PERFORMANCE COMPARISON OF REAL-TIME SINGLE-CHANNEL SPEECH DEREVERBERATION ALGORITHMS

Feifei Xiong[1,4], *Bernd T. Meyer*[2], *Benjamin Cauchi*[1,4], *Ante Jukić*[3,4], *Simon Doclo*[1,3,4], *Stefan Goetze*[1,4]

[1]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[2]Johns Hopkins University, Center for Language and Speech Processing, Baltimore, USA
[3]University of Oldenburg, Department of Medical Physics and Acoustics, Oldenburg, Germany
[4]Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

This paper investigates four single-channel speech dereverberation algorithms, i.e., two unsupervised approaches based on (i) spectral enhancement and (ii) linear prediction, as well as two supervised approaches relying on machine learning which incorporate deep neural networks to predict either (iii) the magnitude spectrogram or (iv) the ideal ratio mask. The relative merits of the four algorithms in terms of several objective measures, automatic speech recognition performance, robustness against noise, variations between simulated and recorded reverberant speech, computation time and latency are discussed. Experimental results show that all four algorithms are capable of providing benefits in reverberant environments even with moderate background noises. In addition, low complexity and latency indicate their potential for real-time applications.

Index Terms— Speech dereverberation, supervised learning, objective measure, speech recognition, real-time application

1. INTRODUCTION

Reverberation, caused by multiple acoustic reflections inside an enclosed space, has been shown to be detrimental to speech perception and automatic speech recognition (ASR) systems [1, 2]. Speech dereverberation is therefore desirable for speech communication applications aiming at improving perceptual speech quality, speech intelligibility and ASR performance [2, 3, 4]. Single-channel speech dereverberation approaches include acoustic channel equalization [5, 6], spectral enhancement [7, 8, 9], probabilistic model-based dereverberation [10, 11], and time-frequency (TF) masking [12, 13] adopted from the field of speech separation in computational auditory scene analysis (CASA) [14]. Since channel equalization is a non-blind speech dereverberation method, requiring an accurate estimate of the room impulse response (RIR) between the speech source and the microphone [6, 13], this severely limits its applicability in practice. Spectral enhancement methods aim at estimating a real-valued gain in the short-time Fourier transform (STFT) domain to obtain the clean speech spectral coefficients. These methods are often based on a statistical RIR model, e.g., Polack's RIR model requiring an estimate of the reverberation time (RT) [7] or a more generalized model additionally requiring the direct-to-reverberant ratio (DRR) [8]. The low complexity of spectral enhancement methods is attractive for real-time applications even in spite of their inherent trade-off between late reverberation suppression and speech distortion. Speech dereverberation can also be performed by estimating the late/undesired reverberation using linear prediction, where the filters are typically computed by maximizing sparsity of the output signal in the STFT domain [10, 11]. Real-time capable adaptive implementations based on recursive least squares (RLS) have been proposed in [15, 16]. Motivated by TF masking approaches from the field of CASA, i.e., aiming to achieve speech separation in noisy environments [14], also TF masking approaches for speech dereverberation have been proposed [12, 13]. However, in practice an accurate estimate of the TF mask remains challenging [13, 17].

Recently, the great success of supervised learning using deep neural networks (DNNs) for ASR [18] has been extended to speech enhancement/separation, e.g., in [17, 19, 20], by considering the relationship between the distorted signal (noisy or reverberant speech) and the target signal (clean or anechoic speech) as a nonlinear transformation. From this perspective, speech dereverberation can be formulated as a *supervised* learning problem on the basis of the underlying dereverberation principles derived from the aforementioned *unsupervised* methods. According to the training targets of DNNs, we consider two different supervised algorithms for real-time speech dereverberation, i.e., one based on mapping and one based on masking. More specifically, inspired by the spectral enhancement dereverberation methods, mapping-based methods (e.g., [17]) train a nonlinear mapping between the spectra of the reverberant speech and the spectra of the corresponding anechoic speech. Masking-based methods train DNN to estimate the mask that best approximates a reference mask computed using the clean and reverberant spectra. In contrast to the mapping-based methods which directly consider spectra as DNN features, masking-based methods require dedicated features as input to the DNN in order to improve the DNN's discriminative power w.r.t. the DNN targets, e.g., the ideal binary mask (IBM) [14] or the ideal ratio mask (IRM) [19]. Motivated by findings in masking-based speech separation [21] that the best individual features were both auditory-inspired in matched and mismatched test conditions, herein we propose to use auditory-inspired features that are capable of extracting temporal modulation cues [22] which are known to play an important role in the human auditory system to analyze dynamic acoustic stimuli.

