variance (considering the speech presence) and the residual noise variance. In our iterative procedure, the a posteriori SPP is first initialized as $p_x^{l=0}(t) = q_x(t)$ and, after applying the postfiltering step, it is updated during the E-step at each iteration.

5.2.2 M-step: Estimation of the model parameters

The M-step addresses the MLE estimation of the model parameters given the expectations of the latent variables obtained in the E-step. These model parameters are the Kalman filter parameters, $\mathbf{a}(t)$ and $\sigma_v^2(t)$, and the beamformer parameters, $\mathbf{h}(t)$ and $\Sigma_N(t)$.

The KF parameters are obtained from the second term of the Q function in (5.13). This term is conditioned to that speech is present. Given the stochastic process in (5.3), which follows a Gaussian distribution, the term can be expanded as

$$\begin{aligned}&E_{\Theta,\tau}\left\{\log f\left(|X_1(\tau)|\,|\,\mathbf{x}(\tau-1),\mathscr{H}_x;\mathbf{a}(t),\sigma_v^2(t)\right)\right\}=C_2-\frac{1}{2}\log\sigma_v^2(t)-\\&\frac{1}{2}E_{\Theta,\tau}\left\{\left[|X_1(\tau)|-\mathbf{a}^\top(t)\mathbf{x}(\tau-1)\right]^\top\sigma_v^{-2}(t)\left[|X_1(\tau)|-\mathbf{a}^\top(t)\mathbf{x}(\tau-1)\right]\right\}.\end{aligned}\tag{5.37}$$

The parameters of KF can change quickly, so their estimation should be done in a frame-wise fashion. To this end, we compute the derivatives of the Q function with respect to the KF parameters in the case $\lambda=0$ (i.e. the Q function only considers the log-likelihood at the current frame),

$$\frac{\partial Q_{\lambda=0}}{\partial\mathbf{a}^\top(t)}=p_x(t)E_{\Theta,t}\left\{\sigma_v^{-2}(t)\left[|X_1(t)|-\mathbf{a}^\top(t)\mathbf{x}(t-1)\right]\mathbf{x}^\top(t-1)\right\},\tag{5.38}$$

$$\begin{aligned}\frac{\partial Q_{\lambda=0}}{\partial\sigma_v^{-2}(t)}&=-\frac{1}{2}p_x(t)\left(E_{\Theta,t}\left\{|X_1(t)|^2-\mathbf{a}^\top(t)\mathbf{x}(t-1)\,|X_1(t)|-\right.\right.\\&\left.\left.\mathbf{x}^\top(t-1)\mathbf{a}(t)\,|X_1(t)|+\mathbf{a}^\top(t)\mathbf{x}(t-1)\mathbf{x}^\top(t-1)\mathbf{a}(t)\right\}-\sigma_v^2(t)\right),\end{aligned}\tag{5.39}$$

so only the instantaneous statistics are used. The MLE estimates are obtained by making the derivatives equals to zero, which gives the following expression for the parameters,

$$\mathbf{a}(t)=E_{\Theta,t}\left\{\mathbf{x}(t-1)\mathbf{x}^\top(t-1)\right\}^{-1}E_{\Theta,t}\left\{|X_1(t)|\,\mathbf{x}(t-1)\right\},\tag{5.40}$$

$$\sigma_v^2(t)=E_{\Theta,t}\left\{|X_1(t)|^2\right\}-\mathbf{a}^\top(t)E_{\Theta,t}\left\{\mathbf{x}(t-1)\mathbf{x}^\top(t-1)\right\}\mathbf{a}(t).\tag{5.41}$$

These expressions can also be formulated as

$$\mathbf{a}(t)=\mathbf{R}_x^{-1}(t-1)\mathbf{r}_x(t,t-1),\tag{5.42}$$

$$\sigma_v^2(t)=\sigma_x^2(t)-\mathbf{a}^\top(t)\mathbf{R}_x(t-1)\mathbf{a}(t),\tag{5.43}$$

where

$$\mathbf{R}_x(t-1) = E_{\Theta,t}\left\{\mathbf{x}(t-1)\mathbf{x}^\top(t-1)\right\} = \widehat{\mathbf{x}}(t-1)\widehat{\mathbf{x}}^\top(t-1) + \mathbf{P}(t-1), \tag{5.44}$$

$$\mathbf{r}_x(t,t-1) = E_{\Theta,t}\left\{|X_1(t)|\,\mathbf{x}(t-1)\right\} = \left|\widehat{X}_1(t)\right|\widehat{\mathbf{x}}(t-1) + \mathbf{p}(t,t-1) \tag{5.45}$$

are MMSE estimates of the speech signal correlations. It must be noted that the speech variance $\sigma_x^2(t)$ is used instead of $E_{\Theta,t}\left\{|X_1(t)|^2\right\}$ in (5.43) to avoid the distortion introduced by the SPP masking when estimating the filtered signal $\widetilde{X}_1(t)$. In addition, the subtraction in (5.43) could yield negative values. In such cases, the LPC coefficients are set to zero so that $\sigma_v^2(t) = \sigma_x^2(t)$, and KF reduces to WF.

On the other hand, the beamformer parameters, that can be considered slowly variant, are derived from the third term of the Q function in (5.13). Given that the noise signal follows a multivariate complex Gaussian distribution, the expectation can be developed as

$$\begin{aligned} &E_{\Theta,\tau}\left\{\log f\left(\mathbf{y}(\tau)\,|X_1(\tau),\mathscr{H}_x;\mathbf{h}(t),\Sigma_N(t)\right)\right\} = C_3 - \frac{1}{2}\log|\Sigma_N(t)| - \\ &\frac{1}{2}E_{\Theta,\tau}\left\{\left[\mathbf{y}(\tau)-\mathbf{h}(t)X_1(\tau)\right]^H\Sigma_N^{-1}(t)\left[\mathbf{y}(\tau)-\mathbf{h}(t)X_1(\tau)\right]\right\}, \end{aligned} \tag{5.46}$$

when speech is present, and

$$E_{\Theta,\tau}\left\{\log f\left(\mathbf{y}(\tau)\,|\mathscr{H}_n;\Sigma_N(t)\right)\right\} = C_4 - \frac{1}{2}\log|\Sigma_N(t)| - \frac{1}{2}\mathbf{y}^H(\tau)\Sigma_N^{-1}(t)\mathbf{y}(\tau), \tag{5.47}$$

when speech is absent. The RTF is derived directly under speech presence assumption, while the noise SCM considers both speech presence and absence hypotheses. The derivatives of the Q function with respect to these parameters yield the following expressions,

$$\frac{\partial Q}{\partial \mathbf{h}(t)} = \sum_{\tau=1}^{t}\lambda^{t-\tau}p_x(\tau)E\left\{\Sigma_N^{-1}(t)\left[\mathbf{y}(\tau)-\mathbf{h}(t)X_{1,\tau}\right]X_1^*(\tau)\right\}, \tag{5.48}$$

$$\begin{aligned} \frac{\partial Q}{\partial \Sigma_N^{-1}(t)} &= -\frac{1}{2}\sum_{\tau=1}^{t}\lambda^{t-\tau}\Big[\mathbf{y}(\tau)\mathbf{y}^H(\tau) - p_x(\tau)\Big(\mathbf{h}(t)\widehat{X}_1(\tau)\mathbf{y}^H(\tau) + \\ &\mathbf{y}(\tau)\mathbf{h}^H(t)\widehat{X}_1^*(\tau) - \mathbf{h}(t)S_x(\tau)\mathbf{h}^H(t)\Big)\Big] + \frac{1}{2}\Sigma_N(t)\sum_{\tau=1}^{t}\lambda^{t-\tau}. \end{aligned} \tag{5.49}$$

Then, the MLE estimates for the RTF vector and the noise SCM can be obtained by making the derivatives equals to zero, which yields the following expressions,

$$\mathbf{h}(t) = \frac{\sum_{\tau=1}^{t} \lambda^{t-\tau} p_x(\tau) \mathbf{y}(\tau) \widehat{X}_1^*(\tau)}{\sum_{\tau=1}^{t} \lambda^{t-\tau} p_x(\tau) S_x(\tau)}, \tag{5.50}$$

$$\Sigma_N(t) = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} \left(\mathbf{y}(\tau)\mathbf{y}^H(\tau) - p_x(\tau)\mathbf{h}(t) S_x(\tau) \mathbf{h}^H(t) \right), \tag{5.51}$$

where the relation $\sum_{\tau=1}^{t} \lambda^{t-\tau} = \frac{1-\lambda^t}{1-\lambda}$ is applied. These expressions can be re-formulated as

$$\mathbf{h}(t) = \mathbf{r}_{yx}(t) R_x^{-1}(t), \tag{5.52}$$

$$\Sigma_N(t) = \Sigma_Y(t) - \mathbf{h}(t) R_x(t) \mathbf{h}^H(t), \tag{5.53}$$

where

$$R_x(t) = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} p_x(\tau) S_x(\tau), \tag{5.54}$$

$$\mathbf{r}_{yx}(t) = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} p_x(\tau) \mathbf{y}(\tau) \widehat{X}_1^*(\tau), \tag{5.55}$$

$$\Sigma_Y(t) = \frac{1-\lambda}{1-\lambda^t} \sum_{\tau=1}^{t} \lambda^{t-\tau} \mathbf{y}(\tau)\mathbf{y}^H(\tau) \tag{5.56}$$

are smoothed estimates of the clean speech power spectrum, the cross-correlation between the noisy and clean speech, and the noisy speech SCM, respectively. The advantage of the noise estimator in (5.53) is that it can be updated even in speech presence bins, which allows for a quicker adaptation, especially in non-stationary noisy scenarios.

As can be observed, the estimation of the model parameters presented in this subsection requires an estimate of the speech variance under speech presence assumption, $\sigma_x^2(t)$. This variance could be obtained in the M-step of the REM framework by MLE estimation, but this procedure presents two problems. Firstly, this framework assumes that the parameters are slowly-variant in time [17], an assumption that does not generally hold true for the speech variance distribution. Secondly, the estimation procedure takes into consideration the SPP [183]. This yields a high degree of sparsity in the filtered signal at the postfiltering output, which depends directly on this variance. This results in an enhanced speech signal with severe distortion and degraded perceptual quality and intelligibility. Therefore, we propose a different approach for this speech variance, adapted from the estimation method proposed in [184]. In particular, we estimate the speech variance directly from the speech signal at the beamformer output as

$$\sigma_x^2(t) = G_x(t)\,|Z(t)|^2\,, \tag{5.57}$$

where

$$G_x(t) = \frac{\xi(t)}{1+\xi(t)}\left(\frac{1}{\gamma(t)} + \frac{\xi(t)}{1+\xi(t)}\right) \tag{5.58}$$

is a gain function derived as in [237], and $\gamma(t) = |Z(t)|^2 / \sigma_o^2(t)$ is the a posteriori SNR. The advantage of this gain estimator is that it approximates the Wiener suppression rule at high instantaneous SNR, while the attenuation level is decreased otherwise [237]. Nonetheless, the above estimator needs knowledge about the a priori SNR, which is not available (as it depends on the speech variance). Therefore, we propose the following estimate,

$$\widehat{\xi}(t) = \frac{R_z(t)}{\sigma_o^2(t)}, \tag{5.59}$$

where

$$R_z(t) = \frac{1-\lambda}{1-\lambda^t}\sum_{\tau=1}^{t} \lambda^{t-\tau} p_x(\tau)\,|Z(\tau)|^2 \tag{5.60}$$

is a smoothed estimate of the clean speech power spectrum computed from the MVDR output $Z(t)$ and the a posteriori SPP $p_x(t)$.

5.2.3 DNN-based a priori SPP estimation

As previously explained, the M-step is not adequate for the estimation of quickly time-variant parameters, as in the case of the speech variance. The same problem arises with the a priori SPP. Indeed, Taseska *et al.* showed that the MLE estimation of the a priori SPP is not robust enough in non-stationary noisy environments. This is because the REM framework cannot follow the quick changes in the clean speech signal. As mentioned in the previous chapter, several algorithms have been proposed for the estimation of the a priori SPP [27, 194, 209]. On the other hand, DNNs have been explored for the estimation of speech presence masks [78, 169, 81, 20], and they have also be combined with statistical spatial models [156, 148].

We propose to estimate the a priori SPP using a deep neural network (DNN)-based mask estimator [76], thus avoiding assumptions about the multichannel noisy speech signals. Our DNN mask estimator is based on the one proposed in [78, 72]. The model consists of a unidirectional LSTM layer followed by two fully-connected layers with ReLU activations, and an output layer with sigmoid activation. Thus, the mask estimator is appropriate for an online scenario, as only current and past frames are used for the recurrent neural network. The estimator obtains single speech presence mask for each microphone. These masks are latter combined through a median operation, thus providing the final a priori SPP estimate $q_x(t)$. The input feature vector at each frame is the noisy log-magnitude spectrum,

$$\mathscr{Y}_j(t) \triangleq \left[\log\left|Y_j(t,0)\right| \quad \cdots \quad \log\left|Y_j(t,F-1)\right|\right]^\top, \tag{5.61}$$

where j refers to the microphone index and F is the number of frequency bins. A time-recursive mean normalization is applied to these features before feeding them into the network [72]. During the training phase, the target features were ideal binary masks (IBMs), like those proposed in [78], and the loss function was the binary cross-entropy criterion.

5.2.4 Algorithm overview and implementation issues

The different steps of proposed REM algorithm has been described in the previous subsections. As can be observed, the REM framework is a complex algorithm that involves the definition of the E-step and M-step, and the use of DNN-based mask estimators. To clarify the procedure and remark the important steps, we have summarized the REM framework in Algorithm 1, where the order of the different steps is exposed. Also, we highlight the procedures for the E-step and M-step in the algorithm. It must be remarked that the MLE estimations of the KF parameters as well as the estimation of the speech variance are performed before the postfiltering stage for a correct performance of the algorithm.

The efficient implementation and good convergence of the REM framework also require some additional considerations. In the following, we discuss these practical aspects.

Recursive estimation: During the M-step, we have to deal with expressions that include a sum over frames, with the following structure,

$$R_{\mathscr{B}}(t) = \frac{1-\lambda}{1-\lambda^t}\sum_{\tau=1}^{t}\lambda^{t-\tau}\mathscr{B}(\tau), \tag{5.62}$$

where $\mathscr{B}(t)$ refers to any expression computed at each frame t. For an efficient computation, a recursive procedure can be used instead, as follows,

$$R_{\mathscr{B}}(t) = (1-\alpha(t))R_{\mathscr{B}}(t-1) + \alpha(t)\mathscr{B}(t), \tag{5.63}$$

where

$$\alpha(t) = \frac{1-\lambda}{1-\lambda^t} \tag{5.64}$$

is a time-dependent recursive parameter. As an advantage, we have to save only one previous value.

Algorithm 1 REM algorithm with DNN-based SPP estimation

```
Initialize variables and parameters
for each $t$ in $T$ (total frames) do
    Update $\Sigma_Y(t)$ using $\mathbf{y}(t)$ (5.56)
    Update $\mathbf{h}(t)$ (5.65) and $\Sigma_N(t)$ (5.67) if needed
    Compute $q_x(t)$ using DNN and initialize $p_x^0(t) = q_x(t)$
    for $l = 1$ to $l_{\max}$ do
        Beamformer: Compute $Z(t)$ (5.19) and $\sigma_o^2(t)$ (2.45) (E-step)
        Compute speech variance $\sigma_x^2(t)$ (5.57)
        if using Kalman filter then
            Compute $\mathbf{a}(t)$ (5.42) and $\sigma_v^2(t)$ (5.43) (M-step)
        end if
        Postfilter: Estimate $\widetilde{X}_1(t)$ (5.17) and $P(t)$ (5.18) (E-step)
        Estimate $\widehat{X}_1(t)$ (5.15) and $S_x(t)$ (5.16) (E-step)
        Estimate $p_x(t)$ (5.33) (E-step)
        Update $\Lambda(t)$ (5.66)
        Compute $\mathbf{h}(t)$ (5.52) and $\Sigma_N(t)$ (5.53) (M-step)
    end for
    Update variables for next frame
end for
```

Initialization of the relative transfer function: The RTF estimation in (5.52) is only possible after the processing of the first speech frames. This can degrade the performance of the MVDR beamformer, as it could be not correctly steered towards the target speaker during these first speech frames (or after long speech absence periods). To overcome this problem, we initialize the RTF before the MVDR beamforming step in those bins with no recent speech activity. To this end, we employ eigenvalue decomposition (EVD) [188] on an estimate of the speech covariance matrix,

$$\mathbf{h}^{l=0}(t) = \mathscr{P}\left(\Sigma_Y(t) - \Sigma_N(t)\right), \tag{5.65}$$

where operator $\mathscr{P}(\cdot)$ computes the principal eigenvector of a matrix. The speech activity in each time-frequency bin is quantified by using a weighted recursive sum of the SPP in the previous frames,

$$\Lambda(t) = \lambda \Lambda(t-1) + p_x(t), \tag{5.66}$$

with $\Lambda(0) = 0$. At time t, we use the proposed initialization in those bins where $\Lambda(t-1)$ is below a threshold value Λ_{thr}.

Initialization of the noise covariance matrix: A good initialization of the noise SCM during the initial frames can improve the convergence of the REM algorithm. Moreover, these are usually noise-only frames with low speech activity. The noisy observations of these frames can be used to initialize the noise SCM. Therefore, during the first T_{init} frames, we initialize the noise SCM using the following recursion,

$$\Sigma_N^{l=0}(t) = \beta(t)\Sigma_N(t-1) + (1-\beta(t))\,\mathbf{y}(t)\mathbf{y}^H(t) \tag{5.67}$$

where

$$\beta(t) = 1 + (q_x(t)-1)\,\alpha(t). \tag{5.68}$$

is a recursive factor that prevents updating in speech presence bins. This procedure can be seen as an adaptation of the MCRA method proposed in [194]. For the successive iterations or frames, the noise SCM is directly computed using (5.53).

Updating of the KF parameters: The KF parameters should be updated before using the KF postfilter to correctly track the speech variability. The problem is that this computation requires to obtain $\mathbf{r}_x(t,t-1)$ (5.45), which indeed depends on the KF output. To overcome this, we compute a WF filter with SPP masking to approximate $\mathbf{r}_x(t,t-1) \simeq \widehat{|X_1(t)|}\widehat{\mathbf{x}}(t-1)$ for the first EM iteration. Once the KF parameters are obtained, KF can be applied.

5.3 Experimental results

In this section, we describe the performance evaluation of the proposed algorithm using the simulated corpus for the CHiME-4 database. The fifth microphone on the tablet was used as the reference channel for the different algorithms and evaluations. For STFT computation, a 512-point DFT was applied using a 32 ms square-root Hann window with 50% overlap, which resulted in a total of 257 frequency bins for each time frame. The values of the different parameters used in our algorithm are summarized in Table 5.1.

The DNN-based a priori SAP estimator was trained using the training and development sets, while the evaluations were performed on the evaluation set of the database. The DNN model is comprised of an LSTM layer with 512 units, two fully-connected layers with 512 units each, and an output layer with 257 units. During the training phase, a batch size of five utterances was used. To prevent overfitting, dropout was applied over the hidden layers with a de-activation probability factor of 0.5.

We evaluated the performance of the proposed REM framework using either a Wiener postfilter (WF) or a Kalman postfilter (KF). For comparison purposes, the enhanced signal before (REMWF-BF and REMKF-BF) and after (REMWF and REMKF) the postfilter was considered and evaluated in both approaches (i.e. $Z(t)$ and $\widetilde{X}_1(t)$ outputs). In the next subsections, we will show the results obtained in the experimental evaluations, including

Table 5.1 Hyperparameter values used in the proposed algorithm

Param.	λ	$l_{\max}$	p	Λ_{thr}	T_{init}
Value	0.9	2	2	1.0	10

an analysis of the algorithm performance using oracle estimates, an analysis of the SPP estimators, and the computational performance in terms of the number of EM iterations.

5.3.1 Evaluation results for perceptual quality, intelligibility and signal distortion

We first assessed the goodness of the resulting enhanced signal in terms of three objective performance measures: PESQ for perceptual quality, ESTOI for intelligibility, and scale-invariant SDR for signal distortion. We compared our four variants of the REM approach with three different state-of-the-art enhancement methods:

- MVDR beamforming (MVDR) as described in [194], using the formulation in (2.46) with the rank-1 approximation for the clean speech SCM described in [232].
- Multichannel Wiener filter (MWF) using the previous rank-1 approximation.
- Multichannel Kalman filter (MKF) as proposed in [244], implemented as an MVDR beamforming plus a modulation-domain KF. The KF parameters were obtained by LPC analysis using the enhanced signal at the output of the MWF. For a fair comparison, we updated these parameters each frame, using five previous frames in the estimation.

For a fair comparison, all the methods uses the DNN-based SPP estimator to compute the beamformer parameters. Thus, the recursive procedure of the MCRA approach [194] was applied over the noisy speech signal to compute the noise SCM by means of these SPP masks. On the other hand, the RTF was obtained through eigenvalue decomposition [188] of an estimate of the clean speech SCM.

Tables 5.2, 5.3 and 5.4 show the results obtained by the different evaluated methods for PESQ, ESTOI, and SDR, respectively. These tables include the average results for each noisy environment and the overall average of each method. Also, the 95% confidence intervals are included. The results for the original noisy speech signals are also shown as a reference. As can be observed, the proposed REMWF and REMKF approaches outperform the rest of the methods in terms of speech quality, speech intelligibility, and signal distortion. Moreover, the results obtained for REMWF-BF and REMKF-BF approaches are better than those from the reference MVDR beamformer. The postfiltering step yields an increase in the PESQ

Table 5.2 PESQ results for the different evaluated algorithms. Results are broken down by noise environment.

Method	**Noise**				**Avg.**
	BUS	**CAF**	**PED**	**STR**	
Noisy	1.32	1.24	1.26	1.28	1.27 ± 0.01
MVDR	1.71	1.51	1.58	1.57	1.59 ± 0.01
REMWF-BF	1.87	1.67	1.74	1.69	1.74 ± 0.02
MWF	2.07	1.83	1.94	1.90	1.94 ± 0.01
REMWF	2.19	1.96	2.07	1.98	2.05 ± 0.02
REMKF-BF	1.89	1.68	1.75	1.70	1.76 ± 0.02
MKF	1.97	1.68	1.79	1.78	1.81 ± 0.01
REMKF	**2.22**	**1.98**	**2.10**	**2.01**	**2.08** ± **0.02**

Table 5.3 ESTOI (x100) results for the different evaluated algorithms. Results are broken down by noise environment.

Method	**Noise**				**Avg.**
	BUS	**CAF**	**PED**	**STR**	
Noisy	70.9	65.9	68.7	67.1	68.2 ± 0.5
MVDR	82.9	77.2	79.6	78.5	79.5 ± 0.4
REMWF-BF	86.3	82.2	83.8	82.4	83.7 ± 0.4
MWF	83.3	78.0	80.1	79.1	80.1 ± 0.4
REMWF	86.1	81.7	83.2	82.1	83.3 ± 0.4
REMKF-BF	86.8	**82.6**	**84.3**	**82.9**	**84.1** ± **0.4**
MKF	80.6	74.7	76.8	76.3	77.1 ± 0.4
REMKF	**86.9**	82.5	84.0	**82.9**	**84.1** ± **0.4**

and SDR metrics, while ESTOI keeps similar when comparing with the beamformer output. In our proposed REM framework, the results using the KF postfiltering are slightly better than its WF counterpart, both for the REMKF and REMKF-BF approaches. Regarding the reference methods, both MWF and MKF perform better than the MVDR method in terms of PESQ and SDR metrics. Nevertheless, MKF does not perform better than MWF. Finally, the improvements obtained are similar in the different noisy environments analyzed, with REMKF standing as the best approach.

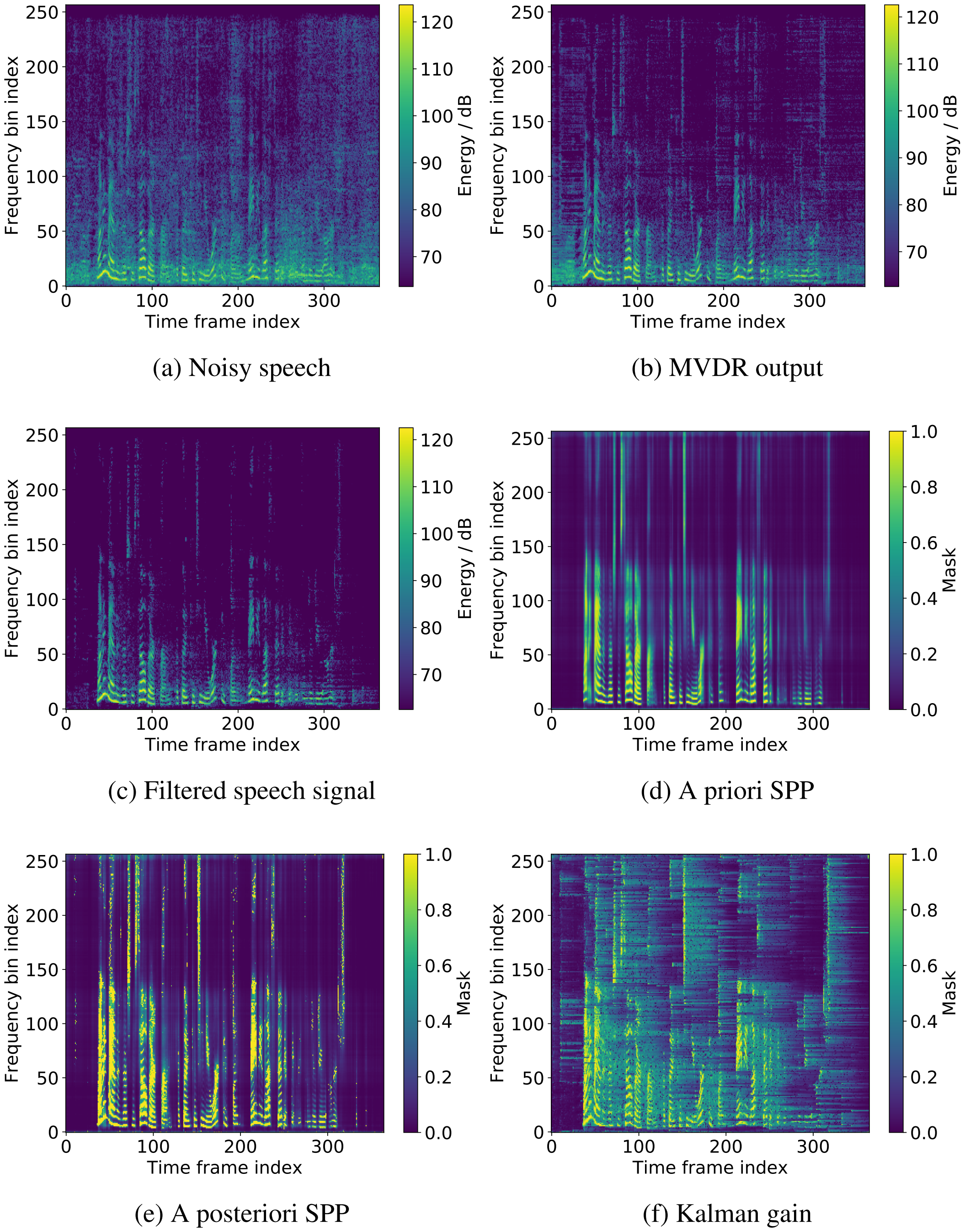
The obtained results show that our REM framework with DNN-based SPP estimates improves the performance of the multichannel speech enhancement task. This can be observed in both beamforming and postfiltering approaches when comparing with the reference methods. The estimation of the beamforming parameters benefits both on the use of the estimated clean speech statistics and an improved SPP estimation using spectral (DNN) and spatial (statistical) models. This remarks the advantage of integrating DNN-based mask

Table 5.4 SDR results (in dB) for the different evaluated algorithms. Results are broken down by noise environment.

Method	**Noise**				**Avg.**
	BUS	**CAF**	**PED**	**STR**	
Noisy	6.79	7.77	8.60	6.86	7.51 ± 0.11
MVDR	11.42	11.12	11.94	10.88	11.34 ± 0.14
REMWF-BF	12.49	12.26	12.83	11.74	12.33 ± 0.14
MWF	13.72	12.44	13.05	12.83	13.01 ± 0.15
REMWF	15.22	14.06	14.47	14.08	14.46 ± 0.16
REMKF-BF	12.63	12.36	12.96	11.88	12.46 ± 0.15
MKF	13.58	11.71	12.39	12.45	12.53 ± 0.16
REMKF	**15.75**	**14.38**	**14.90**	**14.54**	**14.89 ± 0.16**

estimators, which do not require explicit assumptions about the a priori SPP, and statistical spatial models for the noisy speech and noise signals. The postfilter step further enhances the speech signal at the beamformer output, increasing the noise reduction but at the cost of a slight reduction of the speech intelligibility. This translates to an improvement in PESQ and SDR when comparing with the REMWF-BF and REMKF-BF versions, while ESTOI is not severely affected. In addition, independence between the postfiltering step and the SPP masking ensures that no additional signal distortion is introduced in the filtered enhanced signal, which could degrade PESQ and ESTOI results.

Finally, we can compare the Wiener and Kalman postfiltering approaches in our REM framework. As can be concluded, our proposal benefits from using KF, which takes into account the temporal correlations in the amplitude speech signal. Thus, the KF postfilter outperforms the WF performance. This also allows for more accurate computation of the beamforming parameters during the M-step. Nevertheless, the same behavior is not observed when comparing the reference MWF and MKF approaches, where the use of KF degrades the performance. This observation does not match with the results reported in [244]. Nevertheless, the disagreement could be explained in part by the use of longer analysis windows and smaller overlapping between them than in the original implementation, which could lead to a degradation in the MKF performance. The reason we chose the same window length and overlapping in the different compared methods is to integrate the same DNN mask estimator architecture in all of them. Another possible explanation is that the MKF approach applies an LPC analysis on the enhanced speech signal obtained by the MWF, which could still contain residual noise. Besides, several speech frames are needed to obtain the matrices required to solve the involved linear equations system. On the contrary, our proposal directly uses the estimated clean speech statistics of the current frame for the



(a) Noisy speech

(b) MVDR output

(c) Filtered speech signal

(d) A priori SPP

(e) A posteriori SPP

(f) Kalman gain

Fig. 5.2 Noisy and enhanced spectrograms, as well as estimated masks, resulting from the REMKF approach when applied to the audio file F05_444C0214_CAF (cafeteria noise, SDR = 7.23 dB) from the CHiME-4 database.

MLE procedure. Moreover, the SPP masking is also considered, which allows us to better discriminate between speech presence and absence time-frequency bins. As a result, our proposed REM framework achieves a better estimation of the KF parameters.

