IberSPEECH 2022
14-16 November 2022, Granada, Spain

# Online Multichannel Speech Enhancement combining Statistical Signal Processing and Deep Neural Networks: A Ph.D. Thesis Overview

*Juan M. Martín-Doñas*[1,2]*, Antonio M. Peinado*[2] *and Angel M. Gomez*[2]

[1]Fundación Vicomtech, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia-San Sebastián, Spain
[2]Dept. de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, Spain
jmmartin@vicomtech.org

## Abstract

Speech-related applications on mobile devices require high-performance speech enhancement algorithms to tackle challenging, noisy real-world environments. In addition, current mobile devices often embed several microphones, allowing them to exploit spatial information. 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, unfeasible or very limiting. Our contributions focus on different noisy environments where these mobile speech technologies can be applied. These include dual-microphone smartphones in noisy and reverberant environments and general multi-microphone devices for speech enhancement and target source separation. Moreover, we study the training of deep learning methods for speech processing using perceptual considerations. Our contributions successfully integrate signal processing and deep learning methods to exploit spectral, spatial, and temporal speech features jointly. As a result, the proposed techniques provide us with a manifold framework for robust speech processing under very challenging acoustic environments, thus allowing us to improve perceptual quality and intelligibility measures.

**Index Terms**: speech enhancement, array signal processing, deep neural networks, low-latency, speech presence probability

## 1. Introduction

Speech enhancement algorithms aim to improve the perceptual quality and intelligibility of speech signals degraded due to distortions [1], especially environmental noise, but also reverberation or interfering speakers. These techniques are usually implemented in the time-frequency (TF) domain via the short-time Fourier transform (STFT) [2]. Thus, classical single-channel methods, including spectral subtraction, Wiener filtering, or Bayesian estimators [3], can be expressed as a gain function applied to the noisy STFT. The availability of microphone arrays in recent speech devices, such as smartphones or smart speakers, allows us to exploit the additional spatial information through multichannel speech enhancement techniques. The most common strategy consists of using a frequency domain beamforming algorithm [4], which applies spatial filtering, followed by a single-channel postfilter to enhance the speech signal further. The performance of the aforementioned techniques relies on an accurate estimation of the noise power spectral density (PSD) [5], in the single-channel case, and the noise spatial covariances and relative acoustic channels for beamforming [6]. Despite the existence of different estimation algorithms for each case, most popular approaches are based on the computation of the speech presence probability (SPP) for each TF bin in order to discriminate speech- and noise-dominant spectral regions [7].

In the last decade, the revolution of the deep learning paradigm has extended the use of deep neural networks (DNN) for speech enhancement tasks [8]. Regarding the STFT domain, two main approaches have been followed for the single-channel scenario, spectral mapping [9] or masking [10], which differ in the network's target. While the first directly tries to estimate the magnitude spectrum of the clean speech signal, the latter uses the DNN to compute a gain function. In this last case, network outputs can be optimized for a pre-defined target mask or through a loss function that considers the enhanced speech signal [8]. The integration of DNNs with beamforming algorithms can be performed in several ways. Still, a general approach uses DNN spectral mask estimators to compute the needed speech and noise spatial covariances accurately [11, 12, 13]. This is similar to the aforementioned SPP paradigm.

Among the presented speech enhancement approaches, classical signal processing is limited due to the assumptions made, while DNNs depend on the training data and can lack generalization. In addition, common methods assume the availability of the whole speech signal during processing. On the other hand, speech-related applications in mobile devices have to ensure online processing with low latency and computational efficiency. Therefore, this Thesis focuses on developing online multichannel speech enhancement techniques suitable for mobile devices. Our objective is the integration of statistical signal processing algorithms with DNNs efficiently used in parts of the algorithm where assumptions about signal properties are weak. Thus, this yields increases in robustness under challenging, noisy real-world environments while allowing for online processing. More specifically, we focus on four different scenarios to apply these integrated techniques: (1) dual-microphone smartphones in noisy and reverberant environments exploiting power and phase channel differences, (2) the joint estimation of clean speech and acoustic parameters in general multi-microphone devices, (3) multichannel target speaker extraction in multi-talker mixtures by exploiting auxiliary spectral and spatial information, and (4) the training of DNNs for speech enhancement using perceptual considerations of the human auditory system.