This paper aims to systematically investigate the performance and the applicability of the four mentioned single-channel speech dereverberation algorithms for real-time applications. Section 2 briefly introduces the algorithms and Section 3 outlines the experimental procedure. Performance results and detailed discussions of the relative merits will be addressed in Section 4.

2. SPEECH DEREVERBERATION ALGORITHMS

2.1. Spectral Enhancement

Spectral enhancement based dereverberation methods usually assume that late reverberation is uncorrelated to the direct or early

speech component and hence, can be considered as additive disturbance. In the STFT domain the reverberant speech $x[k,\ell]$ for frequency index k and block time index ℓ can be represented as the addition of the (clean) early speech component speech $x_e[k,\ell]$ and the late reverberation speech component $x_l[k,\ell]$, i.e.,

$$x[k,\ell] = x_e[k,\ell] + x_l[k,\ell]\,. \tag{1}$$

When the spectral variance of the late reverberation component can be accurately estimated, speech dereverberation can be achieved, e.g., by estimating the amplitude of the clean speech component in a minimum mean square error (MMSE) sense [23]. Typically, a statistical RIR model is used to estimate the spectral variance of the late reverberation component [7]. In order to cover scenarios in which the speaker-microphone distance is smaller than the critical distance, i.e., the DRR is larger than 0 dB [24], a generalized model [8] is adopted here. To jointly estimate RT and DRR, we have used a trained neural network with a low complexity [25].

A parameterized MMSE spectral magnitude estimator [26] is used to determine the gain function $g[k,\ell]$, where the required a-priori early-to-late-reverberation energy ratio is estimated using the decision-directed approach [23]. Subsequently, the desired speech component $\hat{x}_e[k,\ell]$ is estimated as

$$\hat{x}_e[k,\ell] = \max(g[k,\ell],\, g_{\min})\, x[k,\ell]\,, \tag{2}$$

where $g_{\min}$ is a lower bound for the gain function $g[k,\ell]$, which alleviates speech distortions. An inverse STFT is then used to reconstruct the speech signal in the time domain.

2.2. Dereverberation by Single-Channel Linear Prediction

Dereverberation using linear prediction is based on the assumption that the late reverberant component at the current time can be predicted from the previous microphone signals, which holds exactly for the multi-channel case [10] and is a good approximation for the single-channel case [27]. In the single-channel case, the late reverberation can then be modeled as

$$x_l[k,\ell] = \sum_{\iota=0}^{L_p-1} p_\iota[k,\ell]\, x[k,\ell-\tau-\iota]\,, \tag{3}$$

where $p_\iota[k,\ell]$ denotes the ι-th prediction coefficient at time ℓ, L_p is the number of the coefficients, and τ is the prediction delay to determine the boundary between early reflections and late reverberation. An estimate of the desired signal $\hat{x}_e[k,\ell]$ can be obtained once the prediction coefficients $p_\iota[k,\ell]$ in (3) are accurately estimated as

$$\hat{x}_e[k,\ell] = x[k,\ell] - \hat{x}_l[k,\ell]\,. \tag{4}$$

The unknown prediction coefficients are typically estimated by maximizing sparsity of the output (desired) speech signal in the STFT domain [10, 11]. An online method suitable for real-time implementations, based on the RLS algorithm, has been proposed in [16].