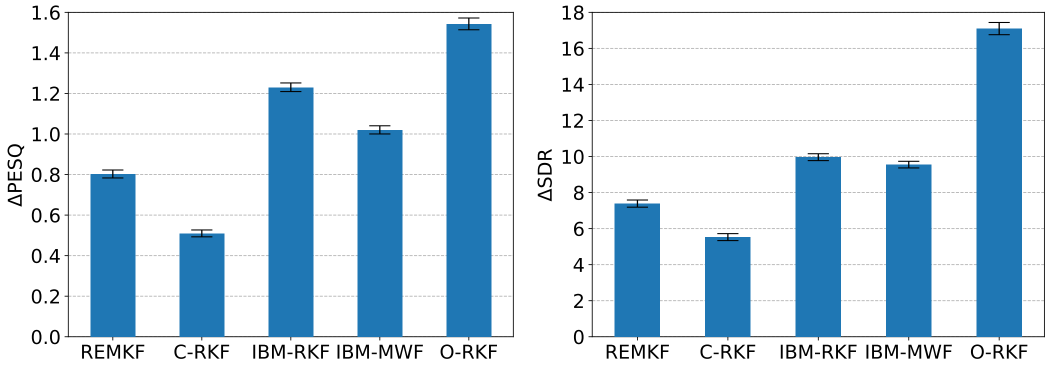
To conclude this section, we show a visual example of a noisy speech signal processed using our REM framework. Thus, Fig. 5.2 shows the example spectrogram of the noisy speech signal (Fig. 5.2.a), the speech signal at the beamformer output (Fig. 5.2.b) and the filtered speech signal at the postfilter output (Fig. 5.2.c) for our REMKF approach. In addition, the corresponding a priori SPP (Fig. 5.2.d), a posteriori SPP (Fig. 5.2.e), and the Kalman gain (Fig. 5.2.f) are also shown.

It can be observed that our approach achieves efficient noise reduction, especially at the output of the postfilter step (Fig. 5.2.c). The postfilter removed much of the residual noise from the beamforming step, especially at medium and high frequencies. Furthermore, the speech formants appear preserved, which is in line with the speech intelligibility and signal distortion results obtained. The a priori SPP (Fig. 5.2.d) estimated by the DNN architecture shows that these deep learning models are suitable for the accurate estimation of the speech presence without needing any assumptions about the spectral properties of the speech and noise signals. Moreover, the resulting a posteriori SPP (Fig. 5.2.e) improves the previous estimation by using statistical spatial models, which helps to discriminate more clearly between speech presence and absence regions. This results in more accurate SPP masks, as shown in Fig. 5.2.e. Finally, it is observed that the Kalman gain (Fig. 5.2.f) is not as sparse as the a posteriori SPP, while it still presents higher values in high SNR bins. In addition, this gain suffers a smooth decay when speech is absent, avoiding the signal distortion introduced by the SPP sparsity. The same behavior is observed for the Wiener gain in the REMWF variant. The explanation is that these gains depend on the speech variance under speech presence, whose estimation is addressed using a different methodology than that used in the M-step. The followed procedure avoids the dependence on the SPP masks. This also supports the decoupling achieved between the postfiltering step and the SPP masking, which allows for a better enhanced signal as well as an accurate MLE estimation of the model parameters.

5.3.2 Performance analysis using oracle estimates

The upper-bound performance of our proposed framework can be analyzed using oracle estimates for both the a posteriori SPP and the acoustic parameters. Fig. 5.3 shows the results, in terms of PESQ and SDR scores, obtained using the REMKF approach in comparison with three variations involving oracle estimates:

- The MWF approach using the oracle IBM masks (IBM-MWF).



(a) PESQ results

(b) SDR results

Fig. 5.3 Results for the evaluation using oracle estimates for the SPP and the acoustic parameters. The improvement with respect to the noisy speech results is showed along with the 95% confidence intervals.

- The REMKF approach using oracle IBM masks for the a priori SPP instead of the DNN-based estimation (IBM-RKF).
- The REMKF approach using oracle estimates for the RTF and the noise SCM (O-RTF), but with the DNN-based a priori SPP estimator.

The goal of this last variant is to clearly distinguish between the contribution of accurate acoustic parameters and good a priori SPP estimates to the performance. Oracle noise estimates were obtained using Eq. (5.56), but knowing the noise signal instead of the noisy speech signal in the recursion. On the other hand, the oracle RTF was derived from Eq. (5.52), where correlations were obtained directly from the clean speech signal (with no need of clean speech and a posteriori SPP estimates). We also considered an additional variant for the REMKF framework, named C-RKF. This variant uses the CDR-based a priori SAP estimator proposed in [209] for the a priori SPP estimation. The idea was to compare the performance of the DNN estimates with respect to that achieved using a classical signal processing approach based on assumptions about the spatial properties of the signals.

The results show that the REMKF approach clearly outperforms the C-RKF approach, which remarks that the DNN leads to a better SPP initialization. This increases the final performance of the REM framework, which takes advantage of a good initialization for the a priori SPP. In the case of the oracle estimators, the IBM-RKF approach performs better than the IBM-MWF approach, especially in terms of PESQ results. This suggests that the robustness of the REM framework is not only due to the availability of accurate SPP estimates but also to the EM iterative procedure followed. Our approaches allows for a more robust computation of the model parameters, thus increasing the performance in non-stationary environments. On the other hand, the O-RKF approach outperforms the rest of the oracle estimators. This shows that a good estimation of the acoustic parameters, in practice, has a larger contribution to the performance than the use of oracle SPP estimates. Moreover, the final performance benefits for the combination of the DNN-based SPP estimates and the use of statistical spatial models derived from the acoustic model parameters.

5.3.3 Performance analysis of the SPP estimators

In this subsection, we further explore the contribution, to the SPP estimation, of DNN estimates integrated with statistical spacial models. In order to compare the DNN-based a priori SPP estimates and the a posteriori SPP estimates obtained by our framework, we considered a binary detector as [209]

$$\widehat{\mathscr{M}_x}(t,f) = \begin{cases} 1 & \text{if } p_x(t,f) > p_{\text{thr}}, \\ 0 & \text{otherwise}, \end{cases} \tag{5.69}$$

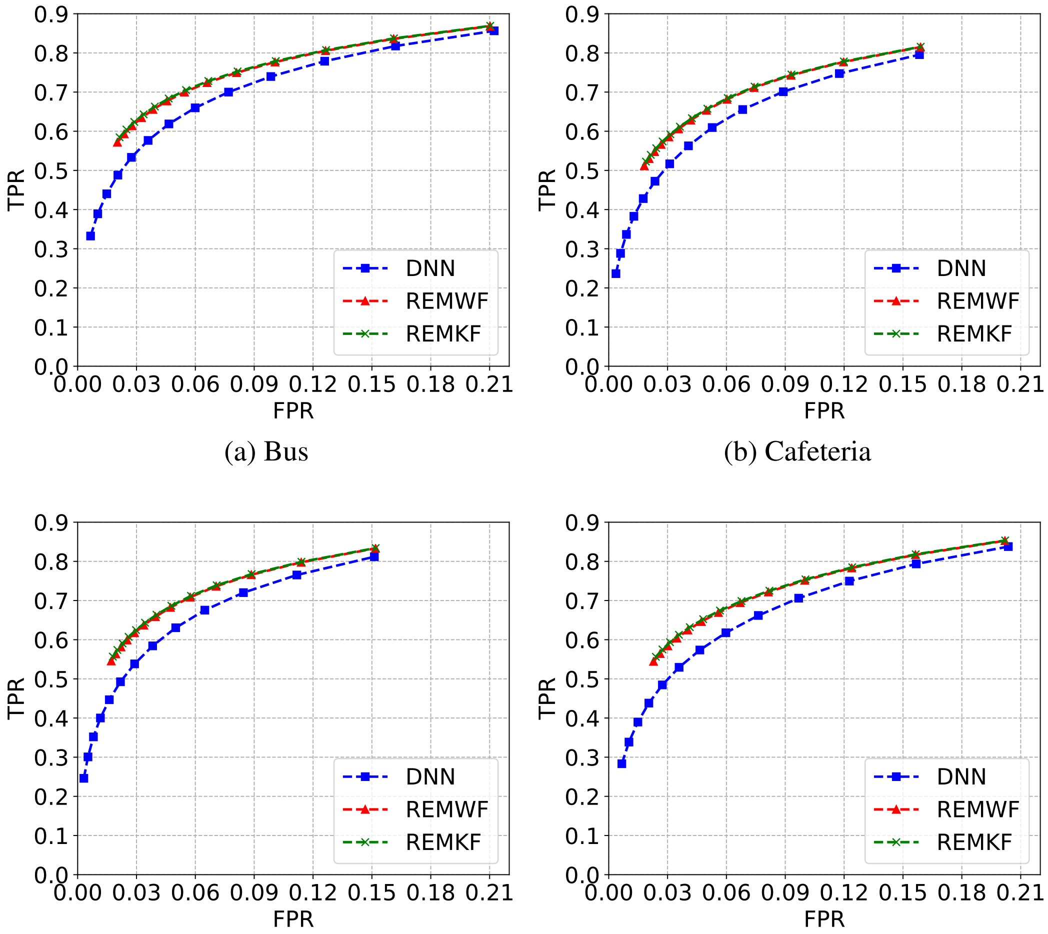
where p_{thr} is a selected threshold. In the case of the DNN estimate, we used $q_x(t,f)$ instead of $p_x(t,f)$. We also defined a ground truth detector, $\mathscr{M}_x(t,f) = 1$ when speech is present ($\mathscr{H}_x$) and zero otherwise. This ideal detector was chosen to be equal to the IBM masks used to train the DNN. Then, we defined the true positive rate (TPR) and the false positive rate (FPR) of the detector for a given utterance as

$$\text{TPR} = \frac{\sum_{t,f}\left[\left(\widehat{\mathscr{M}_x}(t,f) = 1\right) \& \left(\mathscr{M}_x(t,f) = 1\right)\right]}{\sum_{t,f}\left[\mathscr{M}_x(t,f) = 1\right]} \tag{5.70}$$

$$\text{FPR} = \frac{\sum_{t,f}\left[\left(\widehat{\mathscr{M}_x}(t,f) = 1\right) \& \left(\mathscr{M}_x(t,f) = 0\right)\right]}{\sum_{t,f}\left[\mathscr{M}_x(t,f) = 0\right]} \tag{5.71}$$

We evaluated the performance of the binary detector by means of the Receiver Operating Characteristics (ROC) curve [47], which is a representation of TPR vs. FPR for different threshold values. The higher the area under the curve, the best the detector performance.

Fig. 5.4 shows the ROC curves obtained using the DNN estimates and the a posteriori SPP estimates provided by the REMWF and REMKF algorithms in the different noisy environments. We used the values $p_{\text{thr}} \in [0.2, 0.8]$ with 0.05 step to focus on the regions of the curve where the differences are more noticeable. The results were obtained using the evaluation set. As can be observed, the proposed approaches improve the ROC curves with respect to the unprocessed DNN estimates. This shows that the use of spatial statistical



(a) Bus

(b) Cafeteria

(c) Pedestrian street

(d) Street

Fig. 5.4 ROC curves of detectors obtained using the DNN output (a priori SPP) and the a posteriori SPP estimates of the REMWF and REMKF algorithms. The values of the threshold p_{thr} are chosen between 0.2 and 0.8 with a step of 0.05.

models helps to better discriminate between speech presence and speech absence regions in the spectrogram, as both spectral and spatial information are used in the decision. On the other hand, the performance improvements of both REMWF and REMKF approaches are comparable across the different noisy environments.

5.3.4 Analysis of the computational latency and complexity

This subsection is devoted to the performance evaluation of our REMWF and REMKF approaches in terms of the computational complexity and latency. As a complexity measure, we used the number of EM iterations needed for a good performance. Thus, Fig. 5.5.a) shows the improvements achieved in terms of SDR with respect to the noisy speech signal

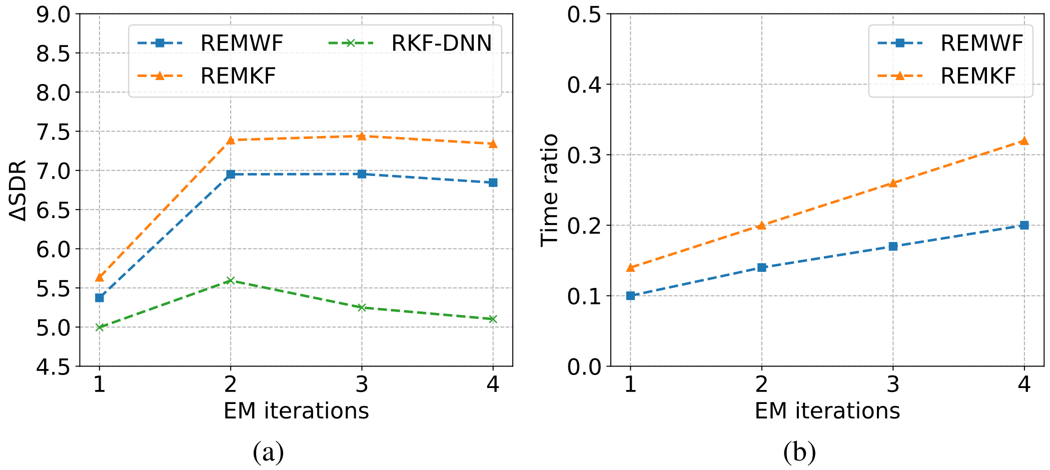


Fig. 5.5 Performance evaluation of our REMWF and REMKF approaches in terms of the number of EM iterations: (a) Improvements on SDR metric, (b) Time needed to process each second of the noisy speech signal.

for a different number of iterations. As can be observed, there is no improvement in the performance after two or more iterations. This fast convergence of the REM algorithm has been already reported in other works [194, 209]. It can be explained by the fact that the a posteriori SPP estimation, in the first iteration, uses the model parameters from previous frames. Then, in the second and following iterations, the model parameters are updated using the information provided by the current frame, which allows for more accurate a posteriori SPP estimates. This is the reason why we chose two EM iterations for the proposal.

We also analyzed the performance of the REMKF algorithm, in terms of the number of EM iterations, when the DNN estimates are directly used for the a posteriori SPP (RKF-DNN). In this case, there is no computation of the a posteriori SPP through (5.33). The objective was to check that the REM framework contributes to improve the SPP estimates obtained from the DNN. It can be observed that the SPP estimation of the REMKF outperforms the RKF-DNN approach. Moreover, the RKF-DNN approach does not show significant improvements after the first iteration as in the case of the REMKF method. This highlights the fact that both frameworks, REM and DNN, can be successfully integrated and help each other to improve the final enhancement.

In addition, we also evaluated the computational latency of our algorithm in terms of number of iterations. To this end, the REMWF and REMKF approaches were evaluated over 220 files from the evaluation set. The ratio between the total time needed to process the file and its duration was computed. The implementations were run on an Intel Core i7-4790 CPU at 3.6 GHz with four cores, 16 GB of RAM, and an Nvidia GeForce GTX 1060 GPU with 6 GB of memory. The GPU was only used for DNN inference of the a priori SPP, while the rest of the algorithm was run on CPU. The obtained average ratios are depicted in Fig. 5.5.b), which shows that the computational time increases almost linearly with the number of iterations. Moreover, it is observed that both algorithms can be executed faster than real-time on a computer with similar settings.

5.4 Summary

In this chapter, we have proposed a recursive EM framework for online multichannel speech enhancement which integrates a DNN-based speech presence estimator. The REM framework estimates the clean speech signal and the a posteriori SPP during the E-step, while the acoustic parameters are computed during the M-step. The procedure is recursively repeated frame by frame, allowing for the online processing of the noisy speech signal. The proposal was formulated using a statistical framework intended for multichannel noisy speech signals and under the narrowband assumption. The speech presence probability at each time-frequency bin is also considered. Moreover, a temporal linear prediction model for the clean speech amplitudes was defined, which accounts for the correlation across frames.

The proposed REM framework was defined from a statistical multichannel model formulated using an exponentially-weighted log-likelihood. In the E-step, the first- and second-order expectations of the clean speech signal are first estimated using an MVDR beamformer followed by a linear postfiltering. Two postfilters have been considered, a Wiener filter and a Kalman filter, the latter being able to exploit the temporal correlations of the clean speech amplitudes by means of an LPC prediction model. An SPP masking is also used during the estimation of the clean speech statistics. This masking is only considered in the M-step to avoid distortions on the output signal. In addition, the a posteriori SPP is estimated using an a priori SPP and the likelihoods of the noisy speech and noise signals. On the other hand, the MLE estimation is applied during the M-step to compute the RTF, the noise SCM and the LPC parameters of the KF postfilter. The clean speech variance can be obtained from the MVDR output by a procedure which allows for a quick adaptation and avoids the distortion introduced by SPP. Moreover, the a priori SPP is estimated using a DNN-based mask estimator. This DNN model exploits the spectral properties of the noisy speech signals and provides a good initialization for SPP in the REM framework. The REM framework could suffer from some convergence issues, so that some recommendations have been proposed and described for its correct performance.

The chapter conclude with an experimental evaluation where different versions of the proposed REM framework are evaluated using a noisy multichannel database recorded with a six microphone tablet. The evaluation through objective metrics showed that the proposal outperforms other state-of-the-art approaches as the MWF and the MKF, which also use the same DNN mask estimator for a fair comparison. The REMKF also achieved better performance than the REMWF version, which demonstrated the benefits of exploiting the temporal correlations along with the spectral and spatial properties of the signals. The performance was further analyzed by considering oracle estimates for the a posteriori SPP and the acoustic parameters. Moreover, the DNN mask estimator was compared with classical signal processing estimators, showing better accuracy and performance for the REM framework. The integration of the DNN estimates with statistical spatial models also led to more discriminative a posteriori SPP predictions, thus increasing the overall performance of the REM approach. Finally, the analysis of the number of EM iterations showed that two of them are enough to achieve competitive performance. Moreover, the low computational latency allows for real-time processing in an online fashion, which makes the REM framework suitable for real-world applications in mobile devices.

Chapter 6

Multichannel speech target extraction based on spatial-beam network

The previous chapters were devoted to our online multichannel speech enhancement contributions used in noisy environments. The main source of distortion in these scenarios is the environmental noise, which may be a very challenging interference to address due to its non-stationary properties. Nevertheless, these noise signals are commonly uncorrelated with the clean speech signal and they can be differentiated from their different spectral patterns. We will now focus on a different scenario of particular interest: multichannel automatic speech recognition (ASR) for conference scenarios with multiple speakers. In this case, an array microphone is used to record the target speaker, who has to be transcribed to get an accurate summary, at each moment. Even though current ASR systems allow for high recognition accuracy, mainly due to the rise of DNN algorithms, these systems suffer in this scenario from two issues. The first issue is the need for low-latency processing in the case of real-time transcriptions. The second and most important issue is the presence of different types of distortion, especially overlapping speakers. Therefore, speech enhancement front-ends are needed to ensure good ASR performance.

Beamforming techniques are the common front-end for ASR systems when array microphones are used to record the noisy speech signals. Most of the state-of-the-art front-ends in these conditions are based on the estimation of time-frequency masks to discriminate between the target speech and noise sources. These masks are used to estimate the acoustic parameters needed for the computation of the beamformer weights. The use of DNN mask estimators is a state-of-the-art approach to deal with this task [76, 78]. Also, these estimators can be designed to allow online processing [81, 256, 72]. The main problem of these DNN models is that their performance degrades in the presence of multiple overlapped speakers, as the mask estimator is unable to discriminate between them due to their similar spectral patterns. Several strategies have been developed in recent years to deal with this scenario, referred to as the source separation problem. Among the most promising approaches, we can highlight three techniques: deep clustering [75], deep attractor networks [22], and permutation invariant training [246]. These techniques are usually applied when we are interested in separate the different speech sources that are mixed in the utterance. However, we are only interested in a single target speaker, so the use of these techniques is not recommended due to their complexity. Another class of techniques directly focuses on the estimation of the masks for a target speaker by using contextual information [254, 226, 227]. For example, the speaker-beam (SpkB) approach [255] proposed a novel network architecture that uses an auxiliary utterance of the target speaker to allow the DNN model to focus on the spectral characteristics of the desired speaker. The problem of the SpkB approach is that its performance degrades in the case of overlapping speakers of the same gender, especially if these speakers are unseen during the training phase. On the other hand, other works are based on the training of a mask estimator to focus on the speaker that uses one of a set of predefined utterances [103, 191]. In [23] a DNN is trained to focus on the direction of the target speaker, which is provided by using oracle information. Finally, the approaches in [32, 228] use the adaptation utterances directly in the ASR back-end to allow target speaker separation without the need for a speech enhancement front-end.

In this chapter, we propose a novel online multichannel target speaker extraction algorithm called spatial-beam, based on the SpkB approach. In our approach, the spatial information provided by a multichannel adaptation utterance from the target speaker is also considered along with the spectral information to force the DNN mask estimator to focus on the target speaker. Two different alternatives are evaluated for our proposal: a spatial pre-processing of the signals before their use in the DNN mask estimator, and the use of additional spatial features for the network. The adaptation utterance does not need to be fixed. Besides, the only assumptions made are that the adaptation utterances do not contain other interfering speakers and there are only slight variations in the position of the target speaker between the adaptation utterance recording and the noisy speech signal to be processed. The advantage of the presented approach is that it does not depend on specific noise conditions during the adaptation or the use of other additional oracle information. Moreover, our approach can exploit the spectral and spatial properties of the target speaker signal, yielding better discriminative performance even in the same gender cases.

The remainder of this chapter is structured as follows. First, the proposed block-online beamformer procedure based on the use of target speaker and noise masks is presented in Section 6.1. Section 6.2 introduces the SpkB mask estimator architecture, and then the spatial-beam approach is described along with its two variants: the spatial pre-processing method and the spatial features method. Finally, the proposal is evaluated in Section 6.3 using objective quality metrics, and its performance is also measured when it is used as a front-end of an ASR system.

6.1 Block-online estimation of beamformer parameters

We first assume the next multichannel additive-distortion model for the noisy speech signal in the STFT domain,

$$\mathbf{y}(t,f) = \mathbf{x}(t,f) + \mathbf{n}(t,f) \tag{6.1}$$

where $\mathbf{x}(t,f)$ and $\mathbf{n}(t,f)$ are, respectively, the multichannel target speech and noise signals. This noise signal mainly accounts for the presence of interfering speakers, although it also contains other distortions as late reverberations and background noise.

The noisy speech signal is processed using a beamforming algorithm to estimate the target clean speech signal. To obtain the beamformer weights, we need an estimation of the clean speech and interfering noise SCMs at each time-frequency bin, namely $\Sigma_X(t,f)$ and $\Sigma_N(t,f)$ respectively. Assuming that the statistics of the signals are slowly-variant, we propose low-latency processing of the noisy speech signal by a recursive estimation of these matrices in blocks of L frames. Therefore, the matrices for the n-th block of frames, which comprises the time frames $t_n \in [(n-1)L, nL-1]$, are obtained as

$$\Sigma_\nu(t_n,f) = \beta_\nu \Sigma_\nu(t_{n-1},f) + (1-\beta_\nu)\Phi_\nu(t_n,f), \tag{6.2}$$

where $\nu = \{X,N\}$ indicates target speech or noise, β_ν is a forgetting factor, and

$$\Phi_\nu(t_n,f) = \sum_{t=(n-1)L}^{nL-1} M_\nu(t,f)\mathbf{y}(t,f)\mathbf{y}^H(t,f). \tag{6.3}$$

This procedure depends on the estimation of target speech and noise dominant time-frequency masks, $M_X(t,f)$ and $M_N(t,f)$ respectively. These masks helps to discriminate between speech presence and noise-only bins to compute the spatial statistics. Thus, the masks have to be estimated in advance, using, for example, a DNN mask estimator. The computation of these speech and noise masks will be presented in the next section.

The above procedure (Eq. (6.2)) also needs an initialization of the target speech and noise SCMs. This initialization can help to achieve a fast convergence for beamforming and a performance improvement in the first frames. A typical choice when there is no knowledge about the signals is to initialize with an identity matrix for the noise SCM and a zero matrix for the target speech SCM. Nevertheless, better performance can be obtained with proper initialization of these matrices, making use of prior information about the spatial characteristics of the signals. Thus, we can assume a diffuse noise field as initialization for the noise SCM as

$$\Sigma_N(t_0, f) = \sigma^2_{N_{\text{init}}}(f)\Gamma_{\text{diff}}(f), \tag{6.4}$$

where $\Gamma_{\text{diff}}(f)$ is the coherence matrix of the diffuse noise field, as defined in (2.52), and $\sigma^2_{N_{\text{init}}}(f)$ is an estimation of the noise PSD obtained from the first block of frames. On the other hand, for the speech SCM, we assume the availability of a noise-free multichannel utterance from the target speaker. The only condition is that the speaker position does not change between the recording of the noisy speech and that of the auxiliary utterance. This auxiliary signal is called adaptation utterance, whose STFT signal is $\mathbf{s}_A(t, f)$. The target speech SCM is then initialized as $\Sigma_X(t_0, f) = \Sigma_{S_A}(f)$, where $\Sigma_{S_A}(f)$ is an offline estimation of the adaptation utterance SCM, computed as

$$\Sigma_{S_A}(f) = \frac{1}{T_A}\sum_t \mathbf{s}_A(t, f)\mathbf{s}_A^H(t, f), \tag{6.5}$$

where T_A is the number of frames of $\mathbf{s}_A(t, f)$. This way, we can exploit the spatial information and the characteristics of the acoustic channels between the speaker and the array microphones.

Finally, the beamformer weights are computed using the estimated SCMs for every block step. We use the version of the MVDR beamformer in (2.46) and a rank-1 approximation for the target speech SCM [232], defined as

$$\widetilde{\Sigma}_X(t_n, f) = \widetilde{\mathbf{h}}(t_n, f)\widetilde{\mathbf{h}}^H(t_n, f) \cdot \frac{\text{tr}\{\Sigma_X(t_n, f)\}}{\text{tr}\{\widetilde{\mathbf{h}}(t_n, f)\widetilde{\mathbf{h}}^H(t_n, f)\}}, \tag{6.6}$$

where

$$\widetilde{\mathbf{h}}(t_n, f) = \Sigma_N(t_n, f)\mathscr{P}\left\{\Sigma_N^{-1}(t_n, f)\Sigma_X(t_n, f)\right\} \tag{6.7}$$

is an approximation of the acoustic transfer function vector in the rank-1 model, with $\mathscr{P}\{\cdot\}$ standing for the principal component of a matrix.

6.2 Spatial-beam target speaker estimation algorithm

In this section, we describe the mask estimation procedure based on the proposed spatial-beam approach to deal with interfering speakers. This approach combines the speaker-beam (SpkB) method with the use of spatial information to better discriminate the target speaker.

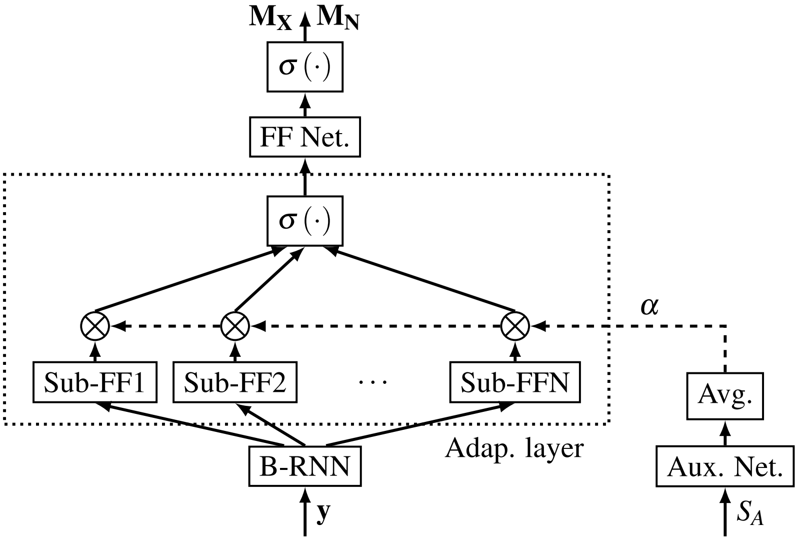


Fig. 6.1 Block diagram of the speaker-beam (SpkB) mask estimator. This model includes a bi-directional RNN (B-RNN), feed-forward (FF) layers and sub-layers, non-linearities $\sigma(\cdot)$ and the auxiliary network.

Moreover, our approach adapts the mask estimator to the online scenario. We first describe the SpkB approach and then we explain how we modify this model in our proposal.

6.2.1 Basis of speaker-beam approach

The SpkB approach was proposed in [255] for mask estimation in scenarios where we are only interested in a target speaker among multiple speakers. The SpkB approach assumes the availability of an adaptation utterance $\mathbf{s}_A(t,f)$ from the target speaker. This adaptation utterance only contains speech from the target speaker, so its spectral information can be exploited to better discriminate among speakers.

The DNN mask estimator for the SpkB approach is depicted in Fig. 6.1. This model is based on the mask estimator proposed in [76]. The DNN estimator of that work includes a bi-directional RNN, several feed-forward layers, and an output layer that provides the masks $M_X(t,f)$ and $M_N(t,f)$. The masks obtained for the individual microphone channels are combined using a median operation. The aforementioned mask estimator is modified to incorporate the spectral information from the target speaker provided by the adaptation utterance. This adaptation is done as follows. First, one of the feed-forward layers is split into several sub-layers. The output of these sub-layers is combined through a weighting vector α, called speaker representation, which provides the spectral speaker information. The speaker representation vector helps the mask estimator to focus on the target speaker, while the rest of the interfering signals are treated as noise. Vector α is obtained from a single-channel adaptation utterance $S_A(t,f)$ by using an auxiliary network, which is fed with the adaptation utterance. The output of the auxiliary network contains the spectral information for the different time frames, which is finally averaged on time to obtain the speaker representation vector. The SpkB mask estimator and the auxiliary network are jointly trained using pairs of multichannel noisy speech and single-channel adaptation utterances. Thus, the estimator learns to exploit the speaker's characteristics.

The SpkB approach exhibits some limitations for its use in a practical online processing task. First, the use of a bi-directional RNNs is not allowed in an online scenario. Moreover, the approach exhibits a performance degradation when applied in scenarios with overlapping speakers with similar spectral characteristics, as in the case of speakers of the same gender. In this case, the mask estimator is not able to discriminate among the target speaker and the interfering ones. Although this problem can be alleviated if both speakers are part of the training data, this is not the general case.

6.2.2 Improvements using spatial information and online processing

Our proposed spatial-beam approach is intended to improve the performance of the SpkB method by solving the aforementioned problems. First, the mask estimator is adapted to online processing by replacing the bi-directional LSTM layer with a single LSTM layer of twice the output size, thus reducing the network information to current and past frames. The input features are the log magnitude spectrum. In addition, the utterance mean and variance offline normalizations are replaced by a recursive mean normalization [72]. The offline normalization is still applied for the adaptation utterance. Moreover, we propose the use of additional spatial information obtained directly from the adaptation utterance. Thus, it is assumed that a multichannel adaptation utterance $\mathbf{s}_A(t,f)$ is available and that the target speaker position has slightly changed between the noisy speech and adaptation utterances. This ensures that the spatial characteristics of the target speaker are similar in both cases. The use of spatial information allows better separation between speakers with similar active frequencies and speech patterns, but different positions. Therefore, we propose two different approaches that exploit the spatial information provided by additional algorithm steps.

Spatial pre-processing: The diagram of this approach, called PreBF, is depicted in Fig. 6.2. PreBF uses a pre-processing beamforming step for the multichannel noisy speech and adaptation utterances before its use in the SpkB block. An offline MVDR beamformer is used as spatial pre-processor. The beamformer weights for this initial beamformer are computed by using the same SCMs used as initialization of the block-online beamforming.