The rest of this paper aims to describe the key aspects of the developed speech enhancement algorithms in this Thesis, the most relevant experimental results obtained and the main conclusions drawn for the research addressed in these works.


10.21437/IberSPEECH.2022-45

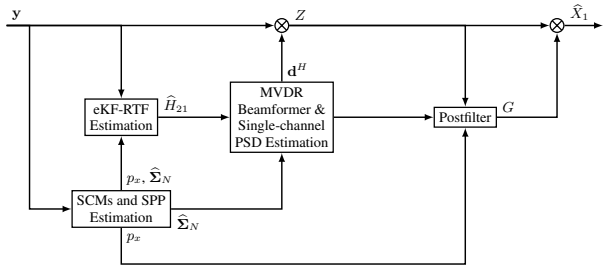



Figure 1: *Overview of the dual-channel speech enhancement algorithm for dual-microphone smartphones.*

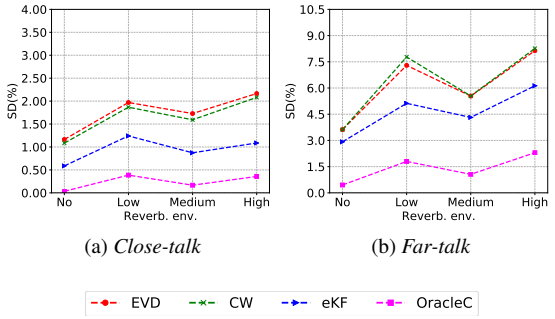
## 2. Thesis overview

In this section, we present four speech enhancement (SE) algorithms developed in this Thesis. First, we give an overview description of these techniques and their contributions to the state-of-the-art. Then, we present the most relevant experimental results and analysis obtained through evaluating these algorithms in different speech corpora. Our evaluations focus on the use of objective quality and intelligibility measures, especially the Perceptual Evaluation of the Speech Quality (PESQ) [14], the Short-Time Objective Intelligibility (STOI) [15] and its extended version (ESTOI) [16], as well as the Scale Invariant Signal Distortion Ratio (SI-SDR) [17].

### 2.1. Dual-channel SE based on extended Kalman filter for channel estimation

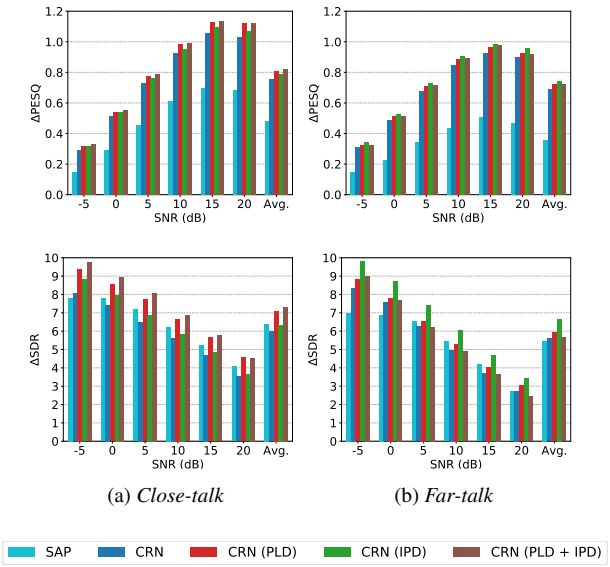
The proposed algorithm for dual-microphone smartphones is depicted in Figure 1. The dual-channel noisy speech signal vector in the TF domain, $\mathbf{y}(t, f)$, is processed by a minimum variance distortionless response (MVDR) beamformer $\mathbf{d}(t, f)$ [7] followed by a Bayesian postfilter $G(t, f)$ [5]. This computation requires knowledge about the relative transfer function (RTF) between the microphone acoustic channels $H_{21}(t, f)$, and the noise spatial covariance matrices (SCM) $\mathbf{\Sigma}_N(t, f)$. The estimation of the RTF variations in reverberant environments is addressed via a proposed extended Kalman filter (eKF) for tracking the channel variations [18, 19] assuming the availability of a priori RTF statistics information. On the other hand, the estimation of the SCMs relies on the computation of the SPP $p_x(t, f)$ through the power level difference (PLD) [20] and interchannel phase difference (IPD) [21] information. To this end, we investigate two different approaches. The first one, based on Bayesian estimation, considers both the likelihoods of the noisy speech given speech presence and absence hypotheses and a priori speech absence probability (SAP), which is computed through the previous dual-channel features [22]. The second one directly performs the SPP estimation using a convolutional recurrent network (CRN) fed with both the log-magnitude noisy spectrum and the dual-channel features [23].