2.3. Mapping-based Supervised Dereverberation

In contrast to the unsupervised dereverberation methods described above, supervised methods rely on the learning abilities of DNNs to generate a transformation from reverberant speech to anechoic speech. A straightforward manner is to train a mapping from the reverberant spectra to the anechoic spectra [17]. The log-spectrogram of the reverberant speech signal is chosen as the DNN input, and its anechoic version is used as the DNN target. Note that the target log-spectrogram is normalized to the range of $[0,1]$ to provide a bounded DNN target. The time domain signal is then synthesized using the phase of the reverberant speech. The DNN operates on a frame-by-frame basis allowing for real-time processing.

2.4. Masking-based Supervised Dereverberation

The masking-based supervised approaches take the TF mask as the DNN target, for which IRM usually provides better performance than other mask types (e.g., IBM) according to the findings from speech separation [19]. The considered IRM which is applied to the reverberant spectrum for dereverberation is defined as (cf. [28]),

$$m[k,\ell] = \frac{|x_e[k,\ell]|^\alpha}{|x_e[k,\ell]|^\alpha + |x_r[k,\ell]|^\alpha}\,, \tag{5}$$

with α introducing a linear/nonlinear warping of the spectrum, e.g., $\alpha = 1$ for magnitude spectrum, $\alpha = 2$ for power spectrum, or $\alpha = 2/3$ motivated by auditory power law. On the other hand, it is important to search a proper feature representation as the DNN input, which is required to be strongly related to the targets as well as be robust in terms of discriminative power. Herein we apply a temporal modulation filter bank which is inspired by speech processing of the human auditory system to the conventional log-mel-spectrogram coefficients, so that temporal modulation knowledge from 0 until 50 Hz can be extracted.

3. EXPERIMENTAL SETUP

We adopt the evaluation test set from the REVERB Challenge [4], which contains six simulated test sets (three rooms with near and far speaker-to-microphone distances, denoted as s1n, s1f, s2n, s2f, s3n, s3f, with in total 2176 utterances), and one additional realistic recording scenario (in total 372 utterances with moderate ambient noise). Stationary background noise is added to the simulated test data with signal-to-noise ratios (SNRs) of $[\infty, 20, 15, 10, 5, 0]$ dB. The reference anechoic speech data of the simulated and realistic test sets are clean speech files from the WSJCAM0 British English corpus [29] and the headset microphone recordings from the MC-WSJ-AV corpus [30], respectively.

All simulations have been performed at a sampling frequency of 16 kHz, where the STFT has been computed using a 25 ms Hann window with 40% overlap, i.e., a hop size of 10 ms. For the spectral enhancement method (denoted SE), $g_{\min}$ in (2) is set to -10 dB, which provides a good compromise between late reverberation suppression and speech distortion (please refer to [3] for other parameter settings). For the single-channel linear prediction method (denoted SCLP), τ and L_p in (3) are set to 2 and 30 with a forgetting factor of 0.99 for the RLS algorithm.

For the mapping-based and masking-based supervised methods (denoted MAP and MASK), feedforward DNNs are used with five hidden layers, each having 2048 rectified linear hidden units. The mean squared error has been used as the cost function for the DNN training. For MAP DNN, a 201-dimensional reverberant spectrogram input (FFT length of 400 in STFT) as well as a length 11 context window (5 for left and 5 for right) is used, and the output layer represents the target 201-dimensional clean spectrogram. For MASK, we use 12 temporal modulation filters (cf. [22]) performing on the 40-dimensional log-mel-spectrogram to generate the auditory-inspired features as DNN input. Additional context window with length 3 (1 for left and 1 for right) is applied as well. The output layer represents the target 201-dimensional IRM in (5) with $\alpha = 1$, which provides slightly better results than other values of α in our pilot experiments. The training data for these two DNNs

was generated by convolving the anechoic utterances with the provided RIRs from the REVERB training set [4] but without additive noises, resulting in 7861 reverberant speech files (SNR = ∞ dB) in total. Note that these RIRs were recorded by an 8-channel microphone array with near and far positions and two different angles in six rooms different from those rooms for the simulated and realistic test sets [4], and all these 192 recorded RIRs (against 48 different recorded RIRs in the simulated test set) were involved in order to boost the generalization of observing various reverberation effects during DNN training.

