On Joint Beamforming and Spectral Enhancement for Robust ASR in Reverberant Environments

Fanuel Melak Asmare[1,2], Feifei Xiong[1], Mathias Bode[2], Bernd T. Meyer[3] and Stefan Goetze[1]

[1]*Fraunhofer Institute for Digital Media Technology IDMT,*
Group Hearing, Speech and Audio Technology (HSA), Oldenburg, Germany
[2]*Jacobs University Bremen, Bremen, Germany*
[3]*University of Oldenburg, Dept. of Medical Physics and Acoustics, Oldenburg, Germany*
melakeegzi@gmail.com, feifei.xiong@idmt.fraunhofer.de

1. Introduction

Robust distant speech recognition greatly benefits those applications for which hands are not free to use and significantly increasesthe convenience of use, e.g., the source can move freely while the microphone(s) is/are placed in a certain position in distance to the speech source. Unfortunately, in distant talking scenarios the acoustic environment adds disturbances to the speech signal, i.e. problems arise due to interfering noises as well as reverberation caused by multiple reflections of the desired speech signal at the room boundaries and other objects in the room. This reverberation effect degrades the speech quality and speech intelligibility [1], as well as deteriorates the performance of automatic speech recognition (ASR) [2].

Reverberant speech can be perceived as sounding distant with noticeable coloration and echo. These detrimental perceptual effects generally increase with increasing the room volume, as well as with increasingthe distance between the source and the microphone [3]. Furthermore, with the spread in the time of arrival of reflections at the microphone, reverberation causes blurring of speech phonemes. Reduction of these detrimental effects is evidently of considerable practical importance. Although several pioneering efforts were made [4][5], it is still very challenging to compensatefor such long-term distortions. Nowadays, improving the robustness of the ASR systems in reverberant environments has been paid increased attention [6][7][8]. Still, overcoming the reverberation problem is paramount for realizing distant-talking speech recognizers.

The focus of this work is on removing the effects of reverberation by pre-processing of the speech signal by a front-end processing procedure before the feature extraction phase of the ASR systems. By applying different speech dereverberation strategies in single- and multi-microphone scenarios, we intend to alleviate the reverberation effect in order to improve the robustness of the ASR systems in different reverberant environments. We also use different short time Fourier transform (STFT) lengths to analyze the separation of early reflection and late reverberation tail w.r.t. the performance of ASR systems.

2. System Combination Strategies

As depicted in Fig.1, system (I) is based on a minimum mean square error (MMSE) estimator as pre-processor in each of the input channels of the beamformer. In this scenario, the late reverberation spectral variance (LRSV) estimation is carried out for the individual channels separately. Since the pre-processor does not change the phase of the signal, the spatial information for the beamformer will not be affected. System (II) uses an independent beamformer to get an enhanced spatially filtered signal followed by a single-microphone dereverberation system. Since the LRSV and the MMSE estimators use this spatially filtered signal, the computational complexity is less. The spatial filtering may, however, cause distortions to these estimator inputs due to the spatial correlation between microphone signals. To avoid the spatial correlation effect on the LRSV estimators [9] in system (II), system (III) uses a spatially averaged LRSV estimate obtained from all the microphone signals refined by the minimum variance distortionless response (MVDR) beamformer. System (IV) illustrates a multi-channel MMSE enhancement scheme [10], which can be decomposed into an MVDR beamformer followed by a single-channel Wiener filter. Here the MMSE estimator is actually a post-filter [11][12].

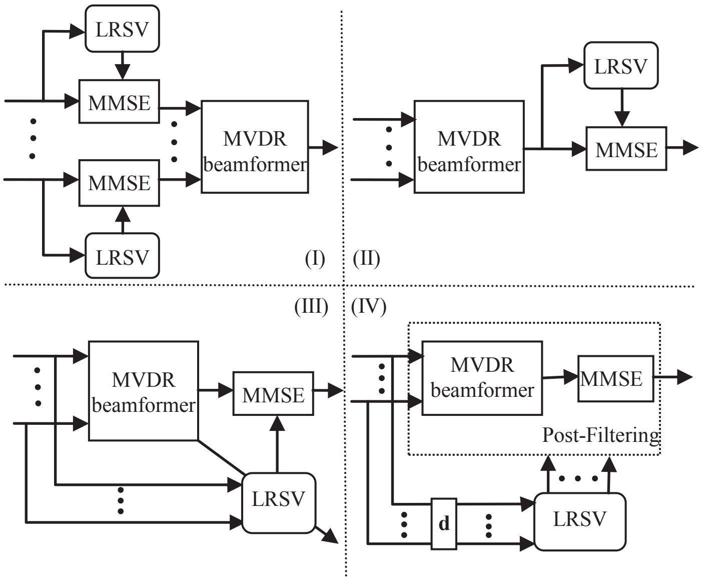

Fig. 1: Four different system combinations (I)-(IV) consisting of the MVDR beamformers and the MMSE estimators, as well as the respective LRSV estimators.