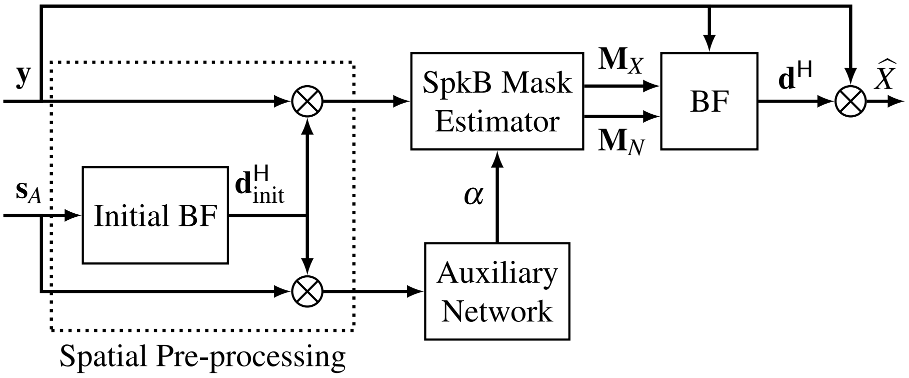


Fig. 6.2 Block diagram of the spatial pre-processing alternative for the spatial-beam approach. The initial beamformer (BF) process the multichannel signals before its use in the SpkB block.

Thus, the weights are obtained as

$$\mathbf{d}_{\text{init}}(f) = \frac{\Sigma_{\text{diff}}^{-1}(f)\widetilde{\Sigma}_{S_A}(f)}{\text{tr}\{\Sigma_{\text{diff}}^{-1}(f)\widetilde{\Sigma}_{S_A}(f)\}}\mathbf{u}_1, \tag{6.8}$$

where Σ_{diff} is the diffuse noise SCM, $\widetilde{\Sigma}_{S_A}$ is the rank-1 approximation of the adaptation utterance SCM, and $\mathbf{u}_1$ is a unit vector pointing to the reference microphone. After the spatial pre-processing stage, the two resulting single-channel signals feed both the SpkB mask estimator and the auxiliary networks, as shown in Fig. 6.2, thus allowing the computation of masks $M_X(t,f)$ and $M_N(t,f)$. The estimated masks are finally employed to obtain final beamformer weights. The spatial pre-processor can exploit the spatial information extracted from the adaptation utterance using its spatial statistics. Although the pre-processing step does not provide an accurate speaker separation on its own (the initial beamformer does not consider the actual statistics of the interfering signals), the overlapping speakers are attenuated so that a more accurate mask estimation is possible.

Spatial features: This alternative is based on the use of additional spatial features at the input of the auxiliary and mask estimation networks. If the spectral magnitude of the target speaker is not discriminative enough information, the phase differences between the microphone signals may still provide useful spatial information to identify the target speaker. Therefore, the auxiliary network can obtain discriminative embeddings for speakers with similar spectral properties by using this additional spatial information. Similarly, speakers with similar positions with respect to the microphone array can be still separated using spectral information.

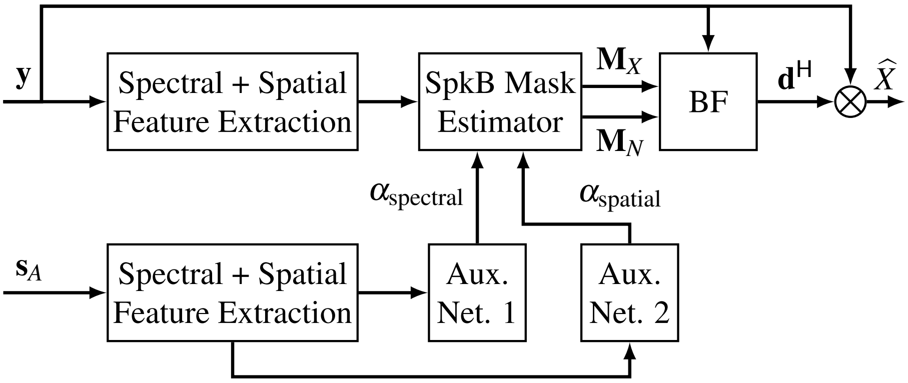

y
Spectral + Spatial Feature Extraction
SpkB Mask Estimator
$\mathbf{M}_X$
$\mathbf{M}_N$
BF
$\mathbf{d}^{\mathsf{H}}$
$\widehat{X}$
α_{spectral}
α_{spatial}
$\mathbf{s}_A$
Spectral + Spatial Feature Extraction
Aux. Net. 1
Aux. Net. 2

Fig. 6.3 Block diagram of the spatial features alternative for the spatial-beam approach.

The block diagram of this approach is depicted in Fig. 6.3. First, both spectral and spatial features are extracted from the multichannel signals. As spatial features, we use the Interchannel Phase Difference (IPD) features, which have shown to achieve good performance in source separation tasks [233]. These IPD features are computed similarly to [233] as

$$\mathrm{cIPD}_{i,j}(t,f) = \cos\left(\theta_i(t,f) - \theta_j(t,f)\right), \tag{6.9}$$

$$\mathrm{sIPD}_{i,j}(t,f) = \sin\left(\theta_i(t,f) - \theta_j(t,f)\right), \tag{6.10}$$

where i and j are microphone channel indices, and $\theta_i(t,f)$ represents the single-channel phase component of either the noisy speech signal or the adaptation utterance. The above procedure involves the computation of these features for a two-channel problem. In the case of more than two channels, each pair of channels is treated like a two-channel problem.

In order to feed the SpkB mask estimator, the spectral and spatial features obtained from the multichannel noisy speech signal are concatenated. In addition, a bottleneck feed-forward layer is introduced at the input of the SpkB mask estimator to reduce the size of the LSTM input. On the other hand, to ensure that the SpkB network uses both the available spatial and spectral information, two independent auxiliary networks are trained. These networks use either spectral or spatial features obtained from the adaptation utterance. The mean pooling at the output of each auxiliary network is now carried out in both time and channel dimensions. Finally, the estimated speaker representation vector for the spectral properties α_{spectral} is used to weight half of the sub-layers of the adaptation layer, while the speaker representation vector for the spatial properties α_{spatial} is used to weight the other half.

6.3 Experimental results

We evaluated the proposed different variants of the spatial-beam approach (PreBF and spatial features) and compared them with the speaker-beam approach and deep attractor networks (DAN). To this end, we used SDR to measure the performance of the enhancement procedure, STOI for speech intelligibility, and WER to test its performance in a conference scenario for ASR. The evaluation was performed on the SMS-WSJ simulated multichannel database, which considers mixtures of two concurring speakers along with reverberation and microphone noises. The adaptation utterances were obtained in the same room acoustics and target speaker positions than the noisy mixture, using another clean speech utterances from the same speaker. Reverberation and microphone noise are still present in these adaptation utterances, but there are not interfering speakers. Moreover, the speaker position can be assumed approximately fixed along every utterance. Thus, offline beamforming can be considered the best solution if low latency is not an issue.

For the STFT computation, a 512-point DFT was used with a Hann window and a 75% overlap, resulting in 257 frequency bins for each time frame. The SpkB mask estimator consisted of a single LSTM layer of 1024 units, an adaptation layer with 30 feed-forward sub-layers and 1024 units each sub-layer, a feed-forward layer with 1024 units, and one output layer. On the other hand, the auxiliary network had two feed-forward layers of 50 units each and an output layer of 30 units, as in [255]. The architecture was trained using the training set of the SMS-WSJ database. As loss function, we chose the binary cross-entropy between the estimated masks and ideal binary masks calculated from the reverberated clean speech signals. For the block-online beamforming, we chose blocks of five frames and a forgetting factor of $\beta_v = 0.95$.

The acoustic model of the ASR back-end was a Wide Residual Network as proposed in [77]. This ASR back-end uses logarithmic Mel filterbanks as input features and it mainly consists of two bi-directional LSTM layers. All hyper-parameters were taken directly from [77]. The acoustic model was combined with a trigram language model from the WSJ baseline provided by the KALDI toolkit [170]. The back-end was trained on the artificially reverberated WSJ utterances without overlapped speech. The decoding was performed without language model rescoring. Although the proposed back-end operates offline for all experiments, as we focus on the front-end processing, it may be replaced by an online version to obtain a fully online system.

Table 6.1 SDR, STOI and WER scores obtained for different initialization of the SCM estimation using ideal binary masks.

Method	**Initialization**		**STOI**	**SDR**	**WER**
	Σ_X	Σ_N		dB	%
Offline	–	–	0.84	12.37	16.40
Online	Zeros	Identity	0.82	10.95	19.89
		Diffuse	0.82	**11.13**	19.60
	Σ_{S_A}	Identity	0.82	10.69	17.88
		Diffuse	**0.83**	11.10	**16.94**

6.3.1 Evaluation of the initialization methods

First, we evaluated the performance of the different proposed SCM initialization strategies (required in Eq. (6.2)) for the block-online beamforming using IBM masks. The results are shown in Table 6.1, where the offline method and the different initializations for the SCMs are compared in terms of STOI, SDR, and WER. For the target speaker SCM, we evaluated both the zero matrix and the use of the adaptation utterance SCM. On the other hand, the noise SCM was either initialized with the identity matrix or the diffuse noise SCM. It is observed that the best initialization performance is achieved for the combination of diffuse noise SCM and adaptation utterance SCM. Moreover, this combination is close to the offline beamformer in recognition accuracy, and it also obtains competitive results in distortion reduction and intelligibility. Therefore, these results show that proper initialization is helpful for beamformer convergence.

6.3.2 Evaluation of the target speaker extraction algorithms

In this subsection, we evaluated the performance of the different mask estimators for target speaker extraction. First, Table 6.2 compares the different DNN mask estimator approaches for offline and online beamforming. In the case of offline beamforming, the spatial-beam approaches used bi-directional RNNs for a fair comparison with the other approaches. The results show that both offline spatial-beam approaches achieve WER scores superior to the state-of-the-art approaches DAN and SpkB. Nevertheless, in terms of signal distortion, the use of spatial information does not lead to improvements in comparison with the DAN approach, although spatial-beam still performs better than the SpkB approach. Moreover, our proposals outperform the other methods in speech intelligibility gain. Thus, the proposed spatial-beam approach achieves competitive enhancement and recognition results, while it

Table 6.2 SDR, STOI and WER scores obtained for different speaker extractors.

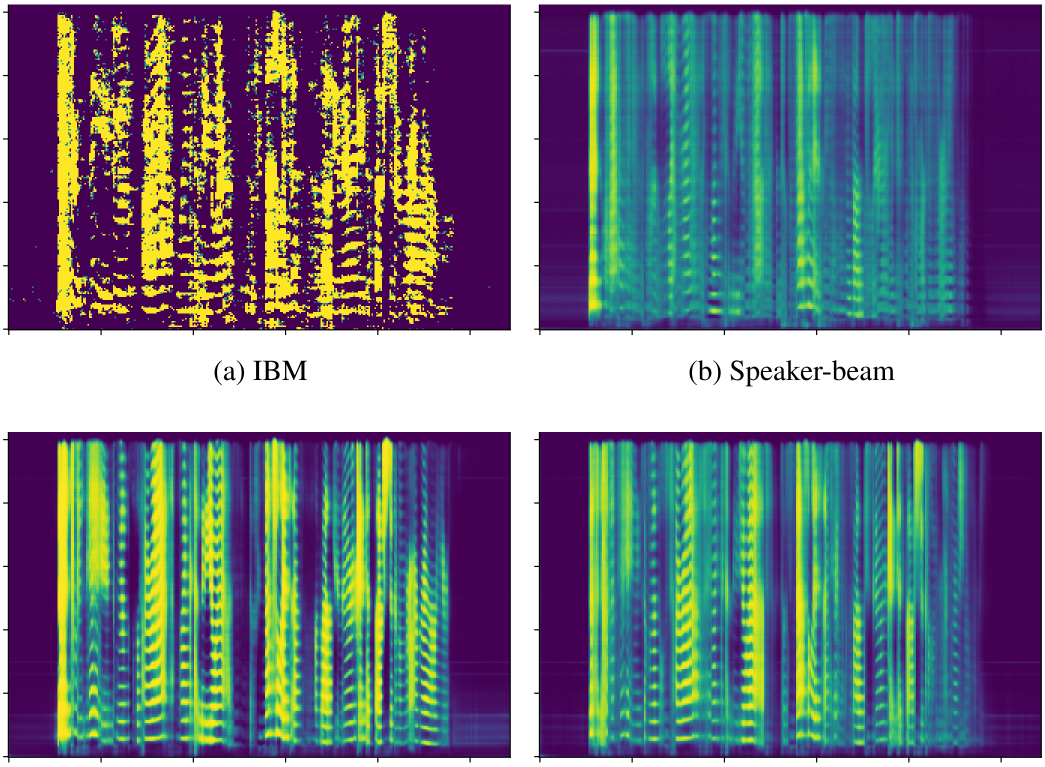
BF	Extractor	STOI	SDR dB	WER %
Offline	Speaker-beam	0.76	8.78	28.66
	DAN	0.78	**11.38**	23.70
	PreBF	**0.80**	10.00	**23.32**
	Spt. Features	**0.80**	9.70	23.50
Online	Online-PreBF	0.74	**5.54**	34.60
	Online-Spt. Features	**0.75**	5.09	**33.61**

Table 6.3 SDR and WER scores obtained for the different speaker extractors. Results are separated for overlapped speaker of the same and different gender.

Method	SDR (dB)		WER (%)	
	Differ.	**Same**	**Differ.**	**Same**
Speaker-beam	10.17	7.25	23.13	34.82
PreBF	10.68	**9.24**	21.21	**25.67**
Spt. Features	**10.92**	8.49	**19.49**	28.52

allows focusing on the estimation of the target speaker. We also tested the PreBF variant using only the spatial pre-processing but not the SpkB mask estimator, obtaining a WER score of 27.03 %. This result shows that the combination of the spatial pre-processing with the speaker information of the SpkB approach outperforms the individual systems, allowing for further interference speaker reduction and better recognition performance. Finally, we evaluated the online versions of the spatial-beam variants, using both an online mask estimator and block-online beamforming. The online approaches have a higher WER than their offline counterparts, mainly due to the block-online updating of the SCM matrices, but also because of the use of a single LSTM layer in the mask estimator. Nevertheless, the online approaches still achieve competitive results for online speech recognition. The use of spatial features stands as the preferred approach for ASR, while the PreBF variant performs better noise reduction.

One of the goals of our approach is to minimize the limitations of the SpkB method when dealing with overlapping speakers of the same gender. Therefore, we split the results into cases where overlapped speakers with different and same gender are found. Thus, we could evaluate how our strategies perform in each scenario. The results for speech distortion and recognition accuracy are shown in Table 6.3, where we compare the SpkB approach with our offline proposals. As can be observed, while the SpkB approach performs well in



(a) IBM

(b) Speaker-beam

(c) PreBF

(d) Online-PreBF

Fig. 6.4 Examples of different estimated target speaker masks when different speaker extractors are applied to the audio file 441c0403_445c0401_45 from the SMS-WSJ database.

the different gender case, its performance severely degrades in the case of speakers of the same gender, which increases the final WER. The use of spatial-beam with spatial features improves the accuracy of the estimator, but its performance is still limited in utterances with speakers of the same gender. This may be caused by the fact that the mask estimator network has difficulties to learn both the spectral and spatial characteristics for the separation task. On the other hand, the use of our PreBF variant is particularly effective in the same gender case, outperforming the spatial features variant in both SDR and WER for the same gender case. Moreover, both spatial-beam approaches achieve similar results in the different gender case. The improvement is especially noticeable in the recognition evaluation, where WER differences between the different and the same gender cases are reduced from 11.69 % to 4.46 %. The PreBF approach has the advantage that the input to the network is already processed, so the mask estimator can exploit the attenuated interfering speakers in the input signal to better discriminate between speakers.

We conclude this subsection with an example of the target speaker masks obtained by our proposed approach. Thus, Fig. 6.4 shows example target speaker masks in an utterance with two overlapped female speakers: the IBM oracle mask (Fig. 6.4.a), the Speaker-beam mask (Fig. 6.4.b), the spatial-beam mask for the PreBF approach in its offline version (Fig. 6.4.c), and, finally, the online version of the PreBF approach (Fig. 6.4.d).

It can be observed that the SpkB approach is not able to obtain an accurate target speaker mask in scenarios with speakers of the same gender. The reason is that it easily confuses the speakers' presence in the different time-frequency bins. This yields smoothed masks where there is not a clear distinction between speakers. On the other hand, the spatial-beam approach can successfully use spatial information to obtain discriminative masks. As a result, the resulting mask is closer to the oracle IBM mask. Moreover, the availability of the complete utterance information in the offline version allows for a more accurate separation. Finally, the mask obtained in the online version of the spatial-beam method does not show that clear distinction between the speakers, but is still able to achieve a good separation performance. Besides, the estimated mask outperforms the one obtained using the SpkB approach, which employs an offline mask estimator. This shows that the use of spatial information helps to separate speakers with similar spectral characteristics, as in the challenging case of same-gender speakers.

6.4 Summary

In this chapter, we have described our spatial-beam approach for the target speaker extraction in multichannel scenarios with overlapped speakers. This method is based on the speaker-beam approach, which uses the spectral information from an adaptation utterance of the target speaker to focus on its spectral characteristics. The spatial-beam approach uses additional spatial information to improve the discriminative performance of the DNN mask estimator. The method was evaluated for an ASR application in meeting scenarios with multiple speakers.

We first presented the block-online beamforming algorithm that estimates the signal from the target speaker. This method uses target speaker and noise masks, estimated by the DNN model, to update the speech and noise SCMs at each block of frames. To improve the convergence, we propose an initialization for both matrices. The initialization is based on a diffuse noise field assumption for the noise SCM and an estimate of the adaptation utterance SCM for the target speech matrix. The estimated matrices are then used to compute the weights of a rank-1 MVDR beamformer.

The speaker-beam mask estimator was then described, introducing first the recurrent network mask estimator and its adaptation to the SpkB approach. This adaptation includes the use of an intermediate adaptation layer with multiple hidden layers, combined by using a speaker representation embedding. This is obtained from the adaptation utterance by using an auxiliary network. After that, the spatial-beam proposal was presented, describing the adaptation to online processing and the use of spatial information from the adaptation utterance. Two different approximations were considered. The first approximation integrates a spatial pre-processing of the noisy speech signal and adaptation utterances using an offline MVDR beamformer. The second approximation uses additional spatial features computed from the noisy speech and adaptation signals.

The proposals were finally evaluated using a multichannel simulated database of overlapped speakers in reverberated rooms. We evaluated the different approaches using objective intelligibility and speech distortion metrics. In addition, we tested the performance of the proposal as a front-end for ASR. First, we evaluated the proposed SCM initializations for the block-online beamformer, showing improvements with respect to the baseline methods, especially in terms of WER scores. Then, we compared the two variants of our proposed spatial-beam approach with the SpkB approach and the DAN estimator. The evaluation was done using both offline beamforming and mask estimation. The results showed that our proposals achieve better recognition results, especially the spatial pre-processing variant. Next, we assessed our techniques for online processing, showing competitive ASR results for a meeting scenario, especially for the spatial features variant. Finally, we compared our spatial-beam approaches and SpkB in terms of the performance for different and same gender scenarios. The results highlighted that our approaches can effectively deal with the same gender case, outperforming the SpkB approach. This demonstrates that spatial information is useful to discriminate between speakers with similar spectral properties.

Chapter 7

Deep learning loss function for the perceptual evaluation of the speech quality

This chapter is dedicated to our contributions in single-channel DNN-based speech enhancement. As overviewed in Chapter 2, two main approaches have been studied in the STFT domain: spectral mapping, where the clean speech signal is directly estimated, and spectral masking, where the DNN estimates a gain function. Both approaches aim to improve the perceptual quality and intelligibility of the enhanced speech signal, outperforming classical approaches based on statistical signal processing. However, although the target of these methods is usually human listeners, the DNN architectures are frequently trained by using mean square error (MSE)-related criteria. This means that none or very weak perceptual criteria are considered during the training stage.

In recent years, several works have investigated the importance of the DNN training of psychoacoustic criteria based on human perception. Thus, the approach proposed in [190] introduced a constant penalty in the loss function against the removal of clean speech signal components. Moreover, the work in [70] proposed a joint DNN training and audible noise suppression framework. Another common strategy is the use of a frequency-dependent weighting in the MSE loss function [239, 71, 111, 123, 101, 252]. The idea is to account for perceptual features that depend on the frequency, as the absolute threshold of hearing, the auditory masking, or the perceptual relevance of each frequency band. Recently, a maximum likelihood approach for DNN training was proposed in [19, 18]. This method models the errors between enhanced and clean speech signals as Gaussian random variables. The approach yields a weighted MSE loss function where the DNN estimator and the statistical model are updated iteratively.

A more direct approach consists of the integration of well-established objective speech quality metrics as criteria to the loss function for the DNN training. Thus, the DNN is optimized using the same metrics that assess its performance. That is the case of the STOI metric, which has been evaluated as an independent loss function [107, 49, 251, 108] and combined with other losses [101]. Other objective metrics have been used as loss functions for speech enhancement and separation approaches, as the case of the ESTOI metric in [155] and the SI-SDR metric in [219]. The recent work in [109] analyzed and compared the performance of several metric-based loss functions for DNN-based monaural speech enhancement in the time-domain. Alternatively, other approaches proposed an indirect optimization of objective quality metrics via reinforcement learning [106], gradient approximation [249], or using a DNN model to learn the related loss function [50, 51].

In this chapter, we propose an adaptation of the PESQ algorithm, which is one of the best known objective metrics for speech quality evaluation, as a loss function for DNN-based speech enhancement methods. We call this approach Perceptual metric for the speech quality evaluation (PMSQE). To the best of our knowledge, this was the first proposal of a loss function based on the PESQ algorithm. To adapt the metric as a loss function, the loudness-based disturbance terms described in the PESQ standard are simplified and adapted for the gradient-based optimization. Moreover, these terms are computed on a per-frame basis from the clean and enhanced speech spectra. The proposal is evaluated for monaural speech enhancement using the most common approaches: spectral mapping and spectral masking. In addition, the PMSQE loss is evaluated when used in combination with other state-of-the-art loss functions.

The remainder of this chapter is structured as follows. The PMSQE loss function is first described in Section 7.1, where the different steps to compute the disturbance terms from the speech signals are indicated. Next, the use of the proposed loss function for DNN-based speech enhancement in its different variants (spectral mapping and spectral masking) is explained in Section 7.2. Finally, the experimental framework and the results obtained using the proposed loss in each of these DNN enhancement approaches are reported in Section 7.3 and 7.4, respectively.

7.1 Perceptual metric for the speech quality evaluation

The perceptual metric for the speech quality evaluation (PMSQE) is a perceptually-inspired loss function based on the well-known PESQ objective metric. This loss is then intended to take into consideration relevant perceptual effects, as loudness differences, perceptual masking, and threshold properties, during the training of a DNN architecture. Therefore, the idea

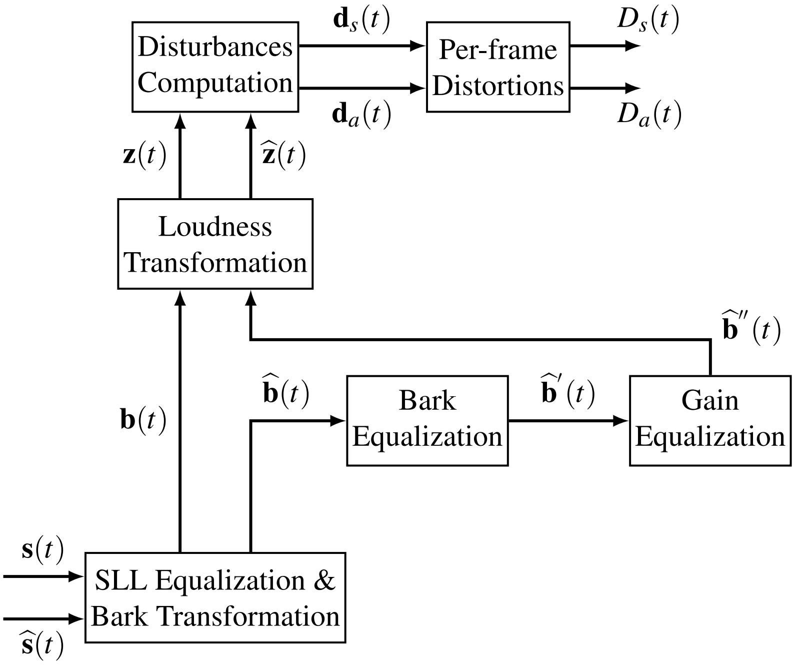
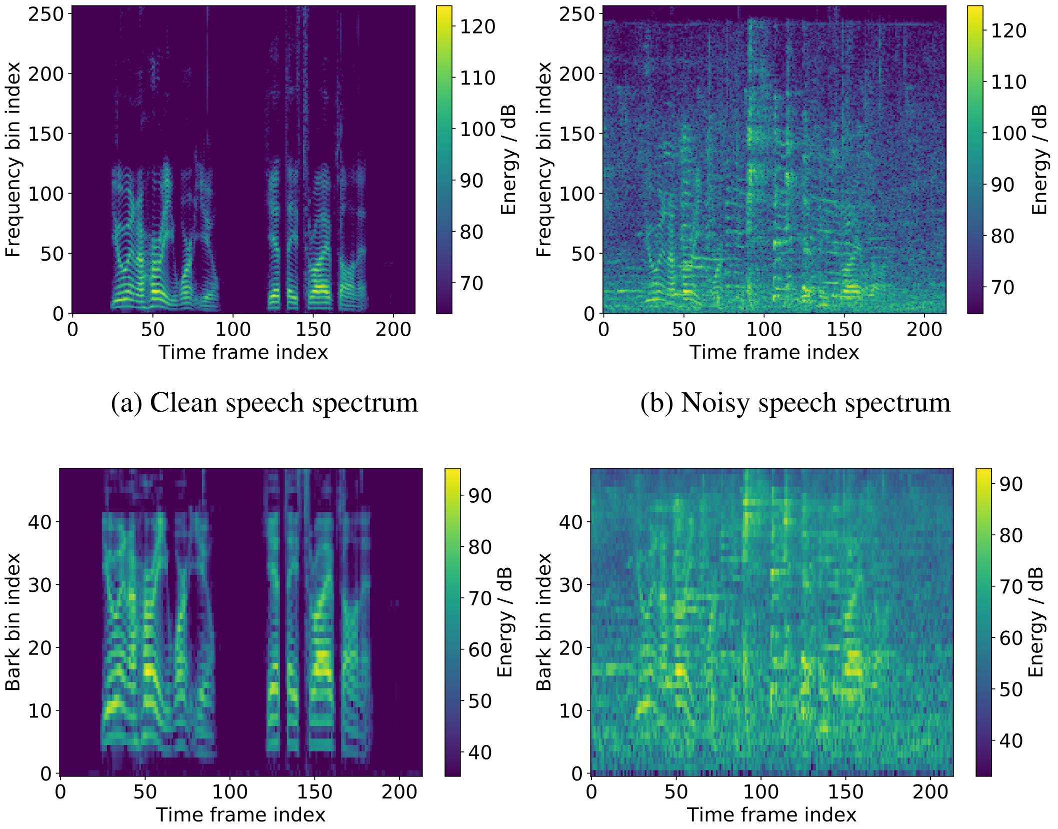


Fig. 7.1 Block diagram of the proposed PMSQE loss function, indicating the pipeline of the algorithm for the computation of the disturbance terms.

is to maximize the perceptual speech quality of the estimated clean speech signal obtained by the DNN model. Moreover, it is expected that this maximization yields improvements when evaluating with speech quality metrics, as PESQ.

The PMSQE loss consider the magnitude spectrum of the clean speech signal STFT, $|X(t,f)|$, and an estimation of this clean speech spectrum, $\left|\widehat{X}(t,f)\right|$. Then, inspired by the PESQ metric, two disturbance terms are estimated, thus modeling speech distortion in a perceptual domain as symmetrical and asymmetrical disturbances. The symmetrical disturbance, $D_s(t)$, considers the absolute difference between the estimated and true clean loudness spectra when auditory masking effects are accounted for. On the other hand, the asymmetrical disturbance, $D_a(t)$, is computed from the symmetrical disturbance but weighting the positive and negative loudness differences differently. This is because negative differences (omitted or attenuated spectral components) are perceived differently than positive ones (additive noise) owing to masking effects. A single value per disturbance term is obtained for each time frame. Thus, the PMSQE works in a frame-wise fashion.

Fig. 5.1 depicts a diagram of the proposed PMSQE loss function, where the clean and enhanced speech signals are processed to compute the two disturbance terms. In the next subsections, we will describe the computation of these disturbance terms from the clean and estimated speech spectra. This computation involves spectral pre-processing and perceptual



(a) Clean speech spectrum

(b) Noisy speech spectrum

(c) Clean Bark spectrum

(d) Noisy Bark spectrum

Fig. 7.2 Example of the clean and noisy speech power spectra (in logarithmic scale) and their corresponding Bark spectra.

domain transformation, the use of frequency and gain equalizations, and the final computation of the symmetrical and asymmetrical disturbances.