The performance of the considered algorithms is evaluated in terms of speech quality and intelligibility, ASR performance and real-time applicability. The used objective measures include cepstral distance (CD) [31], frequency-weighted segmental signal-to-noise ratio (FWSSNR) [31], normalized speech-to-reverberation modulation energy ratio (SRMR) [32], perceptual evaluation of speech quality score (PESQ) [33] and short-time objective intelligibility score (STOI) [34]. Following the REVERB ASR *1ch* procedure [4], GMM/HMM is replaced by DNN/HMM as the acoustic modeling (cf. [3]). Besides the clean condition training (Cln) using only clean speech, we have also used multi-condition training (Mc) which is generated by using the reverberant speech (using only *1ch* of the RIRs provided by the REVERB training script, but without additive noises). Mc is also generated by using the dereverberated speech processed by each dereverberation algorithm on the reverberant training data (same amount of files as Cln for fair comparison). The real-time factor (RTF, defined as the computation time divided by the duration of the processed speech file) and the latency (determined as the time delay between the input and output speech) are evaluated as well.

4. RESULTS AND DISCUSSION

4.1. Objective Measures

For the simulated test set in a noise-free condition, Fig. 1 (a)-(e) illustrates the improvements of the considered objective measures for speech quality and intelligibility compared to the unprocessed reverberant speech data (as reference). In general, the average scores across all conditions (rooms and speaker-to-microphone distances) indicate that all considered algorithms are capable of improving the objective measures, although the algorithms behave quite differently w.r.t. different measures and reverberant conditions. SE generally yields minor but consistent improvements for all reverberant conditions, except in terms of speech intelligibility (predicted by STOI). On the other hand, SCLP only seems to yield improvements for high reverberant conditions (s2n, s2f, s3n, s3f) and even degrades the performance for low reverberant conditions (s1n, s1f), probably due to distortions caused by overestimating the late reverberation. On average, the performance improvements of SE and SCLP are quite similar (except for SRMR). Similarly, the performance improvement of MAP is quite inconsistent across different reverberant conditions, probably due to inconsistencies between the estimated normalized magnitude spectrogram and the unprocessed phase. On average, MAP hardly seems to yield any performance improvement (except for CD and SRMR). MASK outperforms all other considered algorithms (except in terms of SRMR improvement), showing that DNNs incorporating auditory-inspired features seem to be able to accurately estimate the IRM.

In addition, Fig. 1 (f) depicts the word error rate (WER) reduction for clean condition training, reflecting the proximity between features computed on dereverberated signals and clean reference features. It is clearly observed that on average more than 10% WER

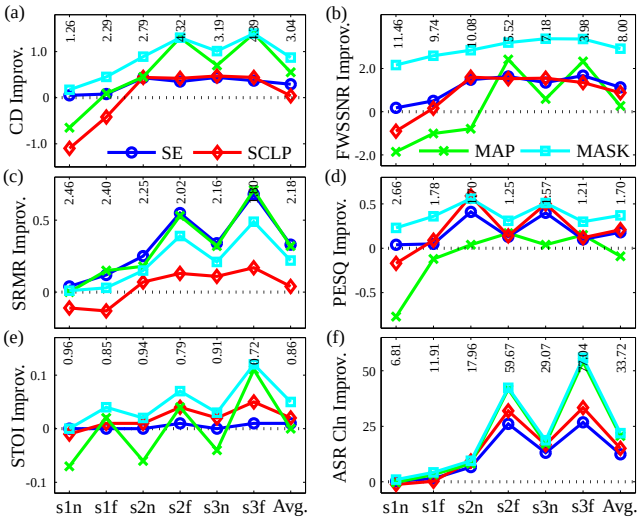

Fig. 1. Improvements of the objective measures for SE, SCLP, MAP and MASK for the simulated test set in a noise-free condition (SNR = ∞ dB) compared to the reference scores of the unprocessed reverberant speech data (at the top of each panel).

reduction can be achieved by all considered algorithms, where both supervised methods (MAP, MASK) perform very similarly and yield a relative WER reduction of nearly 50% compared to the unsupervised methods (SE, SCLP). This indicates that DNN-based dereverberation algorithms are able to preserve the clean speech feature variations that match with the clean trained model quite well.