The proposed technique was evaluated in a simulated dual-channel speech database that considers a smartphone used in different noisy and reverberant environments [18]. Two different user positions were considered, *close-talk* and *far-talk*, which consider a different distance between the speaker and the smartphone (conversational and hands-free). We first compare the proposed eKF estimator with common approaches for RTF estimation, such as eigenvalue decomposition (EVD) and covariance whitening (CW) [7]. The distortionless property of the MVDR beamformer holds when an accurate estimation of the




(a) *Close-talk* (b) *Far-talk*

Figure 2: *SD results for the different RTF estimators when used for MVDR beamforming.*




(a) *Close-talk* (b) *Far-talk*

Figure 3: *PESQ and SDR results from the evaluation of the CRN-based SPP mask estimator with the different input features. The plots show the increments obtained on the metrics with respect to noisy speech.*

RTF is achieved. Therefore, in Figure 2 the RTF estimators are compared in terms of the speech distortion (SD) [4] in the clean speech at the MVDR output. The results show lower distortion for the eKF estimator, closer to an oracle estimation, indicating a more precise RTF estimation. Finally, in Figure 3 we compare the performance of the two SPP estimators in terms of final noise reduction. It is observed that the CRN performs better than its statistical counterpart, while the dual-channel features provide useful information. Thus, PLD features help in close-talk condition, while IPD features are more relevant in the far-talk scenario.

### 2.2. Multichannel SE based on recursive expectation-maximization with DNN speech presence priors

We will now describe the proposed multichannel speech enhancement approach for general multi-microphone devices. To increase the robustness of the approach, we addressed the joint estimation of the speech statistics and the acoustic parameters using a Bayesian framework. A possible solution is to perform maximum likelihood (ML) estimation, but it has no closed-form solution [24]. In addition, the variables should be computed

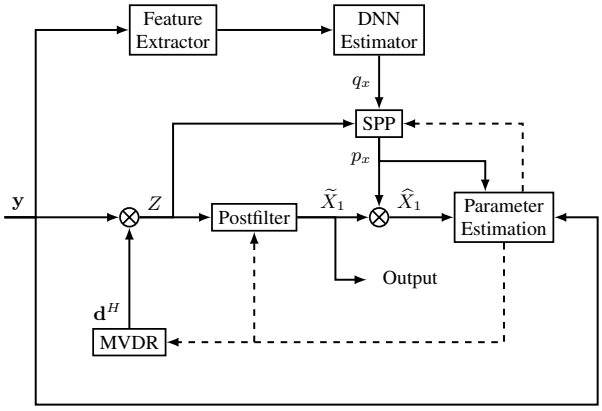


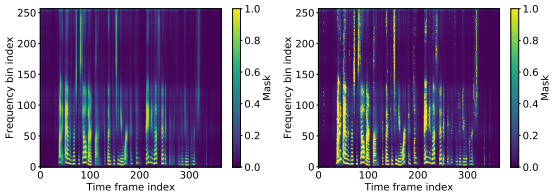

Figure 4: *Block diagram of the proposed REM algorithm for multichannel speech enhancement, depicting the most relevant parts. The dashed lines indicate the feedback due to the M-step.*