7.1.1 Standard listening level and perceptual domain transformation

The first step in the PMSQE algorithm is to transform the power spectrum of the clean and estimated speech signals into a perceptual domain. The clean power spectrum is converted to the Bark frequency scale using the following transformation,

$$\mathbf{b}(t) = \mathbf{B} \cdot G_S \cdot \mathbf{s}(t), \tag{7.1}$$

where vector

$$\mathbf{s}(t) = \left[|X(t,0)|^2, \ldots, |X(t,F-1)|^2 \right]^\top \tag{7.2}$$

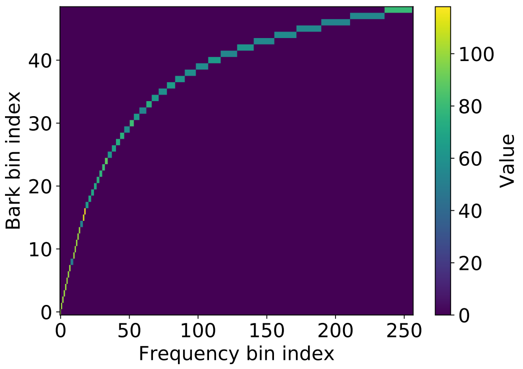


Fig. 7.3 Representation of the Bark matrix coefficients for 16 KHz sampling rate.

consists of the clean power spectrum for a given frame t, $\mathbf{B}$ is the Bark transformation matrix with dimensions $Q \times F$ (Q is the number of Bark bands), and

$$G_S = \frac{P_c}{\frac{1}{T}\sum_t \left(\mathbf{h}^\top \cdot \mathbf{s}(t)\right)}, \tag{7.3}$$

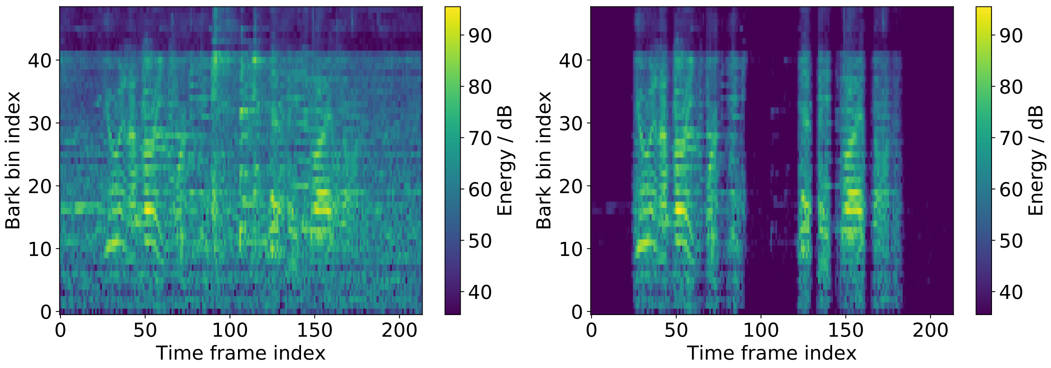
is an utterance gain that equalizes the clean speech signal to a standard listening level (SLL) [3]. The weighting vector $\mathbf{h}$ performs band-pass filtering in the human voice frequency range (from 350 to 3250 Hz), and P_c is a correction factor accounting for STFT parameters such as the window type. Similarly, the estimated Bark spectrum $\widehat{\mathbf{b}}(t)$ can be obtained as

$$\widehat{\mathbf{b}}(t) = \mathbf{B} \cdot G_{\widehat{S}} \cdot \widehat{\mathbf{s}}(t), \tag{7.4}$$

where $\widehat{\mathbf{s}}(t)$ is the estimated clean power spectrum and $G_{\widehat{S}}$ is the SLL equalization gain for the estimated spectrum (computed using $\widehat{\mathbf{s}}(t)$). Fig. 7.2 shows the clean and noisy speech spectra for a given utterance and their corresponding Bark spectra obtained after applying the SLL equalization and the Bark matrix. The Bark matrix is also represented in Fig. 7.3. It can be observed that most of the coefficients of the Bark spectra focus on the low frequencies of the linear spectrum. These regions are mainly dominated by the speech formants. The Bark matrix also emulates the logarithmic behavior of the human auditive system.

Next, the estimated Bark spectrum is equalized to compensate for some effects that are not relevant to the human listener [3]. The frequency equalized Bark spectrum is obtained as

$$\widehat{\mathbf{b}}'(t) = \mathbf{k} \odot \widehat{\mathbf{b}}(t) \tag{7.5}$$



(a) Freq. equalization

(b) Freq. and gain equalization

Fig. 7.4 Example of the effect of the different equalizations on the estimated Bark spectra.

where $\mathbf{k}$ is an utterance vector of band-dependent values and $\odot$ is the element-wise vector multiplication. The gain and frequency equalized bark spectrum is then obtained as

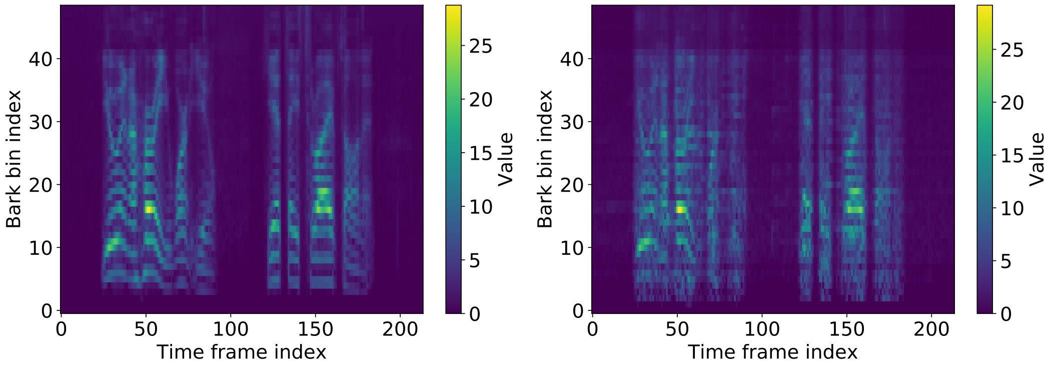
$$\widehat{\mathbf{b}}''(t) = g(t) \cdot \widehat{\mathbf{b}}'(t) \tag{7.6}$$

where $g(t)$ is a per frame value that applies the gain equalization. The computation of these equalizers is explained in the next subsection. Fig. 7.4 shows the effects of the different equalizations applied to the estimated Bark spectrum of the analyzed example. The frequency equalization attenuates that Bark bands where the clean speech signal has low power. On the other hand, the gain equalization attenuates that time frames where the speech source is not active. Therefore, these equalizations help to focus on the speech-active regions in the case of an additive noise signal.

Finally, the clean $\mathbf{b}(t)$ and (equalized) estimated $\widehat{\mathbf{b}}''(t)$ Bark spectra are converted to a sone loudness scale by using Zwicker's law [257]. For example, the clean speech loudness is obtained as

$$\mathbf{z}(t) = z_l \cdot (2 \cdot \mathbf{p}_b)^{\gamma_b} \odot \left[\left(\frac{1}{2} + \frac{\mathbf{b}(t)}{2 \cdot \mathbf{p}_b} \right)^{\gamma_b} - 1 \right], \tag{7.7}$$

where z_l is a loudness scaling factor, $\mathbf{p}_b$ is the vector of absolute threshold powers, and γ_b is the vector of modified Zwicker powers. The division and power operations between vectors are done element-wise. On the other hand, the components of $\mathbf{z}(t)$ where $\mathbf{b}(t) < \mathbf{p}_b$ are directly set to zero. The same procedure is followed to obtain $\widehat{\mathbf{z}}(t)$ from the equalized Bark spectrum $\widehat{\mathbf{b}}''(t)$. Fig. 7.5 shows the resulting clean and estimated loudness spectra from the previous example utterance. The effect of the additive noise in the loudness spectra can be appreciated, distorting the speech formants and increasing the loudness energy between them.



(a) Clean loudness spectrum

(b) Noisy loudness spectrum

Fig. 7.5 Example of the loudness spectra obtained after applying the Zwicker's law to the Bark spectra.

7.1.2 Equalizations in Bark domain

There are some effects in the Bark domain that are not perceived by human listeners as speech quality degradations, i.e. a time-invariant non-severe filtering or short-term gain variations. Therefore, we can compensate for these effects in the estimated Bark spectrum by means of equalization. As explained in the previous subsection, two equalizations are proposed, both based on the PESQ algorithm: frequency and gain equalization.

Before explaining both equalizations, the concept of audible power must be introduced. The audible power at each frame is the sum of the values of those Bark bands with enough power to be perceived by a human listener. To compute it, we first define a vector that indicates which bands are above an audible threshold,

$$\mathbf{u}^{(\alpha)}(t) = \mathscr{U}\left(\mathbf{b}(t) - \alpha \mathbf{p}_b\right), \tag{7.8}$$

where α is a scaling factor and $\mathscr{U}(\cdot)$ is the element-wise step function, which equals one when the argument is greater than zero, and zero otherwise. The audible power for the clean spectrum is finally obtained as

$$A^{(\alpha)}(t) = \mathbf{b}^\top(t) \cdot \mathbf{u}^{(\alpha)}(t). \tag{7.9}$$

The frequency equalizer compensates constant filtering over the frames of the estimated Bark spectrum. This vector is obtained as the ratio between the per-band total power of the clean and estimated Bark spectrum,

$$\mathbf{k} = \frac{\sum_t \beta(t)\mathbf{b}_{\text{th}}(t) + \varepsilon_k}{\sum_t \beta(t)\widehat{\mathbf{b}}_{\text{th}}(t) + \varepsilon_k}, \tag{7.10}$$

where ε_k is a bias value to stabilize the ratio against very small values, $\beta(t)$ is an audible power-based voice activity detector to remove silent frames from the computation ($\beta(t) = 1$ if $A^{(100)}(t) > 10^7$, and zero otherwise) and

$$\mathbf{b}_{\text{th}}(t) = \mathbf{b}(t) \odot \mathbf{u}^{(100)}(t), \tag{7.11}$$

$$\widehat{\mathbf{b}}_{\text{th}}(t) = \widehat{\mathbf{b}}(t) \odot \mathbf{u}^{(100)}(t), \tag{7.12}$$

are the thresholded clean and estimated Bark spectra, respectively. As it can be observed, the masked bands at each frame are obtained from the clean Bark spectrum. The vector division is performed element-wise and the final values are bounded in the range $[-20, 20]$ dB.

To obtain the gain equalizer, we first compute an auxiliary gain as the ratio between the clean and estimated audible power at each frame,

$$\widetilde{g}(t) = \frac{A^{(1)}(t) + \varepsilon_g}{\widehat{A}^{(1)}(t) + \varepsilon_g}, \tag{7.13}$$

where ε_g is a bias value and the estimated audible power $\widehat{A}^{(1)}(t)$ is computed from the frequency-equalized Bark spectrum,

$$\widehat{A}^{(1)}(t) = \mathbf{b}'^{\top}(t) \cdot \mathscr{U}\left(\mathbf{b}'(t) - \mathbf{p}_b\right). \tag{7.14}$$

The final values are bounded in the range $\left[3 \cdot 10^{-4}, 5\right]$. Then, a convolutional layer with fixed parameters is used on the obtained gain values to apply a smoothness over the time dimension using first-order low-pass filtering, thus yielding the final gain

$$g(t) = 0.8 \cdot \widetilde{g}(t) + 0.2 \cdot \widetilde{g}(t-1), \tag{7.15}$$

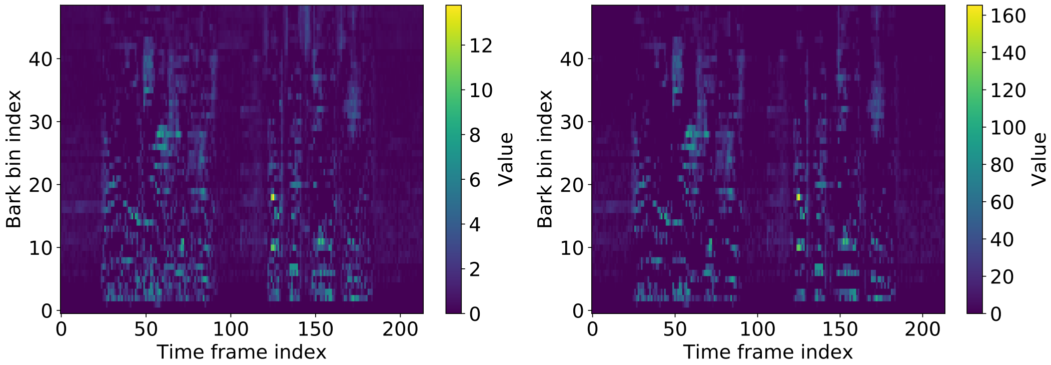
as described in the PESQ algorithm [3].

7.1.3 Disturbances computation

Finally, we compute the symmetrical and asymmetrical disturbances, $D_s(t)$ and $D_a(t)$, respectively, used for the DNN optimization. First, we have to obtain the symmetrical and asymmetrical disturbance vectors, $\mathbf{d}_s(t)$ and $\mathbf{d}_a(t)$ respectively, from the loudness spectra.

The symmetrical disturbance vector is obtained as the absolute difference between the loudness spectra as

$$\mathbf{d}_s(t) = \max\left(|\widehat{\mathbf{z}}(t) - \mathbf{z}(t)| - \mathbf{m}(t), \mathbf{0}\right), \tag{7.16}$$



(a) Symmetrical disturbance

(b) Asymmetrical disturbance

Fig. 7.6 Example of the symmetrical and asymmetrical disturbance vectors per frame.

where $\mathbf{0}$ is a zero-filled vector of length Q, and

$$\mathbf{m}(t) = 0.25 \cdot \min\left(\widehat{\mathbf{z}}(t), \mathbf{z}(t)\right), \tag{7.17}$$

is a center-clipping factor. This center-clipping takes into account the psychoacoustic process by which small spectra differences are inaudible when loud signals are present. The absolute value, maximum and minimum operators are applied element-wise over the vectors.

The asymmetrical disturbance vector is computed from the symmetrical disturbance vector as

$$\mathbf{d}_a(t) = \mathbf{d}_s(t) \odot \mathbf{r}(t), \tag{7.18}$$

where $\mathbf{r}(t)$ is a vector of asymmetric ratios that weights differently positive and negative differences. This vector is obtained from the Bark spectra as

$$\mathbf{r}(t) = \min\left(\mathbf{r}^{'}(t), 12 \cdot \mathbf{1}\right) \cdot \mathscr{U}\left(\mathbf{r}^{'}(t) - 3\right), \tag{7.19}$$

$$\mathbf{r}^{'}(t) = \left(\frac{\widehat{\mathbf{b}}(t) + \varepsilon_d}{\mathbf{b}(t) + \varepsilon_d}\right)^{\lambda}, \tag{7.20}$$

where ε_d is a bias factor, λ is a power factor, and $\mathbf{1}$ is a one-filled vector of length Q.

Finally, the per-frame disturbance terms are obtained as the following weighted norms [3],

$$D_s(t) = \min\left(\eta(t) \cdot \|\mathbf{w}\|_1^{\frac{1}{2}} \cdot \|\mathbf{w} \odot \mathbf{d}_s(t)\|_2, 45\right), \tag{7.21}$$

$$D_a(t) = \min\left(\eta(t) \cdot \mathbf{w}^{\top} \cdot \mathbf{d}_a(t), 45\right), \tag{7.22}$$

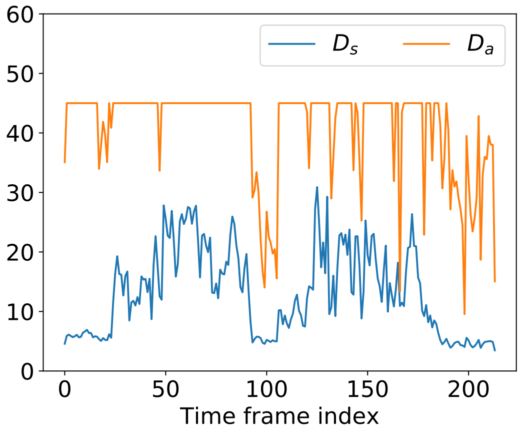


Fig. 7.7 Example of the symmetrical and asymmetrical disturbances obtained at each time frame.

where $\|\cdot\|_p$ is the $\mathscr{L}_p$-norm, $\mathbf{w}$ is a vector with weights proportional to the width of the Bark bands, and $\eta(t)$ is an audible power-based scaling factor computed as

$$\eta(t) = \left(\frac{A^{(1)}(t)+c}{100\cdot c}\right)^{-\xi}, \tag{7.23}$$

where c and ξ are scalar factors. Fig. 7.6 shows the symmetrical and asymmetrical disturbance vectors obtained at each time frame in the example utterance, while Fig. 7.7 show the final per-frame values of the disturbances. It can be observed that the value of the disturbances is higher at those frames where both speech and noise are present. Besides, the asymmetrical disturbance is generally higher than the symmetrical disturbance. It must be taken into account that the asymmetrical disturbance gives more importance to the presence of additive noise signals, as in the case of the analyzed example.

7.2 Integration of the PMSQE as a loss function

The PMSQE disturbances can be used to assist the training of a DNN intended for the estimation of clean speech, which is a common step for different speech processing tasks. We now focus on a speech enhancement scenario, where an enhanced signal is obtained from noisy speech. The PMSQE loss can be used to guide the DNN, thus achieving better speech quality. Moreover, we can incorporate perceptual considerations into DNN training. This can yield improvements in the subjective perception of the speech signal by a human listener.

Table 7.1 Hyperparameters used in the PMSQE loss function for speech enhancement.

Param.	**Value**	**Param.**	**Value**
P_c	2	ε_d	50
ε_k	10^3	λ	1.2
ε_g	$5 \cdot 10^3$	c	10^5
z_l	0.187	ξ	0.04

Table 7.1 shows the value of the different hyperparameters used in the PMSQE loss function for speech enhancement.

As previously mentioned in Chapter 2, there are two main approaches for single-channel DNN speech enhancement in the STFT domain: spectral mapping and spectral masking. In the next subsections, we will show how the PMSQE loss function can be defined for these different approaches.

7.2.1 PMSQE for spectral mapping speech enhancement

We consider now the spectral mapping approach proposed in [242, 243], which predicts the log-magnitude spectral coefficients of a clean speech signal from the ones of a noisy version. Let us define $z_x(t,f)$ as the mean and variance normalized log coefficients of the clean speech signal, as in (2.63), and $\widehat{z}_x(t,f)$ the estimated values obtained from the DNN. The enhanced speech signal is then obtained as

$$\widehat{X}(t,f) = e^{\widehat{\sigma}_y(f)\left(\widehat{z}_x(t,f)+\widehat{\mu}_y(f)\right)+j\theta_y(t,f)} \tag{7.24}$$

where $\widehat{\mu}_y(f)$ $\widehat{\sigma}_y(f)$ are the mean and variance normalization values, respectively, and $\theta_y(t,f)$ is the noisy phase, which is used instead of the unknown clean phase. The log-domain MSE loss function defined in (2.64) is used during the DNN training. This loss function can be re-written, using the definition in (2.63), as

$$\mathscr{L}_{\text{log-MSE}} = \frac{1}{T}\frac{1}{F}\sum_t\sum_f \frac{1}{\sigma_y^2(f)}\left(\log\frac{|X(t,f)|^2}{|\widehat{X}(t,f)|^2}\right)^2. \tag{7.25}$$

As can be observed, this loss function essentially averages a weighted squared log-ratio between the target and enhanced clean speech power spectra across the time-frequency bins.

To take perceptual features into account, we modify the log-MSE loss by incorporating the two PMSQE disturbance terms. These terms are intended to meliorate the log-MSE loss function, as the two disturbance terms can lead to gradient misguidance if applied alone. This is because the spectral mapping approach yields artifacts in some frequencies, which are not taken into account by the disturbances terms but deteriorate the performance of the resulting enhanced signal. In our preliminary experiments, these artifacts commonly appeared in frequencies where there are not speech components. The log-MSE term enforces the target and estimated coefficients to have similar power, avoiding these artifacts that distort the resulting signal. On the other hand, the disturbance terms regularize the log-MSE loss, improving the final speech quality by focusing on the frequency regions that are important to the human auditory system. Thus, the final log-PMSQE loss function can be defined as,

$$\mathscr{L}_{\text{log-PMSQE}} = \mathscr{L}_{\text{log-MSE}} + \frac{1}{T}\sum_{t}\left(\alpha_{LP}\cdot D_s(t) + \beta_{LP}\cdot D_a(t)\right), \tag{7.26}$$

where α_{LP} and β_{LP} are weighting factors experimentally determined. The previous equation can be seen as a multi-objective optimization function where both the log-MSE and the PMSQE disturbances terms have to be jointly optimized.

7.2.2 PMSQE for spectral masking speech enhancement

The spectral masking approach is based on the prediction of a real-valued mask $\widehat{M}_x(t,f)$ that is applied on the noisy speech signal $Y(t,f)$ to obtain an estimate $\widehat{X}(t,f)$. For training, we will follow a signal approximation criterion where the loss function directly measures the error between the enhanced and target clean speech magnitude coefficients.

Regarding the PMSQE loss function, the advantage of the spectral masking approach is that the mask definition prevents the DNN to generate artifacts in frequency bands (masks are defined between zero and one, so they can only attenuate frequency bins). This allows the disturbance terms to be used alone as a loss function. Thus, we define the PMSQE loss as a function that only considers the PESQ disturbance terms as

$$\mathscr{L}_{\text{PMSQE}} = \frac{1}{T}\sum_{t}\left(D_s(t) + \beta_P\cdot D_a(t)\right), \tag{7.27}$$

where β_P is a weighting factor that controls the relative importance of both disturbances.

This PMSQE loss function can also be combined with other losses in order to improve other important characteristics, as speech intelligibility or the distortion level. This will be referred to as multi-objective learning (MOL). The MOL approach defines a loss function that integrates a set $\mathscr{K}$ of loss functions. The final loss function can then be expressed as

$$\mathscr{L}_{\text{MOL}} = \sum_{k\in\mathscr{K}} a_k\mathscr{L}_k, \tag{7.28}$$

where k index every loss function in $\mathscr{K}$, and a_k are the corresponding weighting coefficients. This MOL strategy will be explored with MSE and other objective metric loss functions along with our proposed PMSQE loss.

7.3 Experimental results: Spectral mapping

We first evaluated the spectral mapping approach using our proposed log-PMSQE loss. The DNN was trained and evaluated using the VCTK-Noisy speech database, which contains noisy speech audio files at 8 kHz. We used an STFT of 256-sample frame length with 50% overlap and a Hanning windowing. This yielded frames of 129 frequency components. We chose a feed-forward DNN regressor, as in [243, 140], which included three hidden layers with 2048 rectifier linear units (ReLU) and a linear output layer of 129 units. A temporal context of 4 previous and subsequent frames was applied in the input layer, so the input of the network had a size of 1161 components. The log-magnitude spectrum (LMS) vectors were mean- and variance-normalized using the training set statistics. To prevent overfitting, dropout was applied to hidden layers with a de-activation probability factor of 0.1.

In order to evaluate our proposed loss function, we used a narrowband version of PMSQE. In this version, the Bark matrix only considers the frequencies up to 8 kHz (similarly to the narrowband PESQ [4]). Moreover, the implementation used for spectral mapping with a feed-forward DNN regressor differed from the presented approach in the following aspects:

- SLL normalization and gain and frequency equalizations were applied at a batch-level.
- Gain equalization factor smoothing using convolutional layers was not performed.

In the next subsections, we will show how the hyperparameter factors of the loss function have been determined. Also, the objective and subjective evaluation results will be presented.

7.3.1 Hyperparameter optimization

We first used the Aurora-2 database to optimize the weights α_{LP} and β_{LP} in the log-PMSQE loss in (7.26) with an independent dataset. To reduce the number of possible combinations and make the search easier, we set the same relative weighting between the symmetrical and asymmetrical disturbances as in the PESQ algorithm, i.e. $\beta_{LP} = 0.309\alpha_{LP}$. We also evaluated the effect of the two equalization steps proposed in the PMSQE approach, that is, the gain equalization and the frequency equalization.

Fig. 7.8 shows the average PESQ scores and 95% confidence intervals obtained on the Aurora-2 test set for different values of α_{LP}. We evaluated the proposed log-PMSQE loss

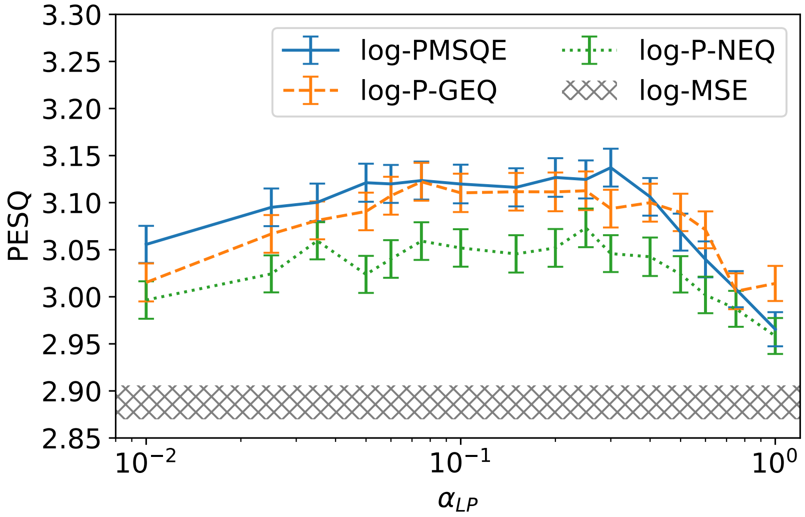


Fig. 7.8 Average PESQ scores with 95% confidence intervals obtained on Aurora-2 test set by the proposed log-PMSQE metric without equalization (log-P-NEQ), with gain equalizations (log-P-GEQ) and with all the equalizations (log-PMSQE). Several values for α_{LP} are evaluated. Results from the log-MSE loss function are also shown (95% confidence interval band).

without equalization (log-P-NEQ), with gain equalization only (log-P-GEQ), and including both the gain and the frequency equalization (log-PMSQE). The results obtained with the log-MSE loss function are also shown as a reference. As can be observed, the proposed log-PMSQE, including the equalizations, performs the best in general, yielding a plateau in performance when α_{LP} is around 0.1. Therefore, we selected the log-PMSQE method with the value of $\alpha_{LP} = 0.1$ for the rest of the evaluation.

7.3.2 Objective evaluation results

Tables 7.2 and 7.3 show the performance of our approach in comparison with other loss functions in terms of objective perceptual quality evaluated using the PESQ and SDR metrics. The tables show the average results obtained for each SNR level. The noisy speech scores (Noisy) are reported as a reference. We compare the log-PMSQE approach with the log-MSE loss function and another two perceptually oriented losses, both proposed in [101]: the Mel-frequency weighted log-MSE loss (wlog-MSE) and a variant of this loss that includes a regularization by spectral variation similarity (wlog-MSE-SVS). The different methods can be applied on a per-frame basis in the spectral domain.

The results show that our proposal achieves the best results in terms of PESQ, yielding an absolute average increase of 0.12 points in the PESQ score with respect to the other

Table 7.2 PESQ scores obtained for the noisy and the DNN enhanced speech signal with different loss functions over the VCTK-Noisy test set.

Method	**SNR (dB)**						**Avg.**
	-5	**0**	**5**	**10**	**15**	**20**	
Noisy	1.62	1.82	2.10	2.42	2.73	3.00	2.28
log-MSE	1.77	2.12	2.47	2.76	3.00	3.20	2.55
wlog-MSE	1.77	2.14	2.50	2.80	3.04	3.25	2.58
wlog-MSE-SVS	1.77	2.11	2.47	2.78	3.04	3.26	2.57
log-PMSQE	**1.89**	**2.27**	**2.62**	**2.89**	**3.13**	**3.34**	**2.69**

Table 7.3 SDR values (in dB) obtained for the noisy and the DNN enhanced speech signal with different loss functions over the VCTK-Noisy test set.

Method	**SNR (dB)**						**Avg.**
	-5	**0**	**5**	**10**	**15**	**20**	
Noisy	-	-	-	-	-	-	-
log-MSE	-2.62	3.03	7.17	10.15	12.03	12.91	7.11
wlog-MSE	-2.74	2.89	7.22	10.47	12.59	13.62	7.34
wlog-MSE-SVS	-3.06	2.80	7.46	**11.08**	**13.62**	**15.02**	**7.82**
log-PMSQE	**-1.53**	**4.14**	**7.85**	9.94	10.89	11.19	7.08

compared losses under *unseen* noise conditions. Furthermore, the wlog-MSE and wlog-MSE-SVS approaches yield PESQ scores almost identical to log-MSE. The good results on PESQ can be expected as our proposal directly optimizes a loss based on PESQ, which also demonstrates that the DNN is correctly guided. On the other hand, the compared techniques outperform our proposal in terms of average SDR. Thus, it seems that the improvement achieved on the perceptual speech quality is at the cost of some speech distortion. This is especially noticeable at high SNRs where the log-PMSQE approach cannot improve the results. Nonetheless, the average SDR reduction is small in comparison with the results obtained using the log-MSE, while significant improvements can be observed at low SNRs. Therefore, we can assert that our proposal performs well in general and outperform related metrics in difficult scenarios with low SNRs.

7.3.3 Subjective evaluation results

Finally, we conducted a subjective listening test to evaluate the perceived quality of the enhanced signals by human listeners. We followed a *Comparative Mean Opinion Score* (CMOS) evaluation [2]. The listeners were asked to compare pairs of enhanced speech utterances in terms of the overall perceived quality using a Likert-style scale from -3 to 3 (-3:

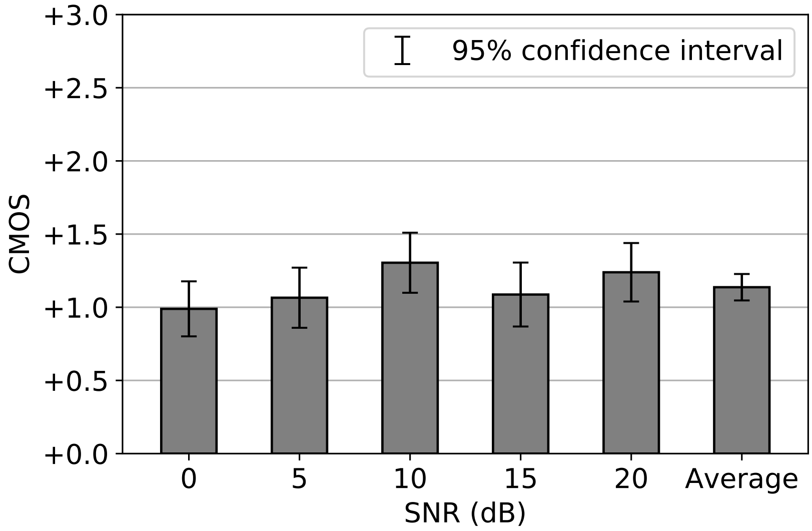


Fig. 7.9 CMOS scores averaged per SNR level when comparing speech signal enhanced with log-PMSQE with respect to wlog-MSE-SVS.

the 1st speech signal sounds much better than the second one, ..., 0: both speech signals sound similar, ..., 3: the 2nd speech signal sounds much better than the first one). The subjective test was conducted in a quiet room using professional headphones and a web-based interface. A total of twenty-three listeners with normal hearing and no previous speech processing knowledge participated in this test. Each listener evaluated a total of 20 randomly-chosen enhanced speech pairs from the test set (a pair for each type of noise and SNR condition, with the SNR range limited from 0 to 20 dB). The enhanced methods compared were the proposed log-PMSQE approach and the wlog-MSE-SVS metric, which obtain the best objective results on average. The speech signal pairs were presented to the listeners in a random order to control a possible order effect bias.

Fig. 7.9 shows the average CMOS scores obtained by our proposal for each SNR level and on average. Positive CMOS scores indicate a preference for our proposed approach. As can be observed, the listening test confirms that our proposal outperforms wlog-MSE-SVS in terms of subjective quality with real listeners in every tested SNR condition. This reveals a noticeable increase of the perceived quality on average, with a global result slightly above $+1.0$ CMOS score. Besides, the scores for the different SNRs evaluated keep similar. These results demonstrate that the improvements on the PESQ metric are not achieved by cheating the objective metric to obtain better PESQ scores. Thus, the good objective metric results also translate to a better perceived quality by the final human listeners.

Table 7.4 Architecture of the CRN applied to spectral masking estimation. The feature size is indicated in the form *feature maps* × *frames* × *freq. channels*. The hyperparameters column refers to *kernel size*, *stride* and *output channels*. For the LSTM layers, the number of hidden units is also indicated.