4.2. ASR Performance

It is well known that multi-condition training is able to improve the performance of ASR systems. In order to further improve the ASR performance when the training set is not large enough to learn the feature variations caused by reverberation (like REVERB), speech dereverberation as front-end processing is a common strategy to remove reverberation-associated feature variances from both training and test data, cf. [3]. Fig. 2 compares the WER for clean condition training (Cln) and multi-condition training (Mc). For the unprocessed test data, it can be observed that Mc leads to a significant WER reduction of more than 20% on average compared to Cln. Moreover, the speech dereverberation algorithms (except for MAP) are able to further reduce the WERs by 1-1.5% on average over the unprocessed case. Compared to Cln, SCLP consistently performs slightly better than SE (except for the low reverberant conditions, i.e. s1n and s1f), and MASK consistently performs better than MAP. It is interesting to note that in contrast to Cln, MAP and MASK are not superior to SE and SCLP when using Mc. This is presumably

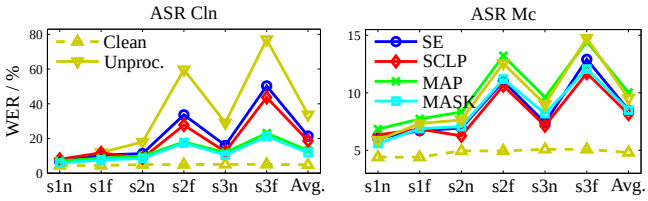

Fig. 2. ASR performance for the simulated test sets in a noise-free condition with clean and multi-condition training; clean speech with Cln is considered as the optimal result with an Avg. WER of 4.80%.

due to the fact that more reverberant feature variations are removed by MAP and MASK, which may benefit ASR with Cln but may hinder DNN generalization with Mc since too much variability from the training data has been removed.

4.3. Influence of Background Noise

Since background noise typically negatively affects the performance of dereverberation algorithms, in this section we investigate the robustness of the considered dereverberation algorithms (without any dedicated denoising strategies) against noise. For the simulated test set, Fig. 3 shows the improvement of the average scores (averaged across all reverberant conditions) for different SNRs. Although benefits can still be clearly observed for all considered algorithms, their robustness to noise is quite different.

For SE, SCLP and MASK, the improvements for most objective measures gradually decrease with decreasing SNR, meaning that the effectiveness of dereverberation degrades as the noise level increases. For example, for CD, PESQ and ASR (Mc) no obvious improvements can be observed at SNRs lower than 10 dB. On the other hand, the SRMR improvement of SE, SCLP and MASK seems to be relatively constant for different SNRs, presumably because the ratio of the modulation frequency energies is not so much affected by stationary noise. For MAP, it is interesting - but not entirely unexpected - that in general the improvements for most objective measures increase with decreasing SNR. In other words, MAP behaves quite robustly to noise, which may be explained by the underlying principle that the MAP DNN is trained to map non-target spectrogram information (i.e. reverberant and silent parts) to zero. Since the MAP DNN probably considers noise as non-target information, this results in denoising to some extent. Overall, MAP performs the best in terms of CD and FWSSNR across all SNRs, as well as the best in terms of SRMR, PESQ and ASR (Cln) at low SNRs (below 10 dB). At moderate SNRs (above 10 dB), SE provides the best SRMR measure and MASK leads to the best PESQ and ASR (Cln) scores. SCLP and MASK still seems to result in the best improvement of ASR (Mc) and STOI, respectively, even at low SNRs.