Table 1: *PESQ, ESTOI and SI-SDR results for the different variants of the proposed REM algorithm*

| Method | PESQ | ESTOI (%) | SDR (dB) |
|---|---|---|---|
| Noisy | 1.27 | 68.2 | 7.51 |
| MVDR | 1.59 | 79.5 | 11.34 |
| MWF | 1.94 | 80.1 | 13.01 |
| MKF | 1.81 | 77.1 | 12.53 |
| REMWF | 2.05 | 83.3 | 14.46 |
| REMKF | **2.08** | **84.1** | **14.89** |

in an online fashion. Therefore, we proposed using a recursive expectation-maximization (REM) iterative procedure [25], which can be performed with only a few iterations per frame. The algorithm is depicted in Figure 4 and consists in a two-step procedure per iteration. In the E-step, the filtered speech signal $\widetilde{X}_1(t,f)$ is obtained through MVDR beamforming plus postfiltering, where two different postfilters are evaluated: Wiener and Kalman filters. The a posteriori SPP $p_x(t,f)$ is also obtained through Bayes' rule using the noisy speech likelihoods and the a priori SPP $q_x(t,f)$. In the M-step, the acoustic parameters for the beamforming and the postfilter are computed using the clean speech statistics and SPP by ML estimation. Furthermore, a recurrent neural network is used to compute the a priori SPP given the noisy speech signal, thus increasing the robustness and improving the convergence of the REM algorithm.

The REM framework was evaluated using the multichannel CHiME-3 database [26], which comprises noisy speech signals captured using a 6-microphone tablet in different public spaces. Table 1 shows the results obtained with the two variants of the REM approach with Wiener (REMWF) and Kalman (REMKF) postfilters. Our approach is compared with MVDR as well as multichannel Wiener filter (MWF) [27] and multichannel Kalman filters (MKF) [28] with a DNN-based mask estimation for the acoustic parameters [11]. These results show that the proposed REM framework outperforms other approaches, with the Kalman filter achieving the best results, which remarks the importance of exploiting additional temporal information. In addition, Figure 5 shows an example of the a priori and the a posteriori SPP. We can observe how the DNN prediction can be improved during the REM procedure by considering the statistical spatial information.




(a) *A priori SPP* (b) *A posteriori SPP*

Figure 5: *Example of a priori and a posteriori SPP obtained from the DNN and the REMKF algorithm.*

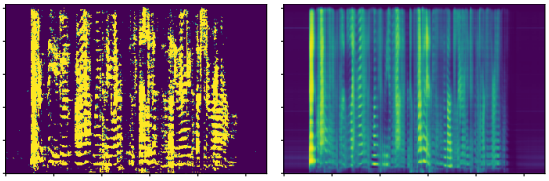
Table 2: *SDR, STOI and WER scores obtained for different speaker extractors*

| 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** |

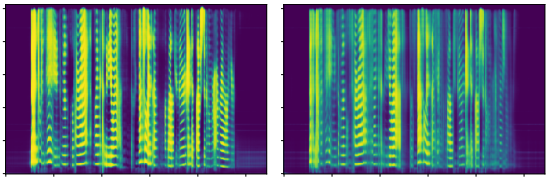
### 2.3. Multichannel target speaker separation based on spatial-beam network

We will now focus on a different scenario with a target speaker and other overlapped interfering speakers. Our objective is the estimation of the target speaker using beamforming with DNN-based mask estimators for the computation of the covariance matrices. The main problem is that the network cannot discriminate among different speakers. To solve this issue, the Speaker-beam approach was previously proposed [29]. It consists of using an adaptation utterance from the target speaker to compute a speaker vector using an auxiliary network. This speaker vector is used in the DNN mask estimator to adapt the network and compute the target speaker mask. The main limitation of this approach is the degradation suffered by unseen speakers with similar voice patterns (e.g., same gender). To overcome this issue and also adapt the system for online processing, we proposed the Spatial-beam variant [30], which considers block-online beamforming and mask estimation, as well as the use of spatial information from the multichannel adaptation utterance. Two different alternatives are considered: (1) pre-beamforming (PreBF), which computes an offline MVDR beamformer from the adaptation utterance to enhance the input signals to the Speaker-beam mask estimator, and (2) the use of additional IPD spatial features from both the noisy speech and adaptation utterances for the Speaker-beam network.