Layer Name	**Input size**	**Hyperparameters**	**Output size**
conv_1	1 × T × 257	1 × 3, (1, 2), 8	8 × T × 128
conv_2	8 × T × 128	1 × 3, (1, 2), 16	16 × T × 64
conv_3	16 × T × 64	1 × 3, (1, 2), 32	32 × T × 32
conv_4	32 × T × 32	1 × 3, (1, 2), 64	64 × T × 16
conv_5	64 × T × 16	1 × 3, (1, 2), 128	128 × T × 8
reshape_1	128 × T × 8	-	T × 1024
lstm_1	T × 1024	1024	T × 1024
lstm_2	T × 1024	1024	T × 1024
reshape_2	T × 1024	-	128 × T × 8
deconv_5	256 × T × 8	1 × 3, (1, 2), 64	64 × T × 16
deconv_4	128 × T × 16	1 × 3, (1, 2), 32	32 × T × 32
deconv_3	64 × T × 32	1 × 3, (1, 2), 16	16 × T × 64
deconv_2	32 × T × 64	1 × 3, (1, 2), 8	8 × T × 128
deconv_1	16 × T × 128	1 × 3, (1, 2), 1	1 × T × 257

7.4 Experimental results: Spectral masking

The PMSQE loss function was then evaluated for a spectral masking approach. The different experiments were performed using the TIMIT-1C database, which contains simulated noisy speech signals at 16 kHz. For the computation of the STFT, a 512-point DFT was applied using a 32 ms square-root Hann window with a 50% overlap. This resulted in a total of 257 frequency bins for each time frame. The different loss functions were evaluated using a convolutional recurrent network (CRN) similar to the one used in Chapter 4. A description of the CRN network architecture is provided in Table 7.4. Apart from the different number of parameters, the CRN employed in these experiments used two LSTM layers to better exploit the temporal information. A dropout layer was used at the input of each LSTM layer with a deactivation probability of 0.5. The input features were the log-magnitude spectrum of the noisy speech signal. A recursive mean normalization was applied before feeding the feature map into the network [72]. For the training setup, we used a batch size of 10 utterances, and the sequences were zero-padded to have the same number of frames.

In the spectral masking case, we evaluated the wideband version of the PMSQE loss function, which uses the Bark matrix up to 16 kHz. In addition, the different normalizations and equalizations were applied utterance-wise. We used the value $\beta_P = 0.309$ in the PMSQE loss $\mathscr{L}_{\text{PMSQE}}$ described in (7.27). The PMSQE loss was compared with other related loss functions proposed in the literature:

- The MSE loss function $\mathscr{L}_{\text{MSE}}$ for spectral masking using the signal approximation approach, as described in (2.68).

- The ESTOI loss function $\mathscr{L}_{\text{ESTOI}}$ [155], based on the ESTOI objective metric described in Chapter 3. The objective is to maximize the ESTOI score, which correlates with speech intelligibility perceived by a human listener. Thus, the negative value of the ESTOI metric was used to compute the loss function. This loss function was evaluated in [155] for magnitude spectral masking in a speech enhancement task. We adapted this loss function to our STFT framework using an analysis window of 24 STFT frames, which is equivalent to the 384 ms window of the metric implementation [95]. In addition, we extended the number of third-octave bands from 15 to 18 to cover the 8 kHz frequency range, as in [155]. Finally, we also proposed the implementation of the same mechanism for removing silent frames in the loss computation than the original metric. This VAD applies thresholding to the frame power computed from the STFT clean spectrum.

- The scale-invariant SDR loss function $\mathscr{L}_{\text{SDR}}$ [219], implemented using Eq. (3.13). The objective is to maximize the SDR, so the negative value is taken in the SI-SD equation during the DNN training for backpropagation. To compute the loss, the ISTFT is first applied to the enhanced spectrum to obtain the time-domain enhanced signal. Then, the enhanced time-domain signal is compared with the clean time-domain signal.

- The multi-objective learning strategy in (7.28) was also explored by combining different loss functions during the DNN training.

The trained DNN models were evaluated for speech enhancement using the following objective metrics: PESQ, ESTOI, and SI-SDR. Thus, the objective was to assess the performance of the different loss functions in their respective and other related metrics. Table 7.5 shows the overall mean results obtained for the noisy speech signals of the test set for both *seen* and *unseen* noisy environments. These results will be taken as a reference to compare the different approaches in terms of the gains obtained with respect to the noisy speech results.

In the next subsections, we will evaluate the performance of the proposed PMSQE loss function for spectral masking. First, we will analyze the impact of the different equalizations on the improvements achieved by the PMSQE loss. Next, the PMSQE loss will be compared with the common MSE loss and the previously presented state-of-the-art loss functions, based

Table 7.5 Results provided by the different objective metrics for the noisy speech signals of the test set in the TIMIT-1C database. Results are broken down by SNR and noise condition (seen or unseen during training phase).

Objective Metric	**Noises**	**SNR (dB)**					
		-5	**0**	**5**	**10**	**15**	**20**
PESQ	Seen	1.06	1.12	1.25	1.53	1.90	2.41
	Unseen	1.16	1.27	1.50	1.82	2.22	2.77
ESTOI	Seen	0.37	0.49	0.61	0.73	0.84	0.91
	Unseen	0.41	0.53	0.65	0.77	0.86	0.93
SDR	Seen	-5.76	-0.75	4.22	9.19	14.02	18.68
	Unseen	-5.75	-0.76	4.22	9.16	14.00	18.67

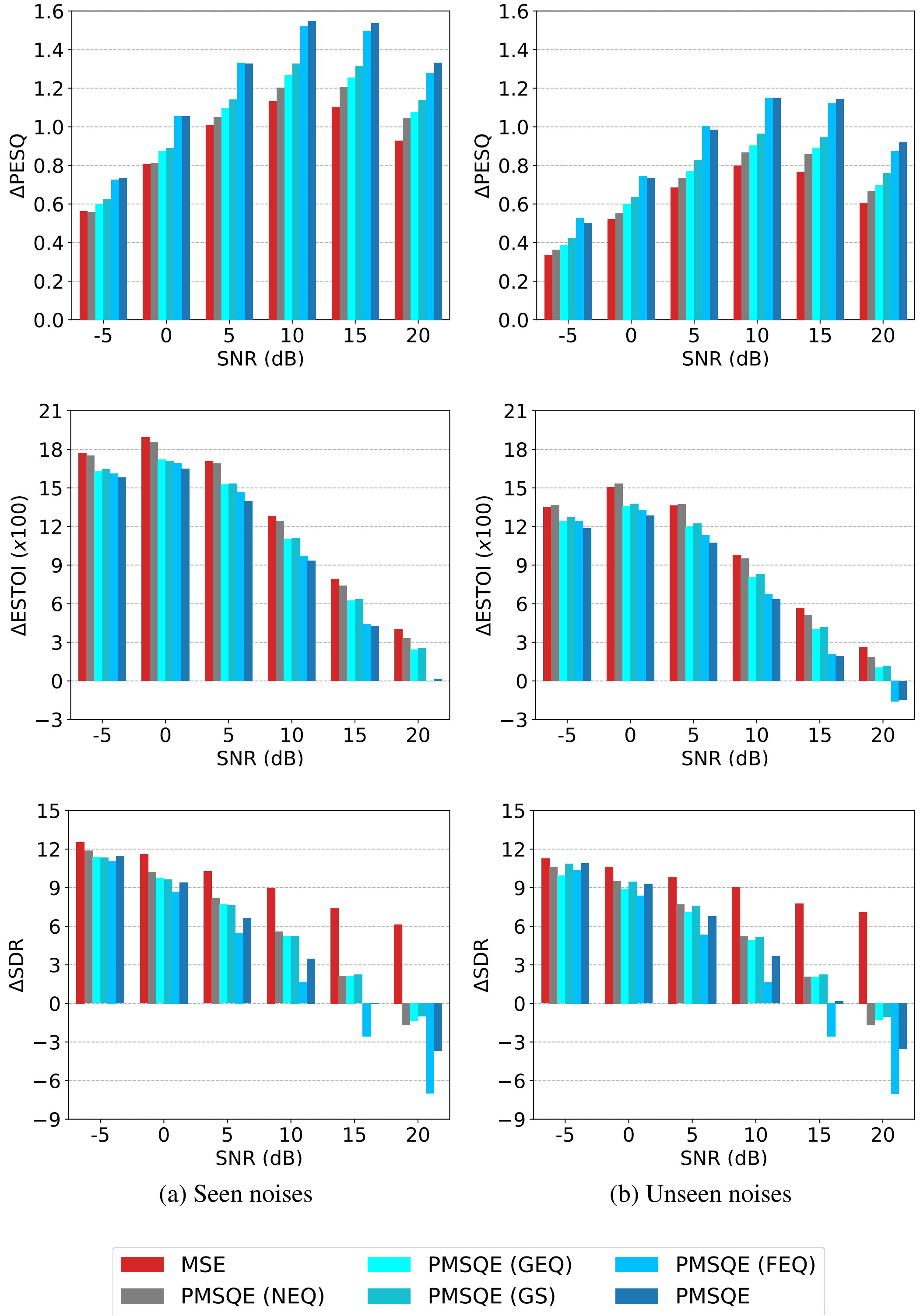
on objective metrics. Finally, we will explore the multi-objective learning approach for mask estimation.

7.4.1 Analysis of PMSQE performance

In this subsection, the performance of the PMSQE loss function and its different equalizations was analyzed for spectral masking. Thus, different variants of the PMSQE loss were tested. These variants are the PMSQE without equalizations (NEQ), two versions including only gain equalization, without the gain smoothing (GEQ) and including the gain smoothing (GS), and, finally, the variant including only frequency equalization (FEQ). Fig. 7.10 shows the obtained results for the different PMSQE variants. Additionally, MSE loss results are included as a baseline.

It is observed that the mask estimator trained with PMSQE outperforms that of MSE when PESQ is evaluated. These results confirm that PMSQE training is an effective approach to increase speech quality, which is a direct consequence of maximizing a speech quality objective metric. On the other hand, this improvement implies a slight reduction in ESTOI and a clear degradation in terms of SDR, which is especially noticeable at high SNRs. This effect can be explained because the PMSQE loss focuses on noise reduction over speech distortion, while SDR is more sensitive to waveform changes, such as speech artifacts and removed speech components. Therefore, training with PMSQE yields mask estimates that aggressively eliminate background noise at the expense of increased speech distortion. The results trend for the *seen* and *unseen* noise conditions is comparable, with smaller improvements for *unseen* noises, as expected.

These results also show how the different PMSQE equalizations contribute to the performance. First, the NEQ variant gives the lower gains for PESQ, but it also has a general minor degradation effect on the other evaluated metrics. The GEQ variant slightly improves



(a) Seen noises

(b) Unseen noises

Fig. 7.10 PESQ, ESTOI and SDR results for the PMSQE loss function and its variants using different equalizations for spectral masking. The MSE loss results are included for comparison purposes. The plots only show increments with respect to the results over noisy speech.

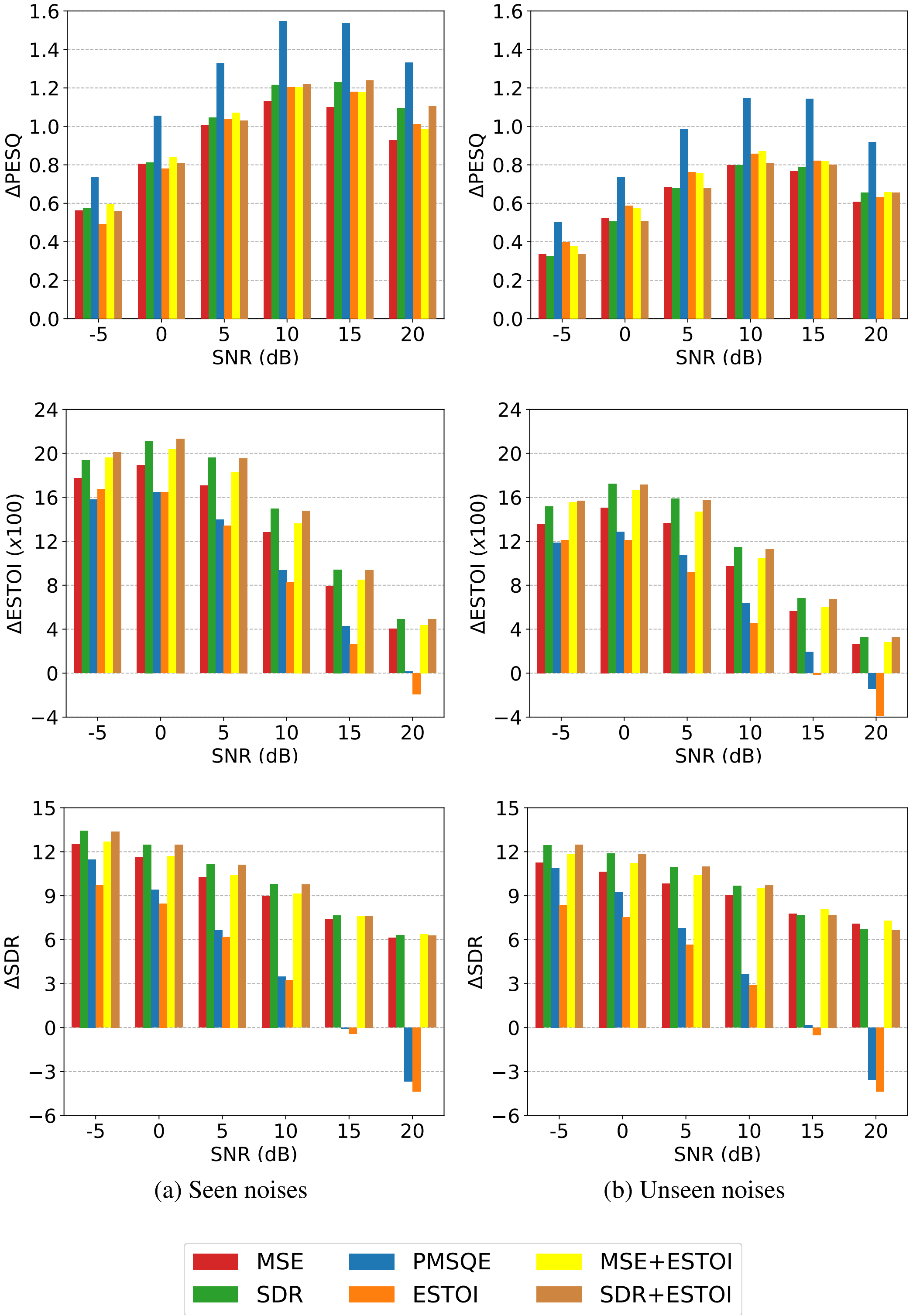
the PESQ results while the remaining metrics keep similar. Moreover, it can be observed that the gain smoothing in the GS variant is beneficial as it increases the performance of gain equalization. Finally, the FEQ variant is the one that contributes most to PESQ improvements, achieving similar PESQ scores than the complete PMSQE loss. Nevertheless, this frequency equalization also introduces high speech distortion when applied alone, thus degrading ESTOI and SDR results. The combination of gain and frequency equalization keeps this PESQ improvement and good performance on the remaining metrics. Therefore, we can conclude that the PMSQE loss, including the different equalizations, has the best average results for spectral masking.

7.4.2 Evaluation of the different loss functions for spectral masking

The PMSQE loss was then compared with the aforementioned ESTOI and SDR losses. The results, including the MSE loss, are shown in Fig. 7.11. In addition, these results include the combination of the ESTOI loss with either the MSE or the SDR losses in a MOL approach. The weighting coefficients in (7.28) for these combinations were $a_{\text{MSE}} = 10^{-8}$ and 1 for the other losses.

The results show that the PMSQE loss achieves the best PESQ results among the different individual losses, which is consistent with the training objective. In the case of the other metrics, the SDR loss achieves the best scores on average, clearly outperforming MSE, PMSQE, and ESTOI losses. The SDR loss aims to reduce speech distortion, which also has a positive impact on speech intelligibility. These results also agree with those of recent papers as [109]. Surprisingly, the ESTOI loss does not share the observed benefits of training with the same loss function as the metric evaluated. The ESTOI loss only achieves improvements on PESQ with respect to the SDR and MSE losses in some cases, especially in *unseen* conditions. Nevertheless, it suffers from a high degradation for the other objective metrics, showing the worst results among the tested losses. These results can be explained by the fact that the ESTOI loss does not consider the time-frequency regions where speech is absent or the instantaneous power of the speech signal, which results in the introduction of artifacts. Thus, the ESTOI loss is not a good choice to improve ESTOI scores when used alone for DNN training.

On the other hand, it can be observed that the combination of the ESTOI loss with MSE or SDR allows for better performance. The combination of MSE and ESTOI losses yields improvements over the MSE loss in the different evaluated metrics and conditions. The use of the SDR loss along with ESTOI shows a similar or even better performance than the previous combination. This is noticeable on the ESTOI and the SDR metrics for low and medium SNRs. Nevertheless, this combination does not yield significant improvements with



(a) Seen noises

(b) Unseen noises

Fig. 7.11 PESQ, ESTOI and SDR results for the different compared loss functions for spectral masking. The ESTOI loss is also evaluated when combined with the MSE and the SDR losses. The plots only show increments with respect to the results over noisy speech.

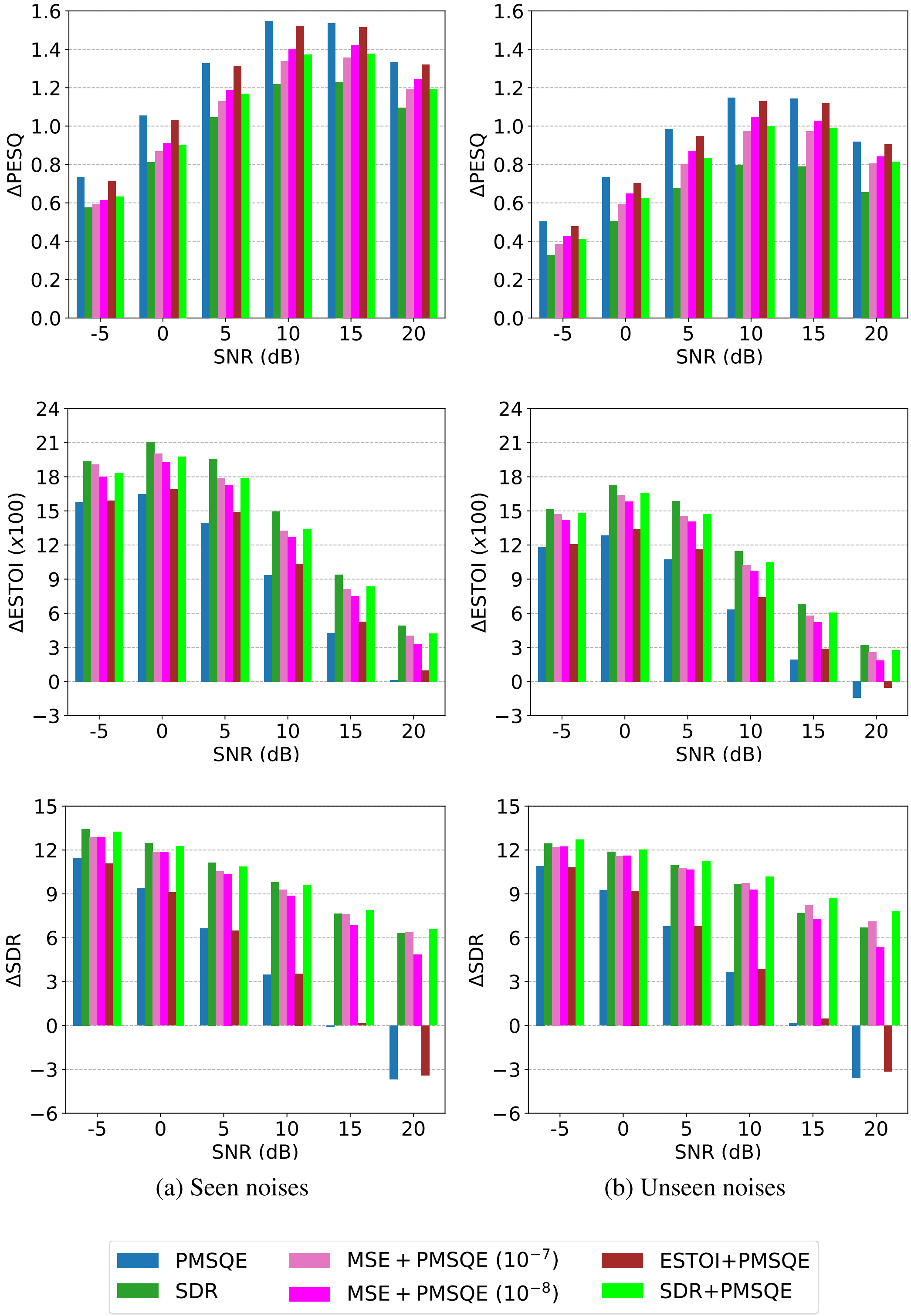
respect to the standalone SDR loss. These results suggest that the SDR loss does not need an additional minimization objective to improve speech intelligibility. To sum up, the PMSQE loss clearly is the best choice in terms of PESQ scores, while the SDR loss exhibits the best performance on ESTOI and SDR metrics.

7.4.3 Multi-objective learning results for spectral masking

To conclude the experimental evaluation, we analyzed the use of the PMSQE loss in a MOL approach for spectral masking. First, the combination of the PMSQE loss with the other evaluated losses was evaluated. Fig. 7.12 shows the results obtained with the different combinations of the PMSQE loss: MSE+PMSQE, ESTOI+PMSQE, and SDR+PMSQE. Regarding the MOL weighting coefficients of Eq. (7.28), the MSE loss was evaluated with coefficients 10^{-7} and 10^{-8} in the MSE+PMSQE approach. On the other hand, the remaining losses (PMSQE, ESTOI, and SDR) are evaluated with a weight of 1 in the different approaches. The results for the standalone PMSQE and SDR losses are also showed for comparison purposes.

These results yield a general observation: the combination of different loss functions can achieve a trade-off between the improvements achieved using the individual losses. Therefore, the combination is usually worst in the related metric with respect to the individual loss, but it achieves competitive performance across the different objective metrics. This is noticeable in the case of the MSE+PMSQE approach, where we have a trade-off between the improvements for PESQ and the other metrics. This trade-off can be controlled by means of the a_{MSE} weight, which changes the relative relevance of both terms. This allows having both a good PESQ performance, comparable to that of the PMSQE loss, and better ESTOI and SDR results. In the case of the ESTOI+PMSQE approach, the variations with respect to PMSQE are smaller, but the ESTOI metric is clearly improved at the expense of a slight reduction in PESQ scores. Finally, the results for the SDR+PMSQE approach are particularly interesting. This strategy achieves comparable results to the MSE+PMSQE approach in terms of PESQ and ESTOI metrics, while it performs better for the SDR metric, especially for *unseen* noise conditions. In general, the SDR+PMSQE approach allows for good trade-off performance in the different evaluated objective metrics.

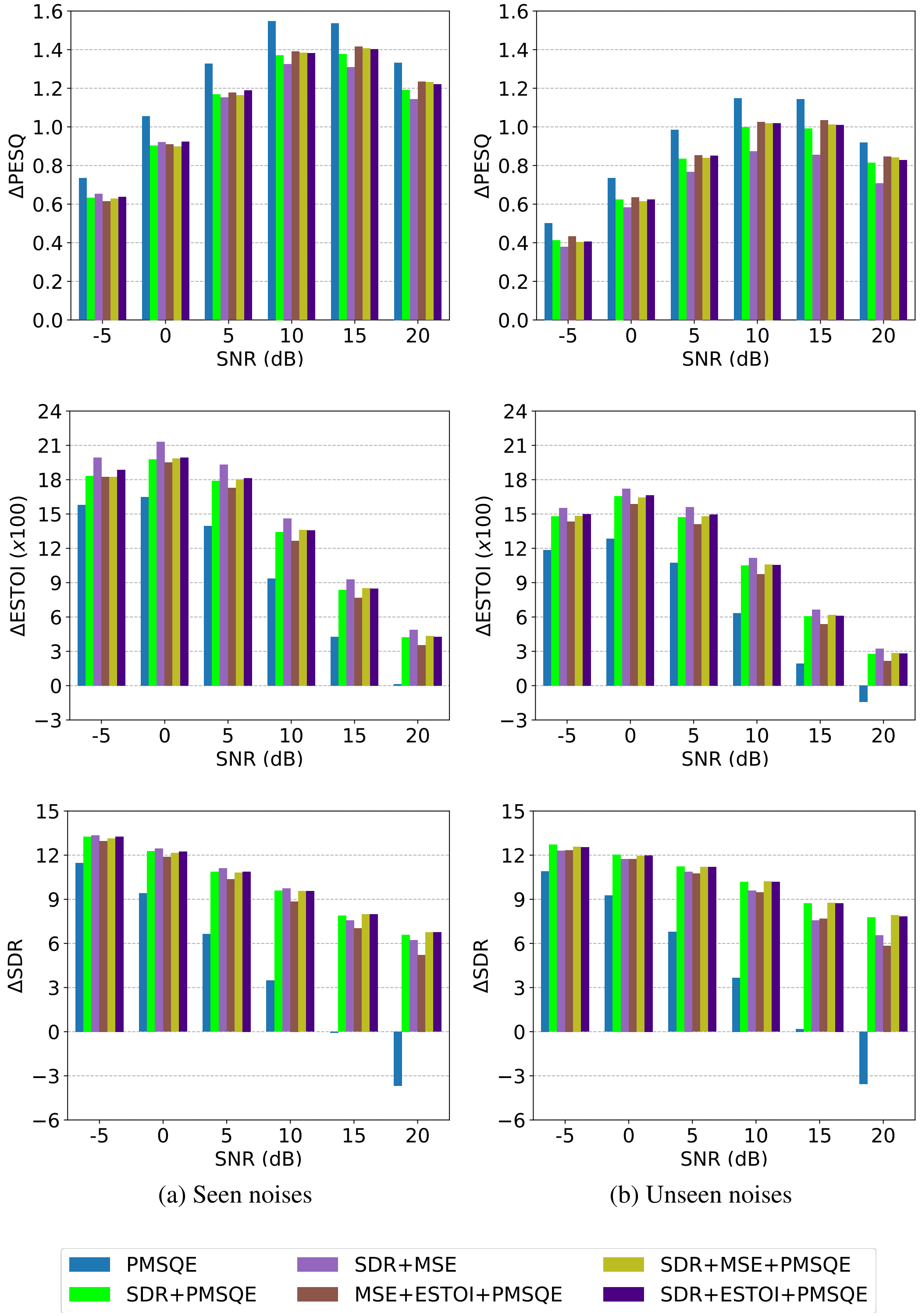
Finally, we evaluated the MOL approach using several loss functions at the same time. Fig. 7.13 compares the results obtained using MOL approaches with two loss functions (SDR+PMSQE and SDR+MSE) and three loss functions (MSE+ESTOI+PMSQE, SDR+MSE+PMSQE, and SDR+ESTOI+PMSQE). The results obtained with the PMSQE loss are also included. The weighting coefficients for the different losses in the MOL approach were $a_{\mathrm{MSE}} = 10^{-8}$ and 1 for the remaining weights.



(a) Seen noises

(b) Unseen noises

Fig. 7.12 PESQ, ESTOI and SDR results for the combination of the PMSQE loss with the other loss functions. The SDR and PMSQE losses are also evaluated. The plots only show increments with respect to the results over noisy speech.



(a) Seen noises

(b) Unseen noises

Fig. 7.13 PESQ, ESTOI and SDR results for the multi-objective learning strategy. The plots only show increments with respect to the results over noisy speech.

As in the previous evaluations, the PMSQE loss remains the best loss function in terms of PESQ scores. The SDR+MSE approach achieves the best ESTOI results, while its performance for PESQ and SDR degrades in *unseen* noise conditions. On the other hand, the MSE+ESTOI+PMSQE approach shows competitive PESQ results, but it does not reach the other losses in the rest of the objective metrics. Finally, the three approaches that include the SDR and PMSQE losses have a similar performance on the different metrics, obtaining the best SDR results and good performance for the other metrics. These results also suggest that the use of additional functions along with the SDR+PMSQE approach does not provide significant gains on speech enhancement performance. Moreover, the SDR+PMSQE approach stands as the best trade-off between speech quality, speech intelligibility, and signal distortion among the evaluated methods.

7.5 Summary

In this chapter, we have proposed a perceptually-motivated loss function to evaluate speech quality in DNN-based speech enhancement methods. The proposed approach is based on the combination of two disturbance terms inspired by the well-known PESQ algorithm: the symmetrical and asymmetrical disturbances. These terms account for different perceptual considerations in the speech signal. To obtain the disturbance terms, the clean and enhanced speech power spectra are first equalized to a standard listening level and then converted to a Bark domain. The Bark spectrum of the enhanced signal is then equalized and the Bark spectra are converted to a loudness domain. Two different equalizations are used to remove non-relevant effects for the perceptual quality: frequency equalization and gain equalization. Disturbance vectors are obtained from the loudness spectra and the disturbance terms are finally computed using weighted norms on a per-frame basis.

The PMSQE approach can be integrated as a loss function for DNN-based speech enhancement. For spectral mapping, the log-spectral domain approximation was chosen, and the PMSQE loss was combined with the log-MSE loss to ensure a good convergence without speech artifacts. In the case of spectral masking, the PMSQE loss was formulated without the need for additional loss terms. Anyway, multi-objective training was also presented for the combination of PMSQE with other loss functions.

The spectral mapping approach was first evaluated using a fully-connected DNN regressor. We first optimized the hyperparameters for the loss function and the combination of equalizations. Then, the proposed log-PMSQE loss was evaluated using objective quality metrics and compared with the log-MSE and other perceptual-related losses. The proposal showed improvements in terms of the perceptual quality performance when the PESQ metric was evaluated, and competitive performance on speech distortion. A subjective analysis with human listeners was then conducted using the CMOS standard procedure to confirm the results obtained using objective metrics. The subjective results showed that the participants had a clear preference for the perceptual quality obtained with our proposal.

Finally, the spectral masking approach was evaluated using a convolutional recurrent DNN mask estimator. The evaluation was conducted using different objective metrics for speech quality and intelligibility. The PMSQE loss was first analyzed in terms of the performance obtained for its different equalizations. The results showed that both equalizations, including gain smoothing, contribute to an increase of the PESQ scores with acceptable speech distortion. The PMSQE loss was then compared with other related losses as MSE, ESTOI, and SI-SDR. The PMSQE approach outperformed the other losses for PESQ results, but it suffered from degradation for the other metrics. Therefore, we finally evaluated the combination of the PMSQE proposal with the other losses. The results showed that the use of the PMSQE approach along with the SDR loss provided good performance and competitive results for the different objective metrics. Thus, the SDR+PMSQE stood as the best choice for the training of DNN-based spectral masking methods among the different evaluated approaches.

Chapter 8

Conclusions

In this work, we have carried out a study of online multichannel speech enhancement combining statistical signal processing and deep neural networks. The conclusions drawn from all the work developed in this Thesis are presented in Section 8.1. Finally, Sections 8.2 and 8.3 are devoted to summarize the contributions and future work, respectively.

8.1 Conclusions

A number of conclusions can be drawn from all the work developed in this Thesis. Some of the most relevant are listed down below:

- The integration of classical statistical signal processing and deep neural network estimators has been proven as a powerful tool for speech enhancement processing. The proposed techniques outperform other approaches that use only one of the previous approximations or do not conveniently integrate them. This combined strategy has allowed the design of speech enhancement algorithms with high control of the different steps of the speech processing pipeline. Thus, in those steps where the assumptions about the signal statistics can be difficult to justify due to the complexity of the problem (i.e., the case of non-stationary environments for the noise estimation or the speech presence uncertainty) we can exploit the powerful modeling capability of deep learning techniques. Therefore, we can skip the limitations imposed by the statistical framework, thus improving the final performance.