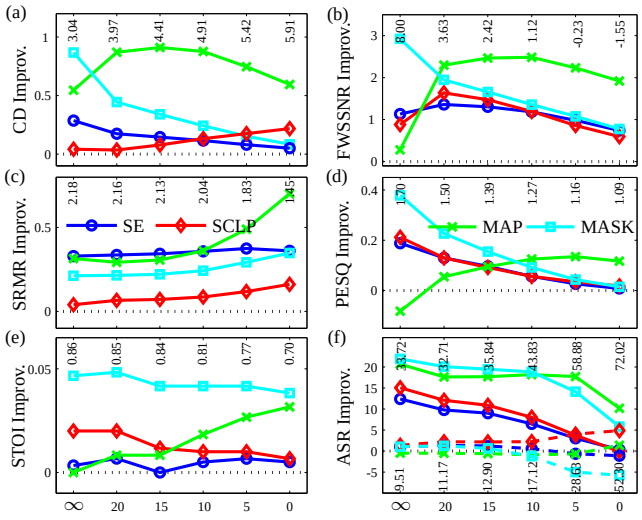

Fig. 3. (a)-(e): Improvements of the objective measures compared to the average reference scores of the unprocessed reverberant data (at the top of each panel) for different SNRs; (f): WER improvements for ASR with Cln (solid curves, reference WERs at the top) and Mc (dashed curves, reference WERs at the bottom).

4.4. Performance with Realistic Recordings

In order to evaluate the robustness of the considered dereverberation algorithms against variations that are not reproducible by simulation, Table 1 summarizes the performance for realistic recordings from a meeting room with moderate stationary ambient noise [4, 30]. Most results are in accordance with the results for the simulated reverberant and noisy data from Section 4.3, i.e., the best CD, PESQ, STOI and ASR (Cln) scores are obtained by MASK and MAP, while the best SRMR score is obtained by SE and the best ASR (Mc) score is obtained by SCLP. On the other hand, although Fig. 3 (b) suggests that MAP could offer the best FWSSNR score and SE the worst score, this is the other way around for the recordings, where the best FWSSNR score is obtained by SE.

Table 1. Objective measures and ASR performance with realistic recordings; WER is 6.98% for the headset mic recordings with Cln.

	CD	FWSSNR	SRMR	PESQ	STOI	WER Cln	WER Mc
Unproc.	4.87	-2.01	1.59	1.17	0.66	73.86	30.11
SE	4.72	**-0.21**	**2.49**	1.22	0.64	63.13	26.53
SCLP	5.02	-1.79	1.77	1.22	0.70	57.38	**25.97**
MAP	**3.72**	-3.27	2.46	**1.29**	0.75	39.86	31.28
MASK	4.44	-1.02	2.09	1.26	**0.76**	**36.55**	27.64

4.5. RTF and Latency

Table 2 lists the average RTF based on all test sets and the latency of each considered dereverberation algorithm. SE and SCLP were implemented in Matlab running on *GenuineIntel x86_64 64bits CPU 2.0 GHz* platform. For MAP and MASK, besides the feature extraction and the STFT analysis/synthesis implemented in Matlab, the estimated magnitude spectrograms or IRMs were obtained via DNNs compiled with a *Tesla K20c NVIDIA GPU 5 GB*. The inherent STFT latency is one block length (here 25 ms) for all algorithms. Due to the context window for DNN input, additional latencies of 5 and 1 block length (cf. Section 3) are required for MAP and MASK, respectively. MASK needs further latency of 24 block length because of the longest modulation filter with length of 49 for the auditory-inspired feature extraction. As a result, it shows that SE, MAP and MASK cost very low complexity, while SCLP is the most demanding computationally. MASK has longer latency compared to other methods, which could be reduced by selecting shorter modulation filters. Overall, RTFs below 1.0 and the low latencies indicate that all four dereverberation algorithms are fit to real-time applications.

Table 2. RTFs and latencies of the four dereverberation algorithms.