The systems were evaluated in a simulated multi-speaker version of the Wall Street Journal database [31]. Table 2 compares Spatial-beam variants with Speaker-beam, and also Deep Attractor Networks (DAN) [32] for source separation. In addition, we evaluated the online versions of the proposed approach. Apart from enhancement metrics such as STOI and SDR, we also evaluated the word error rate (WER) when enhanced signals are used in a DNN-based speech recognition system. It can be observed that the proposed Spatial-beam system outperforms Speaker-beam and achieves similar results with DAN networks while focusing on the target speaker. Moreover, the



(a) *IBM* (b) *Speaker-beam*



(c) *PreBF* (d) *Online-PreBF*

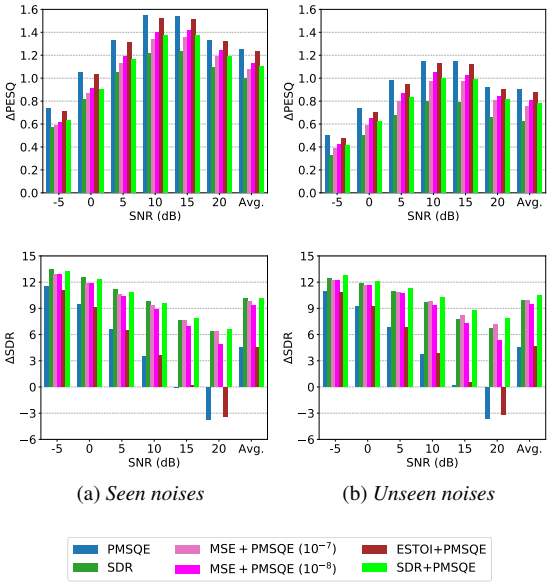
Figure 6: *Examples of different estimated target speaker masks obtained from different DNN speaker extractors.*

performance is still competitive for the more challenging online scenario. Although the two variants perform similarly, PreBF yields better signal distortion while using spatial features improves the recognition accuracy. An example of the target speaker mask obtained for Speaker-beam and Spatial-beam is shown in Figure 6. It can be observed that the speaker mask computed using Spatial-beam is closer to the target ideal binary mask (IBM) used for the DNN training, both for the offline and online mask estimators.

### 2.4. Deep learning loss function for the perceptual evaluation of the speech quality

The last contribution of this Thesis is devoted to the training of DNN-based speech enhancement algorithms exploiting perceptual considerations. To this end, a direct approach consists of integrating well-established objective quality metrics as criteria to the loss function [33]. We proposed a differentiable adaptation of the PESQ algorithm as a loss function, called Perceptual Metric for the Speech Quality Evaluation (PMSQE) [34]. This loss is intended to improve the speech quality, and its computation is performed as follows. First, a standardized listening level enhanced and clean speech spectra are transformed to a perceptual domain by applying a Bark transformation followed by a sone loudness scale. In addition, the Bark spectrum of the enhanced signal is equalized to remove non-relevant effects, such as time-invariant non-severe filtering and short-term gain variations. Finally, two per-frame disturbance terms are computed from the loudness spectra differences: the symmetrical and asymmetrical disturbances. These terms account for masking effects and discriminate spectral differences due to distorted speech and additive noise. The PMSQE loss is essentially obtained as the weighted sum of both terms averaged over time.

The proposed loss function was evaluated for training spectral masking speech enhancement using a CRN network. Thus, a simulated noisy version of the TIMIT database [35] was used for training and testing. Figure 7 shows the results for different objective metrics when the network is trained with PMSQE and other losses based on quality metrics. The results were obtained for both seen and unseen noises during training. The experimental results show that PMSQE loss outperforms in terms of PESQ metric, while the combination of PMSQE with SDR loss gives a generally good performance among objective metrics.




(a) *Seen noises* (b) *Unseen noises*

Figure 7: *PESQ, and SDR results for the PMSQE loss and other metric-based loss functions. The plots show the increments obtained on the metrics with respect to noisy speech.*

## 3. Conclusions

In this Thesis, we have developed a set of online multichannel speech enhancement algorithms suitable for low-latency processing in mobile devices. These techniques are able to combine statistical signal processing and deep neural networks successfully. Therefore, we can overcome the limitations imposed by the statistical framework, thus improving the final performance. This allows us to efficiently exploit the spectral, temporal, and spatial information within the noisy speech signals to obtain high-quality speech estimates.