- The dual-channel information provided by dual-microphone smartphones can be successfully exploited to improve speech enhancement performance in mobile devices. Although the beamforming approach reduces its utility when the number of microphones is small, the use of a proper postfilter can yield competitive performance regarding other state-of-the-art techniques in smartphones. This postfilter takes advantage of the good estimates obtained from the multichannel information. Moreover, dual-channel properties can be also exploited for an accurate estimation of the acoustic parameters. In the case of the RTF, the a priori statistics and the proposed KF state-model successfully exploits the relation between the clean speech signals at both microphones. For noise estimation, the power level difference between microphones provides a good approach to identify noise segments in CT conditions. On the other hand, the spatial information about the noise field, mainly contained in the signal phase, can be a good discriminator in FT conditions. These cross-channel features can also be exploited by a DNN mask estimator.

- The speech signal in the STFT domain shows important temporal correlations that can be exploited, especially when designing online approaches. Regarding the signal processing approaches, the Kalman filter has demonstrated to be a useful framework to model these temporal relationships across time frames. The proposed eKF estimator has provided more accurate estimates and better tracking capabilities of the RTF between the smartphone microphones than the well-known sub-space approaches. Moreover, the Kalman postfilter has been used to successfully take advantage of the correlations between the clean speech amplitudes, improving the noise reduction capabilities when used in combination with an MVDR beamformer. These temporal properties have also been exploited through the use of recurrent neural networks because of their ability for temporal modeling when dealing with spectrograms.

- The availability of accurate knowledge about speech presence probability in the time-frequency domain was a crucial element in the performance of the proposed algorithms. This information can be used to obtain more accurate noise estimates or even to improve noise reduction in speech absent bins. Among the different approaches, DNN mask estimators have shown astounding performance in the computation of accurate SPP estimates. These deep learning models have better discrimination capabilities between speech and noise dominant bins, and they have shown to be successful in these binary classification tasks. The use of DNN SPP estimators has an important impact when dealing with highly non-stationary noises, as they can quickly adapt to the changes of the signal statistics.

- The proposed REM framework for online multichannel speech enhancement has demonstrated a better noise reduction and speech distortion performance than related state-of-the-art techniques. The use of the beamforming-plus-postfiltering strategy is enhanced by the joint estimation of the clean speech signal, speech presence probability, and the different acoustic parameters involved. This allows for positive feedback between the estimation of the speech statistics and the computation of the acoustic model parameters in the iterative procedure. An important part of the framework is the use of a priori SPPs given by a DNN mask estimator. These a priori SPPs contributes to the convergence of the algorithm, allowing to better discriminate between speech presence and absence bins. Moreover, the statistical framework can be used to refine these SPP estimates by employing a complementary statistical spatial model.

- Another important aspect for most of the proposed approaches in this Thesis is a good parameter initialization, especially for the first frames of the signal. This can be a decisive factor for the proper convergence of the different techniques. For example, the performance of the REM framework does not depend solely on the a priori SPP estimates, but also on the initialization details for the noise statistics and the RTF. Moreover, the use of proper statistics initialization for the block-online beamforming in the spatial-beam approach showed a clear improvement in the ASR results.

- The design of speech processing algorithms for scenarios where multiple speakers overlap is a challenging task where even deep learning discriminator models have difficulties. The use of array microphones along with auxiliary information has shown to be a good mechanism when we are interested in a target speaker. In this case, we can design better DNN mask estimators that exploit both the spectral and spatial information of the multichannel noisy speech and auxiliary signals. This allows for more accurate target speech masks, which can be used for the computation of the beamformer. While the spectral information helps to separate speakers with different frequency patterns, the spatial information can be more decisive for spatially-separated speakers who have similar spectral characteristics, as in the case of speakers with the same gender.

- The DNN-based speech enhancement for single-channel scenarios is a current topic in speech processing research. An approach to improve the quality and intelligibility performance of these DNN estimators is the use of training loss functions based on perceptual considerations about the human auditory system. Among different approaches, the integration of objective quality metrics as deep learning loss functions has shown competitive performance. Thus, our proposed PMSQE approach has shown improvements in perceptual quality performance with better PESQ results. These improvements also translate to better-perceived quality by human listeners. Moreover, the combination of different losses can yield to estimators that provide good results among different quality metrics, thus surpassing the limitations of the standalone approaches.

- Between the two most common approaches for single-channel DNN-based speech enhancement, the spectral masking seems to be a better choice than the spectral mapping one. First, the improvements and the generalization capabilities of the DNN mask estimators are larger than the mapping approaches when the obtained results are compared. Also, the metric-based training, as in the case of our PMSQE approach, is favored when DNN mask estimators are employed. The mask estimation is an easier task for the DNN models, especially when the value range for these masks is well-defined. Moreover, the mask mainly decides how to attenuate the signal at each bin in terms of the SNR. This is, to a certain extent, a task similar to the prediction of SPP masks, where DNNs have shown their suitability. Finally, the restricted values of the masks make it difficult for the presence of speech artifacts when training with these losses. This allows the use of more sophisticated loss functions with better convergence behavior.
- Finally, we have focused on the design of online processing algorithms for speech enhancement. The implementation of online algorithms is a challenging task as they can easily suffer from convergence problems and degradation performance during the first frames. Besides, they can only use past information about the signal, so quick changes in the signal statistics can severely affect the quality of the enhanced signal. Nevertheless, the use of DNN models integrated with proper online speech processing techniques can help to overcome these limitations and obtain good performance in difficult non-stationary noisy environments. Moreover, the design of low-latency and lighter computational algorithms is a required characteristic for the implementation of these techniques in small mobile devices which may be used in real-world conditions.

8.2 Contributions

The different technical contributions resulting from our work are summarized in the following:

- An experimental framework for the simulation of dual-microphone speech signals recorded using a smartphone in noisy and reverberant environments, either in CT or FT conditions [144].

- An estimator of the relative transfer function between two microphones using an extended Kalman filter framework [144], which uses a priori RTF information along with temporal and spatial information to track the variability of the acoustic channels.
- A complete dual-channel speech enhancement framework based on beamforming-plus-postfiltering [145, 146], which uses the eKF-RTF estimator along with better estimates of the clean speech statistics for the postfilter using dual-channel information.
- An SPP-based noise estimation algorithm for dual-microphone smartphones, either using statistical spatial models with a priori SAP estimators based on dual-channel information [146], or DNN-based SPP estimates exploiting dual-channel features [147].
- A novel REM framework for online multichannel speech enhancement integrating a DNN-based SPP estimation [143], which jointly estimates the clean speech signal, the a posteriori SPP and the acoustic parameters, and exploits the temporal correlations using a Kalman postfiltering.
- A multichannel speaker detection algorithm for multiple speakers scenario using block-online beamforming and a DNN-based mask estimator adapted to exploit spectral and spatial auxiliary information to better discriminate the target speaker [142].
- A deep learning loss function for the perceptual evaluation of the speech quality [141], which can be used to train DNN-based speech processing algorithms. We evaluated this approach for DNN-based single-channel speech enhancement.

8.3 Future work

In this Thesis, we have presented different contributions that combine statistical signal processing and deep neural networks. Nevertheless, these integrations are commonly done independently, changing signal processing blocks for DNN estimators to improve the weaker parts of the algorithm. Hence, as future work, it would be interesting to investigate the full integration of signal processing techniques as deep neural network models. Thus, different speech processing blocks, as filters or beamforming techniques, could be implemented using novel convolutional or recurrent architectures. These networks could learn the important statistics from the data while keeping the knowledge about how each processing block is performing. This approach seems very promising to achieve systems with a lower computational burden and with good generalization capabilities.

On the other hand, the proposed approaches are designed to mainly deal with additive environmental noise, while other kinds of distortions (echoes, reverberations, interfering speakers, etc) have been slightly inspected or committed individually. Therefore, the investigation of more complex techniques that can work with different kinds of distortions at the same time is another interesting research line. These approaches should integrate powerful deep learning models to generalize in these difficult environments, and also advanced signal processing techniques that can model the different distortion sources.

Concerning the perceptual training of DNN-based speech enhancement estimators, the noisy phase remains an unsolved problem. Phase processing is an important current research line, and several approaches are under study, including the use of complex masks, end-to-end models to process the time-domain signals or the use of additional phase correction algorithms. Thus, the investigation of the use of these phase-aware techniques along with the proposed perceptual loss functions is a research area to explore. Among the different alternatives, the use of real-valued filterbanks, which can also be learned during the training phase, is an interesting approach for spectral masking in the new transformed domain.

Appendix A

Resumen

En el presente apéndice se recoge un resumen en castellano de la Memoria de Tesis con el objeto de cumplir con la normativa de elaboración proveniente de la Escuela de Posgrado de la Universidad de Granada. Este resumen se estructura en las siguientes secciones. En primer lugar, las secciones *Introducción*, *Objetivos* y *Estructura de la memoria* se corresponden con el Capítulo 1. A continuación, las secciones *Fundamentos del realce de voz en el dominio de la STFT* y *Marco experimental* se corresponden con los Capítulos 2 y 3, respectivamente. Las contribuciones de esta Tesis se presentan en las secciones *Algoritmo de realce de voz bicanal basado en filtro de Kalman extendido para la estimación de la RTF*, *Realce de voz multicanal mediante un algoritmo recursivo de expectación-maximización con presencia de voz a priori basada en DNN*, *Estimación multicanal del habla del locutor objetivo basada en la arquitectura spatial-beam* y *Función de coste para aprendizaje profundo basada en la evaluación de la calidad perceptual de la voz*, las cuáles se corresponden con los Capítulos 4, 5, 6 y 7, respectivamente. Finalmente, la sección *Conclusiones* corresponde al Capítulo 8.

A.1 Introducción

El habla es seguramente el principal y más relevante método de comunicación entre los humanos. Nos permite expresar nuestras ideas, intercambiar información con otras personas, y es una herramienta fundamental en nuestra sociedad. Las comunicaciones por medio del habla se han visto favorecidas en las últimas décadas gracias al advenimiento de la era de la tecnología de la información y la comunicación. Las tecnologías de difusión, como la radio, la televisión o Internet, permiten acceder rápidamente a la información, que en gran parte se da a través del habla. Las comunicaciones telefónicas permiten conversar con las personas a grandes distancias, y los dispositivos móviles han difundido las comunicaciones humanas de manera ubicua y generalizada. Los servicios de comunicación que utilizan Internet también han cobrado importancia en los últimos años mediante aplicaciones informáticas como Skype o Discord, o incluso aplicaciones móviles como WhatsApp o Telegram, que se utilizan comúnmente, manteniendo una continua interconexión con personas de todo el mundo.

Las mejoras tecnológicas también han propiciado avances en el área de las interacciones hombre-máquina. Buenos ejemplos son los asistentes digitales, que pueden interactuar con los humanos para realizar diferentes tareas. Estos asistentes están incluidos en la mayoría de nuestros dispositivos móviles. Además, nuevos altavoces inteligentes han sido desarrollados por diferentes compañías como en el caso de Amazon Alexa, Apple Siri, o Google Home. Estas comunicaciones hombre-máquina requieren tecnologías de voz como el reconocimiento automático del habla y la síntesis de texto a voz. Otro aspecto importante es la seguridad en las tecnologías del habla, en las que se necesitan métodos robustos de verificación de locutor y de anti-spoofing para garantizar la identidad del usuario (por ejemplo, en el caso de las operaciones en un banco electrónico mediante la biometría de la voz). Además, las tecnologías del habla han resultado de gran utilidad en diferentes servicios de salud, como los dispositivos de ayuda auditiva o las interfaces de habla silenciosa. Como vemos, las tecnologías del habla pueden encontrarse en diferentes aspectos de nuestra vida y se espera que crezcan en los próximos años.

Uno de los principales retos de estas tecnologías del habla es su uso en condiciones en las que la señal de voz se ve afectada por diferentes tipos de distorsión. Estas distorsiones pueden provenir de diferentes fuentes, como el ruido ambiental, que es el principal problema en el uso de dispositivos móviles, o la interferencia de otros hablantes. Otro tipo de distorsiones son las debidas a las propiedades acústicas del entorno, como en el caso de los ecos o la reverberación. Estas diferentes fuentes de distorsión pueden degradar gravemente el rendimiento de las tecnologías descritas anteriormente. Por ejemplo, hacen menos inteligibles las comunicaciones móviles entre los seres humanos y afectan a la calidad perceptiva de la voz, lo que resulta especialmente problemático para los oyentes con problemas de audición. Además, el rendimiento de los sistemas de reconocimiento y verificación disminuye en caso de graves distorsiones de la voz, lo que dificulta su utilización en condiciones difíciles de ruido.

El uso de las tecnologías del habla en condiciones de ruido exige algoritmos de procesamiento de voz de alto rendimiento capaces de mejorar la calidad y la inteligibilidad del habla. Este es el objetivo del realce de voz, que se ocupa del diseño de técnicas para estimar la voz limpia a partir de voz ruidosa y distorsionada. El realce de voz es esencial en muchas tecnologías relacionadas con el habla para proporcionar un buen rendimiento en entornos reales. Por lo tanto, la investigación de diferentes métodos y algoritmos para el realce de voz es un campo crucial y todavía desafiante. Los primeros trabajos en esta área van desde el diseño de algoritmos heurísticos hasta el uso de marcos estadísticos para modelar las propiedades de las señales subyacentes involucradas. El uso de estimadores estadísticos, junto con las suposiciones sobre el ruido, permitió el diseño de técnicas con un rendimiento competitivo, especialmente en condiciones estacionarias de ruido. La investigación de modelos estadísticos más potentes ofrecía soluciones de gran potencial, que podían integrarse en las tecnologías del habla gracias al aumento de las capacidades de computación. Además, los dispositivos actuales han empezado a incorporar arrays de múltiples micrófonos para captar la señal de voz. Esto propicia el interés por los algoritmos de procesamiento de voz multicanal que explotan la información espacial de los diferentes micrófonos, mejorando así la reducción de ruido y logrando una baja distorsión de la voz. La combinación de técnicas multicanal con marcos estadísticos dio resultados del estado del arte en múltiples aplicaciones, especialmente en los sistemas de reconocimiento automático del habla. No obstante, el procesamiento clásico de la señal se enfrenta a limitaciones adicionales en escenarios ruidosos desafiantes, como los que implican ruidos no estacionarios, interferencia entre múltiples locutores y la distorsión debida a entornos reverberantes. En este caso, las suposiciones sobre las señales no son lo suficientemente precisas, por lo que el modelado y, por tanto, el rendimiento final, se degrada. Esto da lugar a señales de voz realzadas con una calidad perceptual pobre y una baja inteligibilidad tanto para los humanos como para las máquinas.

En los últimos años, la revolución del aprendizaje profundo ha cambiado la mayoría de las tecnologías humanas actuales. Este paradigma permite el diseño de algoritmos que pueden ser entrenados para aprender cómo realizar sus tareas directamente de los datos. Esto comparte una similitud, de forma general, con la forma en que los humanos aprenden de su entorno. Hoy en día, las redes neuronales profundas son modelos complejos que incluyen varias capas de no-linealidad y millones de parámetros que deben ser aprendidos. Dos factores principales han favorecido el auge de estos algoritmos de aprendizaje automático. El primero es el aumento de los recursos computacionales, especialmente con las mejoras en las unidades de procesamiento gráfico (GPU por sus siglas en inglés), que permiten el entrenamiento en paralelo de estos modelos en menor cantidad de tiempo. El segundo factor, y probablemente el más importante, es la disponibilidad y variedad de una enorme cantidad de datos de entrenamiento, que mejoran la capacidad de generalización de estas redes. Las redes neuronales profundas se han convertido sin duda en el estado del arte en muchas áreas diferentes, incluido el procesamiento de la voz, donde han superado a los enfoques clásicos. Los actuales algoritmos de reconocimiento, verificación y síntesis de voz se diseñan utilizando redes neuronales profundas debido a su asombroso rendimiento. El aprendizaje profundo también se ha aplicado al campo del realce de voz, proporcionando señales de voz realzadas con una alta calidad perceptual, baja distorsión y poco ruido. Sin embargo, una de las principales críticas a las redes neuronales profundas es que actúan como cajas negras, en las que es casi imposible saber cómo se procesan las señales y cómo aprende el algoritmo. La necesidad de una gran cantidad de parámetros y grandes bases de datos es otra limitación de estas técnicas, ya que no tenemos control sobre cómo estamos dimensionando nuestro problema y la forma en que estos modelos generalizarán en condiciones reales. Por último, estos modelos no requieren de conocimiento para la modelización específica del problema, lo que significa que estamos desperdiciando la experiencia acumulada en el campo de la voz. Este conocimiento puede seguir siendo útil para ciertos problemas o para una mejor comprensión del problema a resolver.

Un último aspecto importante de las tecnologías del habla en los dispositivos inteligentes actuales es que también necesitan asegurar la eficiencia computacional y el procesamiento online (es decir, utilizando información actual y pasada) con baja latencia. Si bien el requisito de algoritmos eficientes es esencial para la integración de estas tecnologías en una amplia variedad de dispositivos, el procesamiento online sigue siendo necesario para aplicaciones que se ejecutan en tiempo real con una calidad de servicio adecuada. El diseño de los algoritmos de procesamiento online suele ser más difícil y su rendimiento suele ser inferior al de las técnicas offline. Por lo tanto, la investigación sobre la mejora de estas técnicas y el uso de algoritmos de baja complejidad es otro punto clave que debe ser estudiado por la comunidad del procesamiento de voz.

A.2 Objetivos

Tal y como hemos introducido, los algoritmos de realce de voz son necesarios en los dispositivos móviles para mejorar la calidad perceptual y la inteligibilidad en condiciones de ruido no estacionario. Los dispositivos actuales incorporan múltiples micrófonos, por lo que la información multicanal también puede ser explotada. Por otra parte, las aplicaciones de voz en los dispositivos móviles tienen que asegurar el procesamiento online con baja latencia y eficiencia computacional. Entre los algoritmos de realce de voz en la bibliografía, el procesamiento clásico de la señal está limitado debido a las hipótesis formuladas sobre las estadísticas de la señal, que a menudo no son realistas, mientras que las redes neuronales profundas son cajas negras que requieren grandes cantidades de datos y parámetros, pudiendo carecer de generalización en entornos reales. Esta Tesis se centra en el desarrollo de técnicas de realce de voz multicanal online adecuadas para los dispositivos móviles. Los algoritmos propuestos están diseñados para integrar el uso del procesamiento estadístico de la señal y de redes neuronales profundas en partes específicas del algoritmo. De este modo, podemos aprovechar el procesamiento de señal multicanal para desarrollar técnicas de realce de voz con alto rendimiento y baja distorsión. Además, se pueden utilizar redes neuronales profundas más eficientes en partes del algoritmo en las que las suposiciones sobre las estadísticas y las propiedades de las señales son débiles. Esto puede mejorar la robustez en entornos ruidosos no estacionarios y reales, al tiempo que permite el procesamiento online. Más precisamente, destacamos los siguientes objetivos, cada uno de ellos centrado en un escenario diferente para aplicar estas técnicas integradoras:

1. Desarrollar algoritmos de realce de voz adecuados para smartphones de doble micrófono en entornos ruidosos y reverberantes. Nuestro objetivo es explotar la relación particular entre la voz limpia y el ruido en ambos sensores, logrando una estimación más precisa de los canales acústicos y las estadísticas de ruido.

2. Estudiar la estimación conjunta de la señal de voz limpia, las diferentes estadísticas de la voz y los parámetros acústicos en un marco de realce de voz multicanal online. La idea es aumentar la robustez bajo ruidos no estacionarios explotando conjuntamente las características espectrales, espaciales y temporales de la señal de voz.

3. Mejorar el rendimiento de los algoritmos de realce de voz en escenarios con múltiples locutores. El objetivo es centrarse en un locutor objetivo utilizando información auxiliar de él, permitiendo así simplificar el problema a un entorno ruidoso.

4. Analizar y evaluar el entrenamiento de las redes neuronales profundas para realce de voz utilizando propiedades perceptuales del sistema auditivo humano. Nuestro objetivo es evaluar las conocidas métricas de calidad objetiva como funciones de entrenamiento, mejorando la calidad percibida por los oyentes humanos.

A.3 Estructura de la memoria

Esta Tesis consta de un total de ocho capítulos y un anexo que incluye el presente resumen en castellano. Tras la introducción recogida en el Capítulo 1, los fundamentos teóricos y una revisión del estado del arte se desarrollan en el Capítulo 2, mientras que el marco experimental se describe en el Capítulo 3. Luego, los Capítulos 4, 5, 6 y 7 se dedican a describir nuestras contribuciones en el campo del realce de voz multicanal online. Cada capítulo desarrolla uno de los objetivos previamente enumerados de esta Tesis. Por último, en el Capítulo 8 se resumen las conclusiones finales. Más específicamente:

- En el Capítulo 1 se exponen las tres primeras secciones de este apéndice.

- En el Capítulo 2 se hace una revisión de la bibliografía sobre el realce de voz para presentar los fundamentos teóricos de esta Tesis. Primero, introducimos el análisis y el procesamiento de la señal de voz ruidosa en el dominio tiempo-frecuencia usando la transformada de Fourier de tiempo corto. A continuación, se revisan los algoritmos monocanal basados en el procesamiento clásico de la señal, destacando el problema de la estimación del ruido. Posteriormente, se explican los enfoques de realce de voz multicanal basados en algoritmos de beamforming junto con el uso de técnicas de postfiltrado y la estimación de los parámetros acústicos necesarios. Finalmente, se hace un repaso del uso de las redes neuronales profundas para el realce de voz, se resumen las arquitecturas de red más comunes y se examina el uso de estos modelos para el realce de voz monocanal y multicanal.

- En el Capítulo 3 se describe el marco experimental utilizado en esta Tesis. Esto incluye las bases de datos de voz ruidosa y las métricas de calidad objetiva utilizadas para el entrenamiento y evaluación de las contribuciones propuestas. Además, detallamos la configuración seguida en el entrenamiento de las redes neuronales profundas que se integran en nuestras propuestas.

- En el Capítulo 4 se propone un algoritmo de realce de voz destinado a smartphones de doble micrófono. Este enfoque explota la información bicanal y el modo de uso del smartphone para obtener parámetros acústicos del modelo más precisos. En primer lugar, hacemos una descripción general de nuestro enfoque, que se basa en una arquitectura de beamforming y postfiltrado. Luego, describimos nuestro propuesta de filtro de Kalman extendido para seguir la variabilidad temporal de la respuesta acústica entre los micrófonos. Por último, la estimación del ruido se aborda utilizando la probabilidad de presencia de voz. Se consideran dos aproximaciones: modelos espaciales estadísticos y redes neuronales profundas. Las propuestas se evalúan en una base de datos bicanal con ruido y reverberación obtenida de un smartphone utilizado en posiciones de habla cercana y lejana.

- En el Capítulo 5 se propone un marco recursivo de expectación-maximización para realce de voz multicanal online. Este marco permite la estimación conjunta de la señal de voz limpia, la probabilidad de presencia de voz y los diferentes parámetros acústicos de forma iterativa, mejorando la robustez en entornos ruidosos no estacionarios. En primer lugar, se utiliza un beamformer para explotar la información espacial de las señales de voz ruidosas. Luego, un postfiltro de Kalman utiliza las correlaciones temporales en la señal de voz limpia para mejorar la reducción de ruido. La probabilidad de presencia de voz se estima usando una aproximación que combina un modelo espacial estadístico con un estimador de máscaras basado en redes neuronales profundas. Por último, las estadísticas estimadas se utilizan para la estimación de máxima verosimilitud de los parámetros acústicos del modelo. Nuestra propuesta se evalúa en una base de datos de voz ruidosa multicanal grabada con una tablet en diferentes entornos reales.

- En el Capítulo 6 se describe nuestra técnica para la separación del hablante objetivo en un escenario de múltiples locutores. Este enfoque permite centrarse en un locutor utilizando un estimador de máscaras basado en redes neuronales profundas que integra información auxiliar sobre el locutor deseado. Para ello, la red se mejora con bloques adicionales que explotan las características espectrales y espaciales del hablante. El estimador de máscaras se utiliza junto con un beamformer por bloques online, el cuál se inicia con la información contextual para mejorar la convergencia del sistema. La propuesta se evalúa para reconocimiento automático del habla en escenarios con múltiples locutores.
- En el Capítulo 7 se propone una función de pérdidas para aprendizaje profundo basada en la evaluación perceptutal de la calidad de la voz. Esta función de pérdidas se deriva del algoritmo PESQ para la evaluación perceptual de la calidad de la voz, el cuál es una conocida métrica objetiva de calidad perceptual. Esta propuesta está destinada al entrenamiento de redes neuronales profundas utilizando consideraciones perceptuales, permitiendo de esta forma la mejora en la calidad de voz percibida por los oyentes humanos. Nuestra propuesta se evalúa para el realce de voz monocanal basado en redes neuronales profundas. Consideramos las dos aproximaciones más comunes: el mapeo espectral y el enmascaramiento espectral.
- Finalmente, las conclusiones de esta Tesis se presentan en el Capítulo 8 junto con un resumen de nuestras contribuciones y trabajos futuros.

A.4 Fundamentos del realce de voz en el dominio de la STFT

En este capítulo hemos introducido los fundamentos del realce de voz en el dominio de la transformada de Fourier de tiempo corto (STFT por sus siglas en inglés). Estos fundamentos sirven de base teórica para las diferentes contribuciones que se presentan más adelante en esta Tesis. En primer lugar, hemos revisado la técnica STFT para el procesamiento de las señales de voz en el dominio del tiempo. También hemos explicado cómo reconstruir la señal de voz realzada en eñ dominio del tiempo usando la STFT inversa (ISTFT).

Luego, se presentaron los principales algoritmos clásicos de realce de voz monocanal, remarcando su implementación como funciones de ganancia de valor real en el dominio de tiempo-frecuencia. Hemos cubierto los algoritmos de sustracción espectral, el filtrado de Wiener, y los modelos basados en estimadores Bayesianos. En esta última categoría, nos hemos centrado en los estimadores de mínimo error cuadrático medio (MMSE por sus siglas en inglés) de la amplitud de voz y la estimación a priori de la relación señal a ruido (SNR por sus siglas en inglés). Estos métodos clásicos requieren una estimación de las estadísticas de ruido, por lo que también hemos introducido los estimadores clásicos de ruido de monocanal: desde métodos simples de detección de voz activa, hasta el seguimiento del ruido usando técnicas de estadísticas mínimas y de promedio recursivo mínimo controlado, y finalmente estimadores estadísticos más avanzados que incluyen estimadores MMSE, de máxima verosimilitud y de máximo a posteriori.

A continuación, se han presentado los algoritmos de realce de voz multicanal. Nos hemos centrado en las técnicas de beamforming, especialmente en aquellos que se formulan a partir de las estadísticas de las señales subyacentes. Primero se ha presentado el beamformer de varianza mínima en la respuesta sin distorsión (MVDR por sus siglas en inglés), y luego lo hemos ampliado al enfoque de filtro de Wiener multicanal y al uso de técnicas de postfiltrado. También se ha abarcado la estimación de los parámetros acústicos importantes para el beamformer, como la matriz espacial de covarianzas del ruido y la función de transferencia relativa (RTF por sus siglas en inglés) entre canales acústicos, incluyendo la estimación conjunta mediante el algoritmo de expectación-maximización (EM).

Por último, se han introducido los enfoques basados en redes neuronales profundas (DNN por sus siglas en inglés) para el realce de voz. En primer lugar se han presentado las principales arquitecturas utilizadas para el procesamiento de voz, incluidas las arquitecturas feedforward, recurrentes y convolucionales. Luego, hemos explicado los dos enfoques principales para el realce de voz monocanal basado en DNNs en el dominio de la STFT. Estos son el mapeo espectral, que trata directamente de estimar el espectro de amplitud, y el enmascaramiento espectral, que estima una función de ganancia. En cuanto a este último enfoque, se han descrito las diferencias entre la aproximación de máscaras y la de señal. En la última parte de la sección se ha presentado la integración de las DNNs junto con las técnicas de beamforming. Los principales enfoques descritos han sido el uso de estimadores de máscaras basados en DNNs para la estimación de los parámetros acústicos, la integración de las DNNs en un algoritmo EM y la implementación directa del beamformer utilizando arquitecturas de redes neuronales.

A.5 Marco experimental

En este capítulo hemos introducido el marco experimental utilizado para los diferentes resultados experimentales obtenidos en esta Tesis. En primer lugar, hemos descrito las características de las bases de datos de voz ruidosa utilizadas para entrenamiento y evaluación. Estas bases de datos incluyen grabaciones ruidosas monocanal y multicanal utilizando diferentes dispositivos en entornos reverberantes y ruidosos. También se incluye el caso de múltiples locutores interfiriendo. A continuación, hemos presentado las métricas objetivas utilizadas para evaluar y comparar los diferentes métodos. Estas métricas incluyen la métrica de evaluación de la calidad de voz perceptual (PESQ por sus siglas en inglés), la métrica de inteligibilidad objetiva en tiempo corto (STOI por sus siglas en inglés) y su versión extendida (ESTOI), y la métrica de ratio en la distorsión de voz (SDR por sus siglas en inglés) y el índice de distorsión de señal. Además, la tasa de error de reconocimiento (WER por sus siglas en inglés) también se utilizó para evaluar la precisión de los sistemas de reconocimiento automático del habla, en los que el algoritmo realce de voz se utiliza como front-end.

Para concluir este capítulo, también hemos dado los detalles de la configuración utilizada para el entrenamiento de las diferentes arquitecturas DNN que integran los algoritmos propuestos. En primer lugar, hemos descrito el marco de programación utilizado para trabajar con estos modelos de aprendizaje profundo, incluyendo las bibliotecas de aprendizaje profundo disponibles. Luego, hemos explicado el procedimiento de entrenamiento y las técnicas de optimización utilizadas. Por último, se han presentado algunas técnicas de regularización para mejorar el rendimiento del entrenamiento: la técnica de Dropout y el criterio de parada temprana. Aunque también son posibles otras regularizaciones, los métodos empleados han proporcionado una buena convergencia de los modelos durante su entrenamiento.

A.6 Algoritmo de realce de voz bicanal basado en filtro de Kalman extendido para la estimación de la RTF

En este capítulo hemos presentado el marco de realce de voz basado en filtro extendido de Kalman (eKF por sus siglas en inglés) para la estimación de la RTF, o estimador eKF-RTF. Este estimador está destinado a su uso en smartphones de doble micrófono utilizados en condiciones de habla cercana y lejana. Este marco utiliza un beamformer MVDR seguido de un postfiltro que explota la información bicanal. Hemos introducido primero el marco teórico y los diferentes pasos involucrados en el procesamiento de la voz ruidosa. Además, hemos descrito los dos postfiltros evaluados en nuestra propuesta: el filtro paramétrico de Wiener (pWF por sus siglas en inglés) y el estimador óptimamente modificado de la amplitud logarítmica espectral (OMLSA por sus siglas en inglés). Estos postfiltros utilizan las estadísticas de voz limpia y ruido monocanal junto con la estimación a posteriori de la probabilidad de presencia de voz (SPP por sus siglas en inglés) para mejorar la reducción de ruido. Además, se han propuesto dos estimadores de varianza de la voz limpia monocanal, basados en diferencia de potencia entre los canales o utilizando una estimación basada en el uso del beamformer MVDR.