	SE	SCLP	MAP	MASK
RTF	0.028	0.679	0.035	0.052
Latency	25 ms	25 ms	75 ms	275 ms

5. CONCLUSIONS

In this paper, we presented a comparative study of two unsupervised and two supervised single-channel speech dereverberation algorithms. Results showed that spectral enhancement is capable of providing minor but consistent benefits for all reverberant conditions. Although single-channel linear prediction shows limited benefits in low reverberant scenarios, it performs the best on average for multi-condition trained ASR. In a noise-free condition, the IRM-based supervised approach provides the best overall objective measures and ASR performance using clean condition training, while the spectral mapping based supervised approach is the least sensitive to background noise. As part of future work, subjective listening tests should be performed to evaluate the correlation between the objective performance measures and subjective scores.

6. REFERENCES

[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons Ltd, United Kingdom, 2009.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[3] F. Xiong, B. T. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-End Technologies for Robust ASR in Reverberant Environments - Spectral Enhancement-based Dereverberation and Auditory Modulation Filterbank Features," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 70, pp. 1–18, 2015.

[4] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A Summary of the REVERB Challenge: State-of-the-art and Remaining Challenges in Reverberant Speech Processing Research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 7, pp. 1–19, 2016.

[5] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[6] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-Domain Single-Channel Inverse Filtering for Speech Dereverberation: Theory and Practice," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 5214–5218.

[7] K. Lebart, J.M. Boucher, and P.N. Denbigh, "A New Method based on Spectral Subtraction for Speech Dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, May 2001.

[8] E. A. P. Habets, S. Gannot, and I. Cohen, "Late Reverberant Spectral Variance Estimation based on a Statistical Model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

[9] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR Beamforming and Single-Channel Spectral Processing for Enhancing Noisy and Reverberant Speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 61, pp. 1–12, 2015.

[10] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation based on Variance-Normalized Delayed Linear Prediction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[11] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-Channel Linear Prediction-based Speech Dereverberation with Sparse Priors," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520, 2015.

[12] N. Roman and J. Woodruff, "Intelligibility of Reverberant Noisy Speech with Ideal Binary Masking," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. 2153–2161, 2011.

[13] O. Hazrati, J. Lee, and P. C. Loizou, "Blind Binary Masking for Reverberation Suppression in Cochlear Implants," *J. Acoust. Soc. Am.*, vol. 133, no. 3, pp. 1607–1614, 2013.

[14] D. Wang, *Speech Separation by Humans and Machines*, chapter On Ideal Binary Mask as the Computational Goal of Auditory Scene Analysis, pp. 181–197, Kluwer, Norwell, MA, USA, 2005.

[15] T. Yoshioka and T. Nakatani, "Dereverberation for Reverberation-Robust Microphone Arrays," in *Proc. European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[16] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive Speech Dereverberation using Constrained Sparse Multi-Channel Linear Prediction," *IEEE Signal Processing Letters*, in press, Nov. 2016.

[17] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning Spectral Mapping for Speech Dereverberation and Denoising," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[18] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[19] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets for Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A Regression Approach to Speech Enhancement based on Deep Neural Network," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[21] Y. Wang, K. Han, and D. Wang, "Exploring Monaural Features for Classification-Based Speech Segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[22] N. Moritz, J. Anemüller, and B. Kollmeier, "An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 23, no. 11, pp. 1926–1937, 2015.

[23] Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[24] H. Kuttruff, *Room Acoustics*, Spon Press, London, 4th edition, 2000.

[25] F. Xiong, S. Goetze, and B. T. Meyer, "Joint Estimation of Reverberation Time and Direct-to-Reverberation Ratio from Speech using Auditory-Inspired Features," in *ACE Challenge Workshop, satellite event of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.

[26] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE Spectral Magnitude Estimation for the Enhancement of Noisy Speech," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4037–4040.

[27] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[28] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively Trained Recurrent Neural Networks for Single-Channel Speech Separation," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, USA, Dec. 2014, pp. 577–581.

[29] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, MI, USA, May 1995, pp. 81–84.

[30] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, Nov. 2005, pp. 357–362.

[31] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[32] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An Updated Objective Intelligibility Estimation Metric for Normal Hearing Listeners under Noise and Reverberation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Juan les Pins, France, Sep. 2014.

[33] ITU-T, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," Feb. 2001.

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