The availability of accurate knowledge about the SPP in the TF domain is a crucial element in the performance of the proposed algorithms. Among the different approaches, DNN mask estimators have shown astounding performance in the computation of accurate SPP estimates. Their potential can also be extended to estimating target spectral masks for overlapped speakers and defining gain functions for single-channel spectral masking. This allows the design of low-latency and lighter computational algorithms suitable for mobile devices in real-world conditions.

As a results of this Thesis[1], we have published a total of three JCR journals [34, 22, 25] and four conference papers [18, 19, 30, 23] in InterSpeech, Eusipco and IberSpeech.

For future work, interesting research lines that can be explored are the full integration of DNNs and statistical signal processing, dealing with more complex reverberant environments, and using complex DNNs for spectral masking.

## 4. Acknowledgements

This work has been supported by the project PID2019-104206GB-I00 funded by MCIN/AEI/10.13039/501100011033.

[1] The Thesis is available at https://digibug.ugr.es/handle/10481/66402

# 5. References


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

[2] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[3] M. Parchami, W. P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time Fourier transform domain," *IEEE Circuits and Systems Magazine*, vol. 16, no. 3, pp. 45–77, 2016.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, 2008, vol. 1.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[6] M. Taseska and E. Habets, "Nonstationary noise PSD matrix estimation for multichannel blind speech extraction," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2223–2236, 2017.

[7] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[8] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. on Audio Speech and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[9] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. on Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[10] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on Audio Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[11] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech and Language*, vol. 46, pp. 374–385, 2017.

[12] S. Chakrabarty and E. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[13] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[14] "ITU-T. Rec P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codec," International Telecommunication Union-Telecommunication Standarisation Sector, Tech. Rep., 2007.

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

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

[17] J. Roux, S. Wisdom, H. Erdogan, and J. Hershey, "SDR - Half-baked or Well Done?" in *Proc. of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[18] 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 *Proc. of 2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2474–2478.

[19] ——, "A postfiltering approach for dual-microphone smartphones," in *Proc. of IberSpeech 2018*, 2018, pp. 142–146.

[20] I. López-Espejo, A. M. Peinado, A. M. Gomez, and J. A. González, "Dual-channel spectral weighting for robust speech recognition in mobile devices," *Digital Signal Processing*, vol. 75, pp. 13–24, 2018.

[21] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "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)*, 2018, pp. 1–5.

[22] J. M. Martín-Doñas, A. M. Peinado, I. López-Espejo, and A. Gomez, "Dual-channel speech enhancement based on extended Kalman filter relative transfer function estimation," *Applied Sciences*, vol. 9, no. 12, p. 2520, 2019.

[23] ——, "Dual-channel eKF-RTF framework for speech enhancement with DNN-based speech presence estimation," in *Proc. IberSpeech 2020*, 2021, pp. 31–35.

[24] B. Schwartz, S. Gannot, and E. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2209–2222, 2017.

[25] J. M. Martín-Doñas, J. Jensen, Z.-H. Tan, A. M. Peinado, and A. Gomez, "Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 3080–3094, 2020.

[26] E. Vincent, S. Watanabe, A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535–557, 2017.

[27] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech and Language*, vol. 49, pp. 37–51, 2018.

[28] W. Xue, A. Moore, M. Brookes, and P. Naylor, "Modulation-domain multichannel Kalman filtering for speech enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1833–1847, 2018.

[29] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures." in *Proc. of InterSpeech 2017*, 2017, pp. 2655–2659.

[30] 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 *Proc. of InterSpeech 2019*, 2019, pp. 96–100.

[31] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[32] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 246–250.

[33] M. Kolbaek, Z.-H. Tan, S. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Trans. on Audio Speech and Language Processing*, vol. 28, pp. 825–838, 2020.

[34] 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*, vol. 25, no. 11, pp. 1680–1684, 2018.

[35] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proceedings of the DARPA Speech Recognition Workshop*, 1989, pp. 2161–2170.