El estimador eKF-RTF ha sido entonces presentado para seguir la variabilidad de la RTF entre los micrófonos del smarthpone en ambientes reverberantes. Primero hemos formulado los modelos del espacio de estados tanto para la variabilidad temporal de la RTF como para las observaciones ruidosas en el canal secundario dado el micrófono de referencia. Luego, se utilizó el filtro de Kalman para definir las ecuaciones para la estimación de la RTF en cada trama temporal dadas las predicciones anteriores. El problema de tratar con modelos no lineales se ha abordado mediante el uso de la linealización de de primer orden basada en series de Taylor vectoriales (VTS por sus siglas en inglés). Además, esta aproximación necesita información a priori sobre las estadísticas de la RTF, que se obtuvieron con antelación utilizando un conjunto de entrenamiento de señales de voz bicanal en entornos reverberantes.

Los pasos anteriores requieren de la estimación de las estadísticas de ruido y la SPP en cada valor de tiempo-frecuencia. Para actualizar las estadísticas de ruido se propuso una estimación del ruido basada en la SPP utilizando un promedio recursivo temporal. Se propusieron dos métodos diferentes para la estimación del SPP a posteriori. El primer método utiliza modelos estadísticos espaciales basados en Gaussianas multivariadas. Este modelo espacial también requiere el conocimiento de la probabilidad a priori de ausencia de voz. Por lo tanto, propusimos diferentes estimadores bicanal de esta probabilidad, basados en la diferencia de potencia y la relación espacial entre los micrófonos. Finalmente, abordamos la estimación directa del SPP a posteriori utilizando un estimador de máscaras basado en DNN. Se utilizó una arquitectura convolucional-recurrente y se exploraron características adicionales bicanal (nivel de potencia y diferencia de fase) para mejorar la precisión de la estimación.

Concluimos el capítulo con una evaluación experimental de las diferentes técnicas propuestas. Utilizamos métricas de calidad objetiva y una base de datos de voz ruidosa simulada, grabada por un smartphone de doble micrófono, tanto en condiciones de habla cercana como lejana. Primero evaluamos el rendimiento de los diferentes estimadores de procesamiento de señal propuestos, incluidos los estimadores de ruido basados modelos estadísticos, el estimador eKF-RTF y la estimación de la varianza de la voz limpia monocanal. Los resultados mostraron que nuestros estimadores de ruido basados en la ausencia de voz a priori logran mejores resultados que otros métodos cuando se utilizan con un beamformer MVDR. Nuestro estimador eKF-RTF también mostró una mayor precisión que otros métodos clásicos, con baja distorsión de la señal de voz. En cuanto a los estimadores de varianza de voz monocanal, el basado en el beamformer MVDR superó el enfoque basado en diferencia de potencias en ambas configuraciones de habla. Luego, evaluamos el rendimiento de los postfilteros pWF y OMLSA utilizando estos estimadores, mostrando mejoras con respecto a otros métodos de realce de voz bicanal con respecto a la calidad e inteligibilidad del habla. Finalmente, este postfiltrado también fue evaluado usando la estimación de SPP a posteriori basada en DNN. Se evaluaron diferentes combinaciones de características de entrada a la red bicanal y los resultados se compararon con el enfoque utilizando modelos estadísticos espaciales y las características monocanal para la red. Los resultados mostraron que el estimador de máscaras basado en DNN supera el enfoque basado en modelo estadísticos y que la DNN puede explotar con éxito las características bicanal para mejorar la precisión de la estimación. Concretamente, se prefieren las diferencias de potencia en condiciones de habla cercana, mientras que las diferencias de fase son más discriminatorias en condiciones de habla lejana.

A.7 Realce de voz multicanal mediante un algoritmo recursivo de expectación-maximización con presencia de voz a priori basada en DNN

En este capítulo hemos propuesto un marco recursivo de expectación-maximización (REM) para el realce de voz multicanal online que integra un estimador de presencia de voz basado en DNN. El marco REM estima la señal de voz limpia y el SPP a posteriori durante el paso E, mientras que los parámetros acústicos se calculan durante el paso M. Este procedimiento se repite en cada trama temporal, permitiendo el procesamiento online de la señal de voz ruidosa. La propuesta se formuló utilizando un marco estadístico para la señal de voz ruidosa multicanal bajo la suposición de una RTF multiplicativa. También se consideró la probabilidad de presencia de voz en cada intervalo de tiempo-frecuencia. Además, se definió un modelo de predicción lineal temporal para las amplitudes de voz limpia, el cuál tiene en cuenta la correlación entre las distintos tramas temporales.

A continuación se definió el algoritmo REM a partir de un modelo estadístico, que se formuló previamente utilizando la log-verosimilitud ponderada exponencialmente. En el paso E, las expectativas de primer y segundo orden de la señal de voz limpia se estiman usando un beamformer MVDR seguido de un postfiltrado lineal. Se han considerado dos postfiltros, un filtro de Wiener y un filtro de Kalman, siendo este último capaz de explotar las correlaciones temporales de las amplitudes de voz limpia utilizando un modelo de predicción lineal. También se aplica un enmascaramiento basado en SPP a las estadísticas. Este enmascaramiento sólo se considera en el paso M para evitar las distorsiones en la señal de voz de salida. Además, la SPP a posteriori se estima utilizando una SPP a priori y los modelos espaciales de las señales de voz y de ruido. Por otra parte, la estimación de máxima verosimilitud se utiliza durante el paso M para calcular la RTF, la matriz de covariancas espaciales del ruido y los parámetros de predicción lineal del postfiltro de Kalman. La varianza de voz limpia puede obtenerse utilizando un procedimiento diferente para permitir su rápida adaptación y evitar la distorsión introducida por el SPP. Además, el SPP a priori se estima mediante un estimador de máscaras basado en DNN. La DNN explota las propiedades espectrales de las señales y proporciona una buena inicialización para el SPP. El algoritmo REM puede sufrir algunos problemas durante su convergencia. Por lo tanto, se han tenido en cuenta algunas consideraciones adicionales para el correcto funcionamiento de la implementación.

El capítulo concluye con una evaluación experimental en la que se evaluaron las diferentes versiones del algoritmo REM propuesto utilizando una base de datos multicanal ruidosa grabada con una tablet de seis micrófonos. La evaluación mediante mediciones objetivas mostró que las propuestas superaban a otros enfoques de vanguardia como el MWF y el filtro de Kalman multicanal (MKF por sus siglas en inglés), que también utilizaban la misma DNN para una comparación justa. La aproximación REMKF (basada en filtro de Kalman) también obtuvo mejores resultados que la versión REMWF (basada en filtro de Wiener), lo que demostró la ventaja de explotar las correlaciones temporales junto con las propiedades espectrales y espaciales de las señales. El rendimiento también se analizó utilizando estimaciones oráculo para la SPP a posteriori y los parámetros acústicos. Además, se comparó el estimador basado en DNN con los estimadores clásicos de procesamiento de señal, mostrando una mayor precisión y rendimiento para el algoritmo REM. La integración de las estimaciones de DNN con modelos espaciales estadísticos también logró predicciones más discriminativas de la SPP a posteriori, aumentando así el rendimiento del enfoque REM. Por último, el análisis del número de iteraciones EM mostró que un número bajo de iteraciones es suficiente para lograr un rendimiento competitivo. Además, la baja latencia computacional permite el procesamiento en tiempo real de forma online, lo que hace que el marco REM sea adecuado para las aplicaciones del mundo real en los dispositivos móviles.

A.8 Estimación multicanal del habla del locutor objetivo basada en la arquitectura spatial-beam

En este capítulo hemos descrito la aproximación spatial-beam para la estimación del locutor objetivo en escenarios multicanal con múltiples locutores. Este método se basa en la aproximación speaker-beam (SpkB), que utiliza la información espectral obtenida de una locución de adaptación del locutor objetivo para centrarse en sus características espectrales. La técnica utiliza información espacial adicional para mejorar la capacidad discriminativa del estimador de máscaras basado en DNN. El método se evaluó para una aplicación de reconocimiento automático del habla en escenarios con múltiples hablantes.

En primer lugar, presentamos el algoritmo de beamformer online por bloques de tramas que estima la señal de voz del locutor objetivo. Este método utiliza las máscaras de voz y ruido, estimadas por la DNN, para actualizar las matrices de covarianza espaciales de la voz y del ruido en cada bloque de tramas temporales. Para mejorar la convergencia, proponemos una inicialización para ambas matrices. En el caso del ruido, la inicialización se basa en la suposición de un campo de ruido difuso, mientras que para la voz se usa la información obtenida de la locución de adaptación. Las matrices estimadas se utilizan entonces para calcular los parámetros del beamformer MVDR.

A continuación se describió el estimador de máscaras basado en speaker-beam, introduciendo primero el estimador de máscaras basado en una red neuronal recurrente y su adaptación al enfoque SpkB. Esta adaptación incluye el uso de una capa de adaptación intermedia con múltiples capas ocultas, combinadas mediante el uso de un vector de representación del locutor. Este vector se obtiene a partir de la locución de adaptación mediante el uso de una red neuronal auxiliar. Posteriormente, se presentó la propuesta spatial-beam, describiendo la adaptación al procesamiento online y la utilización de la información espacial obtenida de la locución de adaptación. Se consideraron dos aproximaciones diferentes. La primera aproximación integra un preprocesamiento espacial de la señal de voz ruidosa y la locución de adaptación utilizando un beamformer MVDR offline. La segunda aproximación utiliza características espaciales adicionales calculadas a partir de las señales de voz ruidosa y de adaptación.

Finalmente, las propuestas se evaluaron utilizando una base de datos multicanal simulada con múltiples locutores en salas con reverberación. Evaluamos las diferentes aproximaciones utilizando métricas objetivas de inteligibilidad y distorsión de la voz. Además, probamos el rendimiento de la propuesta como front-end de un sistema de reconocimiento automático del habla. En primer lugar, evaluamos las inicializaciones propuestas para las matrices de correlaciones espaciales, mostrando mejoras en términos de WER. Luego, comparamos las dos variantes de nuestra propuesta spatial-beam con respecto al método SpkB y el uso de redes atractoras profundas. La evaluación se hizo usando beamformer y estimador de máscaras offline. Los resultados mostraron que nuestra propuesta logra mejores resultados de reconocimiento, especialmente la variante de pre-procesamiento espacial. A continuación, evaluamos nuestras propuestas para el procesamiento online, mostrando resultados de reconocimiento competitivos para un escenario con múltiples locutores, especialmente para la variante que emplea características espaciales. Finalmente, comparamos nuestras propuestas y el método SpkB en términos de rendimiento para escenarios con locutores de diferente y mismo género. Los resultados demostraron que nuestras propuesas pueden tratar eficazmente la separación de locutores del mismo género, superando a la técnica SpkB. Esto demuestra que las características espaciales son una fuente de información útil para discriminar entre hablantes con propiedades espectrales similares.

A.9 Función de coste para aprendizaje profundo basada en la evaluación de la calidad perceptual de la voz

En este capítulo hemos propuesto una función de coste para evaluar la calidad perceptual de la voz en los métodos realce de voz basados en DNN. Esta técnica se denomina métrica para la evaluación perceptual de la caldiad de voz (PMSQE por sus siglas en inglés). El método propuesto se basa en la combinación de dos términos de perturbación inspirados en el conocido algoritmo PESQ: las perturbaciones simétrica y asimétrica. Estos términos tienen en cuenta diferentes consideraciones perceptuales en la señal de voz. Para obtener los términos de perturbación, los espectros de potencia de voz limpio y realzado se ecualizan primero a un nivel de escucha estándar y luego se convierten al dominio Bark. El espectro Bark de la señal realzada es entonces ecualizado y ambos espectros Bark son convertidos a un dominio de sonoridad. Se utilizan dos ecualizaciones diferentes para eliminar los efectos no relevantes para la calidad perceptual: ecualización de frecuencia y ecualización de ganancia. Los vectores de perturbación se obtienen de los espectros de sonoridad, y los términos finales de perturbación se calculan finalmente utilizando normas ponderadas en cada trama temporal.

El método PMSQE puede integrarse como una función de coste para el realce de voz basado en DNN. En el caso del mapeo espectral, se eligió la aproximación del dominio log-espectral y se combinaron las funciones de coste PMSQE y log-MSE para asegurar una buena convergencia sin artefactos en la voz. En el caso del enmascaramiento espectral, la función de coste basada en PMSQE se formuló sin necesidad de términos adicionales.

También se presentó la técnica de aprendizaje multiobjetivo para la combinación de PMSQE con otras funciones de coste para el enmascaramiento espectral.

La aproximación de mapeo espectral se evaluó utilizando una DNN feedforward. Primero se optimizaron los hiperparámetros de la función de coste y la combinación de ecualizaciones. Luego, la función propuesta, log-PMSQE, fue evaluada usando métricas de calidad objetivas y comparada con el log-MSE y otras funciones de coste relacionadas con la percepción. La propuesta mostró mejoras en términos del rendimiento de la calidad perceptual cuando se evaluó la métrica PESQ, y un rendimiento competitivo en distorsión de voz. A continuación, se llevó a cabo un análisis subjetivo con oyentes humanos utilizando el procedimiento de puntuación de opinión media comparativa (CMOS por sus siglas en inglés) para confirmar los resultados obtenidos utilizando métricas objetivas. Los resultados subjetivos mostraron que los participantes tenían una clara preferencia por la calidad perceptual obtenida por nuestra propuesta.

Finalmente, la aproximación de enmascaramiento espectral fue evaluada usando un estimador de máscaras basado en una DNN convolucional-recurrente. La evaluación se llevó a cabo utilizando diferentes métricas objetivas para la calidad e inteligibilidad del habla. PMSQE se analizó primero en términos del rendimiento obtenido para sus diferentes ecualizaciones. Los resultados mostraron que ambas ecualizaciones, incluyendo el suavizado de la ganancia, contribuyen a un aumento del rendimiento del PESQ con una distorsión de la voz aceptable. PMSQE fue entonces comparada con otras funciones de coste relacionadas como MSE, ESTOI y SDR invariante a la escala (SI-SDR por sus siglas en inglés). PMSQE superó a las otras aproximaciones en términos de PESQ, pero sufrió una degradación en las otras métricas. Finalmente, evaluamos la combinación de PMSQE con las otras funciones de coste. Los resultados mostraron que el uso de PMSQE junto con SI-SDR proporciona un buen rendimiento y resultados competitivos en las diferentes métricas objetivas. Así pues, la propuesta SDR+PMSQE fue la mejor opción para el entrenamiento de los métodos de enmascaramiento espectral basados en DNN entre los diferentes métodos evaluados.

A.10 Conclusiones

Se pueden extraer varias conclusiones de todo el trabajo desarrollado en esta Tesis. Algunas de las más relevantes se enumeran a continuación:

- La integración del procesamiento estadístico de señales clásico y los estimadores basados en redes neuronales profundas ha demostrado ser una herramienta con grandes capacidades para el realce de voz. Las técnicas propuestas superan a otros enfoques que utilizan sólo una de las aproximaciones anteriores o no las integran convenientemente.

Esta estrategia combinada ha permitido el diseño de algoritmos de realce de voz con un alto control de los diferentes pasos del algoritmo de procesamiento de voz. Así pues, en los pasos en los que las hipótesis sobre las estadísticas de la señal pueden ser difíciles de justificar debido a la complejidad del problema (como el caso de los entornos no estacionarios para la estimación del ruido o la incertidumbre en la presencia de voz) podemos explotar la capacidad de modelización de las técnicas de aprendizaje profundo. Por lo tanto, podemos evitar las limitaciones impuestas por el marco estadístico, mejorando así el rendimiento final.

- La información bicanal proporcionada por los smartphones de doble micrófono puede explotarse con éxito para mejorar el rendimiento del realce de voz en estos dispositivos móviles. Aunque el rendimiento del beamforming es limitado cuando el número de micrófonos es pequeño, el uso de un postfiltro adecuado puede proporcionar un rendimiento competitivo con respecto a otras técnicas bicanal para smartphones. Este postfiltro aprovecha las buenas estimaciones obtenidas de la información multicanal. Además, las propiedades de los dos canales también pueden explotarse para una estimación precisa de los parámetros acústicos. En el caso del RTF, las estadísticas a priori y el modelo del filtro de Kalman propuesto explotan con éxito la relación entre las señales de voz limpia en ambos micrófonos. Para la estimación del ruido, la diferencia de nivel de potencia entre los micrófonos proporciona un buen método para identificar los segmentos de ruido en condiciones de habla cercana. Por otra parte, la información espacial sobre el campo de ruido, contenida principalmente en la fase de la señal, puede ser un buen elemento discriminatorio en condiciones de habla lejana. Estas características multicanal también pueden ser explotadas por un estimador de máscaras basado en DNN.

- La señal de voz en el dominio de la STFT muestra importantes correlaciones temporales que pueden ser explotadas, especialmente cuando se diseñan aproximaciones online. En cuanto al procesamiento estadístico de la señal, el filtro de Kalman ha demostrado ser un marco útil para modelar estas relaciones temporales. El estimador eKF propuesto ha proporcionado estimaciones más precisas y mejores capacidades de seguimiento del RTF entre los micrófonos de un smartphone que otras aproximaciones. Además, el postfiltro de Kalman se ha utilizado para aprovechar con éxito las correlaciones entre las amplitudes de voz limpia, mejorando la reducción de ruido cuando se utiliza en combinación con un beamformer MVDR. Estas propiedades temporales también se han explotado mediante el uso de redes neuronales recurrentes gracias a su capacidad para el modelado temporal cuando se trabaja con espectrogramas.

- La disponibilidad de un conocimiento preciso sobre la probabilidad de presencia de voz en el dominio de la STFT fue un elemento crucial en el desempeño de los algoritmos propuestos. Esta información puede utilizarse para obtener estimaciones más precisas del ruido o incluso para mejorar la reducción del ruido en ausencia de voz. Entre los diferentes enfoques, los estimadores de máscaras basados en DNN han mostrado un rendimiento asombroso en la estimación precisa del SPP. Estos modelos de aprendizaje profundo tienen una mejor capacidad de discriminación entre los segmentos dominandos por voz y los dominados por ruido, y han demostrado tener éxito en este tipo de tareas de clasificación binaria. El uso de los estimadores SPP basados en DNN tiene un impacto importante cuando nos enfrentamos a ruidos altamente no estacionarios, ya que pueden adaptarse rápidamente a los cambios de las estadísticas de la señal.

- El algoritmo REM propuesto para realce de voz multicanal online ha demostrado un mejor rendimiento en reducción de ruido y distorsión de voz que las técnicas del estado del arte relacionadas. El uso de la estrategia de beamforming más postfiltrado se mejora mediante la estimación conjunta de la señal de voz limpia, la probabilidad de presencia de voz y los diferentes parámetros acústicos involucrados. Esto permite una retroalimentación positiva entre la estimación de las estadísticas de voz y el cálculo de los parámetros acústico del modelo de señal en el procedimiento iterativo. Una parte importante de la propuesta es el uso de las SPP a priori dadas por un estimador de máscaras basado en DNN. Estas SPP a priori contribuyen a la convergencia del algoritmo, permitiendo discriminar mejor entre las componentes del espectro donde la voz est´a ausente o presente. Además, el marco estadístico puede utilizarse para perfeccionar estas estimaciones del SPP mediante el empleo de un modelo espacial estadístico complementario. reconocimiento automático del habla.

- El diseño de algoritmos de procesamiento de voz para escenarios donde se superponen varios locutores es una tarea difícil donde incluso los modelos discriminativos basados en aprendizaje profundo tienen dificultades. El uso de múltiples micrófonos junto con información auxiliar del locutor ha demostrado ser un buen mecanismo cuando estamos interesados en un solo hablante. En este caso, podemos diseñar mejores estimadores de máscaras basados en DNN que exploten tanto la información espectral como espacial de la voz ruidosa multicanal y las señales auxiliares. Esto permite una mayor precisión en las máscaras de voz del locutor objetivo, que pueden utilizarse para el cálculo del beamformer. Si bien la información espectral ayuda a separar a los hablantes con diferentes patrones de frecuencia, la información espacial puede ser más decisiva para los hablantes separados espacialmente que tienen características espectrales similares, como en el caso de locutores del mismo género.

- El realce de voz basado en DNN para escenarios monocanal es un tema actual en la investigación del procesamiento de voz. Una aproximación para mejorar la calidad perceptual y la inteligibilidad de voz proporcionada por estos estimadores es el uso de funciones de coste basadas en consideraciones perceptuales sobre el sistema auditivo humano. Entre los diferentes enfoques, la integración de métricas objetivas como funciones de pérdida de entrenamiento ha demostrado un rendimiento competitivo. Así, nuestra propuesta PMSQE ha mostrado mejoras en el rendimiento de la calidad perceptual con mejores resultados de PESQ. Estas mejoras también se traducen en una mejor calidad percibida por los oyentes humanos. Además, la combinación de diferentes funciones de coste puede dar lugar a estimadores que proporcionen buenos resultados entre diferentes métricas de calidad, superando así las limitaciones de los enfoques independientes.

- Entre las dos aproximaciones más comunes para el realce de voz basado en DNN monocanal, el enmascaramiento espectral parece ser mejor opción que el mapeo espectral. En primer lugar, las mejoras y las capacidades de generalización de los estimadores de máscaras basados en DNN son mayores que los métodos de mapeo espectral cuando se comparan los resultados obtenidos. Además, el entrenamiento basado en métricas objetivas, como en el caso del PMSQE, se ve favorecido cuando se emplean estos estimadores de máscaras. La estimación de la máscara es una tarea más fácil para las DNNs, especialmente cuando el rango de valores de estas máscaras está bien definido. Además, la máscara decide principalmente cómo atenuar la señal en cada componente espectral en términos de la SNR. Esta es, hasta cierto punto, una tarea similar a la predicción de las máscaras SPP, donde las DNNa han demostrado su idoneidad. Por último, los valores restringidos de las máscaras dificultan la presencia de artefactos en la voz cuando se entrena con estas funciones de coste. Esto permite el uso de funciones de coste más sofisticadas con una mejor convergencia durante el entrenamiento.

- Finalmente, nos hemos centrado en el diseño de algoritmos de procesamiento online para el realce de voz. La implementación de estos algoritmos es una tarea desafiante, ya que pueden sufrir fácilmente problemas de convergencia y degradación del rendimiento durante las primeros tramas temporales. Además, sólo pueden utilizar información pasada sobre la señal, por lo que los cambios rápidos en las estadísticas de la señal pueden afectar severamente la calidad de la señal realzada. No obstante, el uso de

DNNs integradas con técnicas adecuadas de procesamiento de voz online puede ayudar a superar estas limitaciones y obtener un buen rendimiento en entornos difíciles con ruidos no estacionarios. Además, el diseño de algoritmos de baja latencia y más ligeros computacionalmente es una característica necesaria para la aplicación de estas técnicas en pequeños dispositivos móviles que puedan utilizarse en condiciones reales.

Las diferentes contribuciones técnicas resultantes de nuestro trabajo se resumen a continuación:

- Un marco experimental para la simulación de señales de voz de bicanañ con smartphones usados en ambientes ruidosos y reverberantes, ya sea en condiciones de habla cercana o lejana [144].
- Un estimador de la función de transferencia relativa entre dos micrófonos usando un filtro de Kalman extendido [144], el cuál usa información a priori de la RTF junto con información temporal y espacial para seguir la variabilidad de los canales acústicos.
- Un algoritmo de realce de voz bicanal basado en beamforming más postfiltrado [145, 146], el cuál utiliza el estimador eKF-RTF junto con mejores estimaciones de las estadísticas de voz limpia para el postfiltro empleando información bicanal.
- Un estimador de ruido basado en SPP para smartphones de doble micrófono, ya sea usando modelos estadísticos espaciales con estimadores de la probabilidad de ausencia de voz a priori basados en información bicanal [146], o estimaciones del SPP basadas en DNN que usan características bicanal [147].
- Un algoritmo REM para la integración del realce de voz multicanal online integrando la estimación del SPP usando una DNN [143].Esta aproximación estima conjuntamente la señal de voz limpia, el SPP a posteriori y los parámetros acústicos, y además explota las correlaciones temporales utilizando un postfiltrado de Kalman.
- Un algoritmo de detección de locutor multicanal para escenarios con múltiples locutores utilizando beamforming onine por bloques y un estimador de máscaras basado en DNN, la cuál es adaptada para explotar la información auxiliar espectral y espacial, permitiendo discriminar mejor el locutor objetivo [142].
- Una función de coste para aprendizaje profunda basada en la evaluación perceptual de la calidad de voz [141], que puede ser usada para entrenar los algoritmos de procesamiento de voz basados en DNN. Evaluamos esta aproximación para el realce de voz monocanal basado en DNN.

En esta Tesis hemos presentado diferentes contribuciones que combinan el procesamiento estadístico de señales y las redes neuronales profundas. Sin embargo, estas integraciones se suelen hacer de forma independiente, cambiando los bloques de procesamiento de señal por estimadores basados en DNN para mejorar las partes más débiles del algoritmo. Por lo tanto, como trabajo futuro, sería interesante investigar la integración completa de las técnicas de procesamiento de señal como modelos de redes neuronales profundas. Así pues, podrían aplicarse diferentes bloques de procesamiento de la voz, como filtros o técnicas de beamforming, utilizando nuevas arquitecturas convolucionales o recurrentes. Estas redes podrían aprender las estadísticas importantes a partir de los datos, manteniendo al mismo tiempo el conocimiento sobre el procesamiento realizado en cada bloque. Este enfoque parece muy prometedor para lograr sistemas con una carga computacional menor y con buenas capacidades de generalización.

Por otra parte, los métodos propuestos están diseñados para tratar principalmente el ruido ambiental aditivo, mientras que otros tipos de distorsiones (ecos, reverberaciones, locutores que interfieren, etc.) se han tratado ligeramente o de forma individual. Por lo tanto, la investigación de técnicas más complejas que puedan trabajar con diferentes tipos de distorsiones al mismo tiempo es otra línea de investigación interesante. Estas aproximaciones deberían integrar modelos de aprendizaje profundo para generalizar en estos entornos difíciles, y también técnicas avanzadas de procesamiento de señal que puedan modelar las diferentes fuentes de distorsión.

En lo que respecta al entrenamiento perceptual de los estimadores de realce de voz basados en DNN, la fase ruidosa sigue siendo un problema sin resolver. El procesamiento de la fase es una importante línea de investigación actual, y se están estudiando varios enfoques, entre ellos el uso de máscaras complejas, modelos extremo a extremo para procesar directamente las señales en el dominio temporal o el uso de algoritmos adicionales de corrección de fase. Así pues, la investigación del uso de estas técnicas de corrección de fase junto con las funciones de coste perceptuales propuestas es un área de investigación que hay que explorar. Entre las diferentes alternativas, la utilización de bancos de filtros, que también pueden aprenderse durante la fase de entrenamiento, es un enfoque interesante para el enmascaramiento espectral en el nuevo dominio transformado.

References

[1] (1990). ITU-R. Rec. BS. 562-3, Subjective assessment of sound quality. Technical report, International Telecommunication Union-Radiocommunication Sector.

[2] (1996). ITU-T. Rec. P.800, Methods for Subjective Determination of Transmission Quality - Series P: Telephone Transmission Quality; Methods for Objective and Subjective Assessment of Quality. Technical report, International Telecommunication Union-Telecommunication Standarisation Sector.

[3] (2001). ITU-T. Rec. P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assesment of narrow-band telephone networks and speech codecs. Technical report, International Telecommunication Union-Telecommunication Standarisation Sector.

[4] (2007). ITU-T. Rec P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codec. Technical report, International Telecommunication Union-Telecommunication Standarisation Sector.

[5] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[6] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 22(10):1533–1545.

[7] Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238.

[8] Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

[9] Allen, J. B. and Rabiner, L. R. (1977). A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564.

[10] Batina, I., Jensen, J., and Heusdens, R. (2006). Noise power spectrum estimation for speech enhancement using an autoregressive model for speech power spectrum dynamics. In *Proc. of 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, pages 1064–1067.

[11] Benesty, J., Chen, J., and Huang, Y. (2008). *Microphone Array Signal Processing*, volume 1. Springer.

[12] Benesty, J., Chen, J., Pan, C., et al. (2016). *Fundamentals of Differential Beamforming*. Springer.

[13] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[14] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120.

[15] Braun, S. and Habets, E. (2018). Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(6):1115–1125.

[16] Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proc. of the IEEE*, 57(8):1408–1418.

[17] Cappé, O. and Moulines, E. (2009). Online EM algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

[18] Chai, L., Du, J., Liu, Q.-F., and Lee, C.-H. (2019). Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):1919–1931.

[19] Chai, L., Du, J., and Wang, Y. (2017). Gaussian density guided deep neural network for single-channel speech enhancement. In *Proc. of 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

[20] Chakrabarty, S. and Habets, E. (2019). Time-frequency masking based online multichannel speech enhancement with convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):787–799.

[21] Chen, J., Benesty, J., Huang, Y., and Doclo, S. (2006). New insights into the noise reduction Wiener filter. *IEEE Transactions on audio, speech, and language processing*, 14(4):1218–1234.

[22] Chen, Z., Luo, Y., and Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250.

[23] Chen, Z., Xiao, X., Yoshioka, T., Erdogan, H., Li, J., and Gong, Y. (2018). Multichannel overlapped speech recognition with location guided speech extraction network. In *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 558–565.

[24] Chinaev, A., Haeb-Umbach, R., Taghia, J., and Martin, R. (2013). Improved single-channel nonstationary noise tracking by an optimized MAP-based postprocessor. In *Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7477–7481.

[25] Chinaev, A., Krueger, A., Vu, D. H. T., and Haeb-Umbach, R. (2012). Improved noise power spectral density tracking by a MAP-based postprocessor. In *Proc. of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4041–4044.

[26] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[27] Cohen, I. (2003). Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5):466–475.

[28] Cohen, I. (2004). Relative transfer function identification using speech signals. *IEEE Transactions on Speech and Audio Processing*, 12(5):451–459.

[29] Cohen, I. and Berdugo, B. (2001). Speech enhancement for non-stationary noise environments. *Signal Processing*, 81(11):2403–2418.

[30] Cohen, I. and Berdugo, B. (2002). Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE signal processing letters*, 9(1):12–15.

[31] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[32] Delcroix, M., Zmolikova, K., Kinoshita, K., Ogawa, A., and Nakatani, T. (2018). Single channel target speaker extraction and recognition with speaker beam. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5554–5558.

[33] Dini, D. H. and Mandic, D. P. (2012). Class of widely linear complex Kalman filters. *IEEE Transactions on Neural Networks and Learning Systems*, 23(5):775–786.

[34] Ditter, D. and Gerkmann, T. (2020). A multi-phase gammatone filterbank for speech separation via TasNet. In *Proc. of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 36–40.

[35] Doclo, S. and Moonen, M. (2002). GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, 50(9):2230–2244.

[36] Doclo, S. and Moonen, M. (2007). Superdirective beamforming robust against microphone mismatch. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):617–631.

[37] Doclo, S., Spriet, A., Wouters, J., and Moonen, M. (2005). Speech Distortion Weighted Multichannel Wiener Filtering Techniques for Noise Reduction. In *Speech Enhancement*, page 199–228. Springer.

[38] Drude, L., Heitkaemper, J., Boeddeker, C., and Haeb-Umbach, R. (2019). SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition. *arXiv preprint arXiv:1910.13934*.

[39] Ducharme, G. R., Lafaye de Micheaux, P., and Marchina, B. (2016). The complex multinormal distribution, quadratic forms in complex random vectors and an omnibus goodness-of-fit test for the complex normal distribution. *Annals of the Institute of Statistical Mathematics*, 68(1):77–104.

[40] Duong, N., Vincent, E., and Gribonval, R. (2010). Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840.

[41] Dvorkind, T. G. and Gannot, S. (2005). Time difference of arrival estimation of speech source in a noisy and reverberant environment. *Signal Processing*, 85(1):177–204.

[42] Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121.

[43] Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error-log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445.

[44] Erdogan, H., Hershey, J. R., Watanabe, S., and Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proc. of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712.

[45] Erdogan, H., Hershey, J. R., Watanabe, S., Mandel, M. I., and Le Roux, J. (2016). Improved MVDR beamforming using single-channel mask prediction networks. In *Proc. of 17th Annual Conference of the International Speech Communication (InterSpeech)*, pages 1981–1985.

[46] Esch, T. and Vary, P. (2009). Efficient musical noise suppression for speech enhancement systems. In *Proc. of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4409–4412.

[47] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874.

[48] Freudenberger, J., Stenzel, S., and Venditti, B. (2009). A noise PSD and cross-PSD estimation for two-microphone speech enhancement systems. In *Proc. of 2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 709–712.

[49] Fu, S., Wang, T., Tsao, Y., Lu, X., and Kawai, H. (2018). End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(9):1570–1584.

[50] Fu, S.-W., Liao, C.-F., and Tsao, Y. (2019a). Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality. *IEEE Signal Processing Letters*, 27:26–30.

[51] Fu, S.-W., Liao, C.-F., Tsao, Y., and Lin, S.-D. (2019b). MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *Proc. of 36th International Conference on Machine Learning (ICML)*, pages 3566–3576.

[52] Fu, S.-W., Tsao, Y., and Lu, X. (2016). SNR-aware convolutional neural network modeling for speech enhancement. In *Proc. of 17th Annual Conference of the International Speech Communication (InterSpeech)*, pages 3768–3772.

[53] Fu, S.-W., Tsao, Y., Lu, X., and Kawai, H. (2017). Raw waveform-based speech enhancement by fully convolutional networks. In *Proc. of 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 6–12.

[54] Fukuda, T., Ichikawa, O., and Nishimura, M. (2010). Long-term spectro-temporal and static harmonic features for voice activity detection. *IEEE Journal of Selected Topics in Signal Processing*, 4(5):834–844.

[55] Gannot, S., Burshtein, D., and Weinstein, E. (2001). Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626.

[56] Gannot, S. and Cohen, I. (2004). Speech enhancement based on the general transfer function GSC and postfiltering. *IEEE Transactions on Speech and Audio Processing*, 12(6):561–571.

[57] Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730.

[58] Garofalo, J., Graff, D., Paul, D., and Pallett, D. (2007). CSR-I (WSJ0) Complete. *Linguistic Data Consortium, Philadelphia*.

[59] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1988). Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, 107:16.

[60] Gerkmann, T., Breithaupt, C., and Martin, R. (2008). Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Transactions on Speech and Audio Processing*, 16(5):910–919.

[61] Gerkmann, T. and Hendriks, R. C. (2011). Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1383–1393.

[62] Gerkmann, T. and Hendriks, R. C. (2012). Improved MMSE-based noise PSD tracking using temporal cepstrum smoothing. In *Proc. of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 105–108.

[63] Gerkmann, T., Krawczyk-Becker, M., and Le Roux, J. (2015). Phase processing for single-channel speech enhancement: History and recent advances. *IEEE Signal Processing Magazine*, 32(2):55–66.

[64] Gerven, S. V. and Xie, F. (1997). A comparative study of speech detection methods. In *Proc of. 5th European Conference on Speech Communication and Technology*.

[65] Griffiths, L. and Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on antennas and propagation*, 30(1):27–34.

[66] Gu, F., Zhang, H., Wang, W., and Wang, S. (2017). An expectation-maximization algorithm for blind separation of noisy mixtures using gaussian mixture model. *Circuits, Systems, and Signal Processing*, 36(7):2697–2726.

[67] Habets, E., Gannot, S., and Cohen, I. (2006). Dual-microphone speech dereverberation in a noisy environment. In *Proc. IEEE International Symposium on Signal Processing and Information Technology*, pages 651–655.

[68] Haeb-Umbach, R., Watanabe, S., Nakatani, T., Bacchiani, M., Hoffmeister, B., Seltzer, M. L., Zen, H., and Souden, M. (2019). Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal Processing Magazine*, 36(6):111–124.

[69] Han, K. and Wang, D. (2012). A classification based approach to speech segregation. *The Journal of the Acoustical Society of America*, 132(5):3475–3483.

[70] Han, W., Zhang, X., Min, G., Sun, M., and Yang, J. (2016a). Joint optimization of audible noise suppression and deep neural networks for single-channel speech enhancement. In *Proc. of 2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

[71] Han, W., Zhang, X., Min, G., Zhou, X., and Zhang, W. (2016b). Perceptual weighting deep neural networks for single-channel speech enhancement. In *Proc. of 2016 12th World Congress on Intelligent Control and Automation (WCICA)*, pages 446–450.

[72] Heitkaemper, J., Heymann, J., and Haeb-Umbach, R. (2018). Smoothing along frequency in online neural network supported acoustic beamforming. In *Speech Communication; 13th ITG-Symposium*.

[73] Hendriks, R. and Gerkmann, T. (2012). Noise correlation matrix estimation for multi-microphone speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):211–221.

[74] Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). MMSE based noise psd tracking with low complexity. In *Proc. of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4266–4269.

[75] Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *Proc. of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35.

[76] Heymann, J., Drude, L., and Haeb-Umbach, R. (2016a). Neural network based spectral mask estimation for acoustic beamforming. In *Proc. of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200.

[77] Heymann, J., Drude, L., and Haeb-Umbach, R. (2016b). Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition. In *Proc. of the 4th International Workshop on Speech Processing in Everyday Environments (CHiME16)*, pages 12–17.

[78] Heymann, J., Drude, L., and Haeb-Umbach, R. (2017). A generic neural acoustic beamforming architecture for robust multi-channel speech processing. *Computer Speech and Language*, 46:374–385.

[79] Higuchi, T., Ito, N., Araki, S., Yoshioka, T., Delcroix, M., and Nakatani, T. (2017). Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):780–793.

[80] Higuchi, T., Ito, N., Yoshioka, T., and Nakatani, T. (2016). Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In *Proc. of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5210–5214.

[81] Higuchi, T., Kinoshita, K., Ito, N., Karita, S., and Nakatani, T. (2018). Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 531–535.

[82] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. ., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012a). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

[83] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

[84] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012b). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

[85] Hirsch, H. G. (2005). FaNT - Filtering and Noise Adding Tool. *Niederrhein University of Applied Sciences*.

[86] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.

[87] Holschneider, M., Kronland-Martinet, R., Morlet, J., and Tchamitchian, P. (1990). A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pages 286–297. Springer.

[88] Hornik, K., Stinchcombe, M., White, H., et al. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

[89] Hu, Y. and Loizou, P. C. (2003). A perceptually motivated approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 11(5):457–465.

[90] Hu, Y. and Loizou, P. C. (2004). Incorporating a psychoacoustical model in frequency domain speech enhancement. *IEEE Signal Processing Letters*, 11(2):270–273.

[91] Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (2014). Deep learning for monaural speech separation. In *Proc. of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1562–1566.

[92] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[93] Ito, N., Araki, S., and Nakatani, T. (2016). Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing. In *Proc. of 2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1153–1157.

[94] Ito, N., Vincent, E., Nakatani, T., Ono, N., Araki, S., and Sagayama, S. (2015). Blind suppression of nonstationary diffuse acoustic noise based on spatial covariance matrix decomposition. *Journal of Signal Processing Systems*, 79(2):145–157.

[95] Jensen, J. and Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022.

[96] Jeub, M., Herglotz, C., Nelke, C., Beaugeant, C., and Vary, P. (2012). Noise reduction for dual-microphone mobile phones exploiting power level differences. In *Proc. of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1693–1696.

[97] Jin, W., Taghizadeh, M. J., Chen, K., and Xiao, W. (2017). Multi-channel noise reduction for hands-free voice communication on mobile phones. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 506–510.

[98] Julier, S. and Uhlmann, J. (1997). A new extension of the Kalman filter to nonlinear systems. In *Signal processing, sensor fusion, and target recognition VI*, pages 182–193.

[99] Kallel, F., Ghorbel, M., Frikha, M., Berger-Vachon, C., and Hamida, A. B. (2012). A noise cross PSD estimator based on improved minimum statistics method for two-microphone speech enhancement dedicated to a bilateral cochlear implant. *Applied acoustics*, 73(3):256–264.

[100] Kamath, S., Loizou, P., et al. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proc. of 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 44164–44164.

[101] Kang, T. G., Shin, J. W., and Kim, N. S. (2018). DNN-based monaural speech enhancement with temporal and spectral variations equalization. *Digital Signal Processing: A Review Journal*, 74:102–110.

[102] Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., USA.

[103] Kida, Y., Tran, D., Omachi, M., Taniguchi, T., and Fujita, Y. (2018). Speaker selective beamformer with keyword mask estimation. In *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 528–534.

[104] Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 126(3):1486–1494.

[105] Kingma, D. P. and Ba, J. L. (2015). ADAM: A method for stochastic optimization. In *Proc. of 3rd International Conference for Learning Representations*, pages 1–13.

[106] Koizumi, Y., Niwa, K., Hioka, Y., Kobayashi, K., and Haneda, Y. (2017). DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 81–85.

[107] Kolbaek, M., Tan, Z.-H., and Jensen, J. (2018). Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[108] Kolbaek, M., Tan, Z.-H., and Jensen, J. (2019). 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*, 27(2):283–295.

[109] Kolbaek, M., Tan, Z.-H., Jensen, S., and Jensen, J. (2020). On loss functions for supervised monaural time-domain speech enhancement. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28:825–838.

[110] Kuklasinski, A., Doclo, S., Jensen, S., and Jensen, J. (2016). Maximum likelihood PSD estimation for speech enhancement in reverberation and noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1595–1608.

[111] Kumar, A. and Florencio, D. (2016). Speech enhancement in multiple-noise conditions using deep neural networks. In *Proc. of 17th Annual Conference of the International Speech Communication (InterSpeech)*, pages 3738–3742.

[112] Kumatani, K., McDonough, J., and Raj, B. (2012). Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Processing Magazine*, 29(6):127–140.

[113] Lamel, L., Kassel, R., and Seneff, S. (1989). Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 2161–2170.

[114] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[115] LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer.

[116] LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

[117] Lee, C.-H., Soong, F. K., and Paliwal, K. K. (2012). *Automatic speech and speaker recognition: Advanced topics*, volume 355. Springer Science & Business Media.

[118] Lefkimmiatis, S. and Maragos, P. (2007). A generalized estimation approach for linear and nonlinear microphone array post-filters. *Speech Communication*, 49(7-8):657–666.

[119] Leonard, R. (1984). A database for speaker-independent digit recognition. In *Proc of. ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 328–331.

[120] Li, B., Sainath, T. N., Weiss, R. J., Wilson, K. W., and Bacchiani, M. (2016). Neural network adaptive beamforming for robust multichannel speech recognition. In *Proc. of 17th Annual Conference of the International Speech Communication (InterSpeech)*, pages 1976–1980.

[121] Lin, L., Holmes, W., and Ambikairajah, E. (2003). Adaptive noise estimation algorithm for speech enhancement. *Electronics Letters*, 39(9):754–755.

[122] Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H. M., and Deng, L. (2015). Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine*, 32(3):35–52.

[123] Liu, Q., Wang, W., Jackson, P. J. B., and Tang, Y. (2017). A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions. In *Proc. of 2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1270–1274.

[124] Lockwood, P. and Boudy, J. (1992). Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. *Speech communication*, 11(2-3):215–228.

[125] Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice*. CRC Press, 2 edition.

[126] López-Espejo, I., Gomez, A. M., González, J. A., and Peinado, A. M. (2014). Feature enhancement for robust speech recognition on smartphones with dual-microphone. In *Proc. of 2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 21–25.

[127] López-Espejo, I., González, J. A., Gómez, Á. M., and Peinado, A. M. (2014). A deep neural network approach for missing-data mask estimation on dual-microphone smartphones: Application to noise-robust speech recognition. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 119–128.

[128] López-Espejo, I., Martín-Doñas, J. M., Gomez, A. M., and Peinado, A. M. (2018). Unscented transform-based dual-channel noise estimation: Application to speech enhancement on smartphones. In *Proc. of 2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pages 88–91.

[129] López-Espejo, I., Peinado, A. M., Gomez, A. M., and González, J. A. (2018). Dual-channel spectral weighting for robust speech recognition in mobile devices. *Digital Signal Processing*, 75:13–24.

[130] López-Espejo, I., Peinado, A. M., Gomez, A. M., and Martín-Doñas, J. M. (2016). Deep neural network-based noise estimation for robust ASR in dual-microphone smartphones. In *Proc. of IberSpeech 2016*, pages 117–127.

[131] Lotter, T. and Vary, P. (2005). Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. *EURASIP Journal on Advances in Signal Processing*, 2005(7):354850.

[132] Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (2013). Speech enhancement based on deep denoising autoencoder. In *Proc. of 14th Annual Conference of the International Speech Communication (InterSpeech)*, volume 2013, pages 436–440.

[133] Lu, Y. and Loizou, P. C. (2008). A geometric approach to spectral subtraction. *Speech communication*, 50(6):453–466.

[134] Luo, Y. and Mesgarani, N. (2018). TasNet: Time-domain audio separation network for real-time, single-channel speech separation. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700.

[135] Luo, Y. and Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266.

[136] Markovich, S., Gannot, S., and Cohen, I. (2009). Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1071–1086.

[137] Markovich-Golan, S. and Gannot, S. (2015). Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method. In *Proc. of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 544–548.

[138] Marro, C., Mahieux, Y., and Simmer, K. U. (1998). Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6(3):240–259.

[139] Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing*, 9(5):504–512.

[140] Martín-Doñas, J. M., Gomez, A. M., López-Espejo, I., and Peinado, A. M. (2017). Dual-channel DNN-based speech enhancement for smartphones. In *Proc. of 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6.

[141] Martín-Doñas, J. M., Gomez, A. M., Gonzalez, J. A., and Peinado, A. M. (2018a). A deep learning loss function based on the perceptual evaluation of the speech quality. *IEEE Signal Processing Letters*, 25(11):1680–1684.

[142] Martín-Doñas, J. M., Heitkaemper, J., Haeb-Umbach, R., Gomez, A. M., and Peinado, A. M. (2019a). Multi-channel block-online source extraction based on utterance adaptation. In *Proc. of 20th Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 96–100.

[143] Martín-Doñas, J. M., Jensen, J., Tan, Z.-H., Peinado, A. M., and Gomez, A. (2020a). Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Early Access.

[144] Martín-Doñas, J. M., López-Espejo, I., Gomez, A. M., and Peinado, A. M. (2018b). An extended kalman filter for RTF estimation in dual-microphone smartphones. In *Proc. of 2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2474–2478.

[145] Martín-Doñas, J. M., López-Espejo, I., Gomez, A. M., and Peinado, A. M. (2018c). A postfiltering approach for dual-microphone smartphones. In *Proc. of IberSpeech 2018*, pages 142–146.

[146] Martín-Doñas, J. M., Peinado, A. M., López-Espejo, I., and Gomez, A. (2019b). Dual-channel speech enhancement based on extended Kalman filter relative transfer function estimation. *Applied Sciences*, 9(12):2520.

[147] Martín-Doñas, J. M., Peinado, A. M., López-Espejo, I., and Gomez, A. (2020b). Dual-channel eKF-RTF framework with DNN-based speech presence estimation. In *Submitted at IberSpeech 2020*.

[148] Matsui, Y., Nakatani, T., Delcroix, M., Kinoshita, K., Ito, N., Araki, S., and Makino, S. (2018). Online integration of DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming. In *Proc. of 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 71–75.

[149] McCowan, I. A. and Bourlard, H. (2003). Microphone array post-filter based on noise field coherence. *IEEE Transactions on Speech and Audio Processing*, 11(6):709–716.

[150] McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York.

[151] Meng, Z., Watanabe, S., Hershey, J. R., and Erdogan, H. (2017). Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 271–275.

[152] Michelsanti, D. and Tan, Z.-H. (2017). Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In *Proc. of 18th Annual Conference of the International Speech Communication (InterSpeech)*, pages 2008–2012.

[153] Mozer, M. C. (1995). A focused backpropagation algorithm for temporal. *Backpropagation: Theory, architectures, and applications*, 137.

[154] Murphy, K. P. (1998). Switching Kalman filters. Technical report, U. C. Berkeley.

[155] Naithani, G., Nikunen, J., Bramslow, L., and Virtanen, T. (2018). Deep neural network based speech separation optimizing an objective estimator of intelligibility for low latency applications. In *Proc. of 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 386–390.

[156] Nakatani, T., Ito, N., Higuchi, T., Araki, S., and Kinoshita, K. (2017). Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290.

[157] Narayanan, A. and Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7092–7096.

[158] Nelke, C. M., Beaugeant, C., and Vary, P. (2013). Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability. In *Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7279–7283.

[159] Nugraha, A. A., Liutkus, A., and Vincent, E. (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664.

[160] Ochiai, T., Watanabe, S., Hori, T., Hershey, J. R., and Xiao, X. (2017). Unified architecture for multichannel end-to-end speech recognition with neural beamforming. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1274–1288.

[161] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[162] Parchami, M., Zhu, W. P., Champagne, B., and Plourde, E. (2016). Recent developments in speech enhancement in the short-time Fourier transform domain. *IEEE Circuits and Systems Magazine*, 16(3):45–77.

[163] Park, S. R. and Lee, J. (2016). A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*.

[164] Pascual, S., Bonafonte, A., and Serra, J. (2017). SEGAN: Speech enhancement generative adversarial network. In *Proc. of 18th Annual Conference of the International Speech Communication (InterSpeech)*, pages 3642–3646.

[165] Pearce, D. and Hirsch, H. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. of International Conference on Spoken Language Processing (ICSLP)*.

[166] Peinado, A. M. and Segura, J. C. (2006). *Speech Recognition Over Digital Channels: Robustness and Standards*. John Wiley and Sons.

[167] Pfeifenberger, L., Zöhrer, M., and Pernkopf, F. (2017). DNN-based speech mask estimation for eigenvector beamforming. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 66–70.

[168] Pfeifenberger, L., Zöhrer, M., and Pernkopf, F. (2019a). Deep complex-valued neural beamformers. In *Proc. of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2902–2906.

[169] Pfeifenberger, L., Zöhrer, M., and Pernkopf, F. (2019b). Eigenvector-based speech mask estimation for multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2162–2172.

[170] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

[171] Prasad, R., Saruwatari, H., and Shikano, K. (2004). Probability distribution of time-series of speech spectral components. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 87(3):584–597.

[172] Prechelt, L. (2012). Early Stopping - But When? In *Neural Networks: Tricks of the Trade*, pages 53–67. Springer Berlin Heidelberg.

[173] Qian, K., Zhang, Y., Chang, S., Yang, X., Florencio, D., and Hasegawa-Johnson, M. (2018). Deep learning based speech beamforming. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5389–5393.

[174] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286.

[175] Rahmani, M., Akbari, A., Ayad, B., and Lithgow, B. (2009). Noise cross PSD estimation using phase information in diffuse noise field. *Signal Processing*, 89(5):703–709.

[176] Richardson, F., Reynolds, D., and Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, 22(10):1671–1675.

[177] Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. (2019). SDR - Half-baked or Well Done? In *Proc. of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630.

[178] Rumelhart, D., Hinton, G., and Williams, R. (1986a). Learning internal representation by error propagation, Parallel Distributed Processing. *MIT Press, Cambridge*.

[179] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[180] Sainath, T. N., Weiss, R. J., Wilson, K. W., Li, B., Narayanan, A., Variani, E., Bacchiani, M., Shafran, I., Senior, A., Chin, K., et al. (2017). Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):965–979.

[181] Sainath, T. N., Weiss, R. J., Wilson, K. W., Narayanan, A., and Bacchiani, M. (2016). Factored spatial and spectral multichannel raw waveform CLDNNs. In *Proc. of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5075–5079.

[182] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

[183] Schwartz, B., Gannot, S., and Habets, E. (2017). Two model-based EM algorithms for blind source separation in noisy environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2209–2222.

[184] Schwartz, B., Gannot, S., and Habets, E. A. P. (2015). Online speech dereverberation using Kalman filter and EM algorithm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):394–406.

[185] Schwartz, O. and Gannot, S. (2018). A recursive expectation-maximization algorithm for online multi-microphone noise reduction. In *Proc. of 2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1542–1546.

[186] Schwartz, O., Gannot, S., and Habets, E. (2016). An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1491–1506.

[187] Schwarz, A. and Kellermann, W. (2015). Coherent-to-diffuse power ratio estimation for dereverberation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 23(6):1006–1018.

[188] Serizel, R., Moonen, M., Van Dijk, B., and Wouters, J. (2014). Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(4):785–799.

[189] Shin, J. W., Chang, J.-H., and Kim, N. S. (2005). Statistical modeling of speech signals based on generalized gamma distribution. *IEEE Signal Processing Letters*, 12(3):258–261.

[190] Shivakumar, P. G. and Georgiou, P. (2016). Perception optimized deep denoising autoencoders for speech enhancement. In *Proc. of 17th Annual Conference of the International Speech Communication (InterSpeech)*, pages 3743–3747.

[191] Sivasankaran, S., Vincent, E., and Fohr, D. (2018). Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment. In *Proc. of 19th Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 2703–2707.

[192] So, S. and Paliwal, K. (2011). Modulation-domain kalman filtering for single-channel speech enhancement. *Speech Communication*, 53(6):818–829.

[193] Souden, M., Benesty, J., and Affes, S. (2009a). On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on audio, speech, and language processing*, 18(2):260–276.

[194] Souden, M., Benesty, J., Affes, S., and Chen, J. (2011). An integrated solution for online multichannel noise tracking and reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2159–2169.

[195] Souden, M., Chen, J., Benesty, J., and Affes, S. (2009b). Gaussian model-based multichannel speech presence probability. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):1072–1077.

[196] Souden, M., Delcroix, M., Kinoshita, K., Yoshioka, T., and Nakatani, T. (2012). Noise power spectral density tracking: A maximum likelihood perspective. *IEEE Signal Processing Letters*, 19(8):495–498.

[197] Srinivasan, S., Samuelsson, J., and Kleijn, W. B. (2005). Codebook driven short-term predictor parameter estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):163–176.

[198] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

[199] Subramanian, V. (2018). *Deep Learning with PyTorch: A practical approach to building neural network models using PyTorch.* Packt Publishing Ltd.

[200] Sun, L., Du, J., Dai, L.-R., and Lee, C.-H. (2017). Multiple-target deep learning for LSTM-RNN based speech enhancement. In *Proc. of 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pages 136–140.

[201] Sun, Y., Xian, Y., Wang, W., and Naqvi, S. (2019). Monaural source separation in complex domain with long short-term memory neural network. *IEEE Journal on Selected Topics in Signal Processing*, 13(2):359–369.

[202] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

[203] Tamura, S. and Waibel, A. (1988). Noise reduction using connectionist models. In *Proc. of ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 553–554.

[204] Tan, K., Chen, J., and Wang, D. (2019a). Gated residual networks with dilated convolutions for monaural speech enhancement. *IEEE/ACM transactions on audio, speech, and language processing*, 27(1):189–198.

[205] Tan, K. and Wang, D. (2018). A convolutional recurrent neural network for real-time speech enhancement. In *Proc. of 19th Annual Conference of the International Speech Communication (InterSpeech)*, pages 3229–3233.

[206] Tan, K. and Wang, D. (2019). Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement. In *Proc. of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6865–6869.

[207] Tan, K. and Wang, D. (2020). Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28:380–390.

[208] Tan, K., Zhang, X., and Wang, D. (2019b). Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios. In *Proc. of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5751–5755.

[209] Taseska, M. and Habets, E. (2017). Nonstationary noise PSD matrix estimation for multichannel blind speech extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2223–2236.

[210] Taseska, M. and Habets, E. A. (2014). Informed spatial filtering for sound extraction using distributed microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(7):1195–1207.

[211] Taseska, M. and Habets, E. A. (2016). Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1291–1304.

[212] Tashev, I., Mihov, S., Gleghorn, T., and Acero, A. (2008). Sound capture system and spatial filter for small devices. In *Proc. of 9th Annual Conference of the International Speech Communication (InterSpeech)*, pages 435–438.

[213] Togami, M. and Kawaguchi, Y. (2014). Simultaneous optimization of acoustic echo reduction, speech dereverberation, and noise reduction against mutual interference. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(11):1612–1623.

[214] Truong, V. B., Nguyen, D. M., and Dang, Q. H. (2014). An MC-SPP approach for noise reduction in dual microphone case with power level difference. In *Proc. of 2014 International Conference on Advanced Technologies for Communications (ATC)*, pages 292–297.

[215] Udrea, R. M. and Ciochina, S. (2003). Speech enhancement using spectral oversubtraction and residual noise reduction. In *Proc. of International Symposium on Signals, Circuits and Systems*, volume 1, pages 165–168.

[216] Van Veen, B. D. and Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE Acoustic and Speech Signal Processing Magazine*, 5(2):4–24.

[217] Varela, Ó., San-Segundo, R., and Hernández, L. A. (2011). Combining pulse-based features for rejecting far-field speech in a HMM-based voice activity detector. *Computers & Electrical Engineering*, 37(4):589–600.

[218] Varzandeh, R., Taseska, M., and Habets, E. A. P. (2017). An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation. In *Proc. of 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pages 11–15.

[219] Venkataramani, S., Higa, R., and Smaragdis, P. (2018). Performance based cost functions for end-to-end speech separation. In *Proc. of 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 350–355.

[220] Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469.

[221] Vincent, E., Watanabe, S., Nugraha, A., Barker, J., and Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech and Language*, 46:535–557.

[222] Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on speech and audio processing*, 7(2):126–137.

[223] Vlaj, D., KačIč, Z., and Kos, M. (2012). Voice activity detection algorithm using nonlinear spectral weights, hangover and hangbefore criteria. *Computers & Electrical Engineering*, 38(6):1820–1836.

[224] Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(10):1702–1726.

[225] Wang, D. and Lim, J. (1982). The unimportance of phase in speech enhancement. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30(4):679–681.

[226] Wang, J., Chen, J., Su, D., Chen, L., Yu, M., Qian, Y., and Yu, D. (2018a). Deep extractor network for target speaker recovery from single channel speech mixtures. In *Proc. of 19th Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 307–311.

[227] Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J. R., Saurous, R. A., Weiss, R. J., Jia, Y., and Moreno, I. L. (2019a). VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. In *Proc. of 20th Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 2728–2732.

[228] Wang, Y., Fan, X., Chen, I.-F., Liu, Y., Chen, T., and Hoffmeister, B. (2019b). End-to-end anchored speech recognition. In *Proc. of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7090–7094.

[229] Wang, Y., Narayanan, A., and Wang, D. L. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 22(12):1849–1858.

[230] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

[231] Wang, Y. and Wang, D. L. (2013). Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech and Language Processing*, 21(7):1381–1390.

[232] Wang, Z., Vincent, E., Serizel, R., and Yan, Y. (2018b). Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments. *Computer Speech and Language*, 49:37–51.

[233] Wang, Z.-Q., Le Roux, J., and Hershey, J. R. (2018c). Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

[234] Warsitz, E. and Haeb-Umbach, R. (2007). Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Transactions on audio, speech, and language processing*, 15(5):1529–1539.

[235] Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., and Schuller, B. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *Proc. of International Conference on Latent Variable Analysis and Signal Separation*.

[236] Williamson, D., Wang, Y., and Wang, D. (2016). Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24(3):483–492.

[237] Wolfe, P. and Godsill, S. (2003). Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. *Eurasip Journal on Advances in Signal Processing*, (10):1043–1051.

[238] Xia, B. and Bao, C. (2013). Speech enhancement with weighted denoising auto-encoder. In *Proc. of 14th Annual Conference of the International Speech Communication (InterSpeech)*, pages 3444–3448.

[239] Xia, B. and Bao, C. (2014). Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Communication*, 60:13–29.

[240] Xiao, X., Watanabe, S., Erdogan, H., Lu, L., Hershey, J., Seltzer, M. L., Chen, G., Zhang, Y., Mandel, M., and Yu, D. (2016). Deep beamforming networks for multi-channel speech recognition. In *Proc. of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5745–5749.

[241] Xiao, X., Zhao, S., Jones, D. L., Chng, E. S., and Li, H. (2017). On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3246–3250.

[242] Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*, 21(1):65–68.

[243] Xu, Y., Du, J., Dai, L. R., and Lee, C. H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Speech and Language Processing*, 23(1):7–19.

[244] Xue, W., Moore, A., Brookes, M., and Naylor, P. (2018). Modulation-domain multichannel Kalman filtering for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1833–1847.

[245] Yamagishi, J. (2012). English multi-speaker corpus for CSTR voice cloning toolkit.

[246] Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245.

[247] Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., and Hinton, G. E. (2013). On rectified linear units for speech processing. In *Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3517–3521.

[248] Zelinski, R. (1988). A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proc. of ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 2578–2581.

[249] Zhang, H., Zhang, X., and Gao, G. (2018). Training supervised speech separation system to improve STOI and PESQ directly. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5374–5378. IEEE.

[250] Zhang, X., Wang, Z.-Q., and Wang, D. (2017). A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR. In *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 276–280.

[251] Zhao, Y., Xu, B., Giri, R., and Zhang, T. (2018). Perceptually guided speech enhancement using deep neural networks. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5074–5078.

[252] Zhao, Z., Elshamy, S., and Fingscheidt, T. (2019). A perceptual weighting filter loss for DNN training in speech enhancement. In *Proc. of 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 229–233.

[253] Zheng, C., Liu, H., Peng, R., and Li, X. (2013). A statistical analysis of two-channel post-filter estimators in isotropic noise fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):336–342.

[254] Zmolikova, K., Delcroix, M., Kinoshita, K., Higuchi, T., Nakatani, T., and Černockỳ, J. (2018). Optimization of speaker-aware multichannel speech extraction with ASR criterion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6702–6706.

[255] Zmolikova, K., Delcroix, M., Kinoshita, K., Higuchi, T., Ogawa, A., and Nakatani, T. (2017). Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In *Proc. of 18th Annual Conference of the International Speech Communication (InterSpeech)*, pages 2655–2659.

[256] Zohrer, M., Pfeifenberger, L., Schindler, G., Froning, H., and Pemkopf, F. (2018). Resource efficient deep eigenvector beamforming. In *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3354–3358.

[257] Zwicker, E. and Feldtkeller, R. (1967). *Das Ohr als Nachrichtenempfanger*. S. Hirtzel Verlag Stuttgart.