3. Spectral Enhancement for Dereverberation

The observed reverberant speech signal can be treated as the mixture of early reflections and the late reverberation, expressed in the STFT domain as,

$$X[\ell,k] = X_e[\ell,k] + X_l[\ell,k], \qquad (1)$$

where ℓ is the frame index and k is the frequency bin. The early reflections $X_e[\ell,k]$ are composed of the direct signal and early reflections, which are usually set to 20-80 ms [3]. As well, early reflections play an important role for enhancing the speech intelligibility. In contrast, late reverberation $\mathrm{X_l}[\ell, \mathrm{k}]$ degrades the signal quality [9][13]. In this paper the noise is neglected since the focus is to remove the late reverberation. In order to suppress the late reverberation effect, a spectral weighing function $\mathrm{G}[\ell, \mathrm{k}]$ determined by a parameterized MMSE estimator [14] which shows superior performance compared to the Wiener filter in our pilot experiments, is applied to the magnitude of the reverberant spectral variance in (1), resulting in [15],

$$\hat{\mathrm{X}}_{\mathrm{e}}[\ell, \mathrm{k}] = \max(\mathrm{G}[\ell, \mathrm{k}], \mathrm{G}_{\min})\, \mathrm{X}[\ell, \mathrm{k}], \qquad (2)$$

where $\mathrm{G}_{\min}$ is a lower bound of the weighting function. The a-priori early reflection to late reverberation energy rationecessary in (2) to calculate $\mathrm{G}[\ell, \mathrm{k}]$ is estimated by the decision-directed (dd) approach [16] which performs slightly better than the temporal cepstrum smoothing technique used in [8] for dereverberation in our pilot experiments. By this the LRSV $\lambda_l[\ell, k]$ estimation is required for (2) as also shown in Fig. 1. A generalized statistical reverberation model [17] is used here which separates the direct path from Polack's room impulse response (RIR) model [15], resulting in the spectral variance of the RIR $h[\mathrm{k}]$ in the STFT domain as

$$\lambda_h[\ell, \mathrm{k}] = \begin{cases} \beta_d[k], & \text{for } \ell = 0 \\ \beta_r[k] e^{-\delta[k]\ell \mathrm{R}}, & \text{for } \ell \geq 1 \end{cases} \qquad (3)$$

where the decay coefficient is related to the reverberation time T_{60} by $\delta = 3\ln(10)/T_{60}$. β_d and β_r denote the variances of the direct path and the reverberant part and R represents the STFT frame shift, i.e. the hop size. The direct signal to reverberation ratio (DRR) can then be expressed as [17]

$$DRR = 10\log_{10}\left(\frac{1-e^{-2\delta\ell\mathrm{R}}}{e^{-2\delta\ell\mathrm{R}}}\frac{\beta_d}{\beta_r}\right). \qquad (4)$$

The DRR is related to the clarity index due to the frame shift R [3]. Now the reverberation variancecan be obtained using (4) as [16] (ignoring the frequency index for simplicity)

$$\lambda_r[\ell] = (1-\kappa)\mathrm{e}^{-2\delta\mathrm{R}}\, \lambda_\mathrm{r}[\ell-1] + \kappa \mathrm{e}^{-2\delta\mathrm{R}}\, \lambda_\mathrm{x}[\ell-1], \quad (5)$$

where$\kappa = \beta_d/\beta_r$ is calculated from the DRR in (4), constraint inthe range of (0, 1]. Then, the LRSV is given by

$$\lambda_l[\ell] = \mathrm{e}^{-2\delta\mathrm{R}(\mathrm{N}_1-1)} \lambda_\mathrm{r}[\ell - \mathrm{N}_1 + 1], \qquad (6)$$

where N_1 denotes the number of frames which corresponds to the duration of early reflections of the RIR. An instantaneous estimate of the input reverberant spectral variance $\lambda_\mathrm{x}[\ell]$ in (5) can be obtained by a smoothed version of $|X[\ell]|^2$ as

$$\hat{\lambda}_\mathrm{x}[\ell] = \eta[\ell]\hat{\lambda}_\mathrm{x}[\ell-1] + (1-\eta[\ell])|X[\ell]|^2, \qquad (7)$$

where the smoothing constant η is calculated by $\eta = 1/(1 + 2\delta R)$. According to [9], in order to improve the tracking performance of the reverberant speech onset, η shall be set to be lower than η_att when $|X[\ell]|^2 > \lambda_\mathrm{x}[\ell]$. Note that such an LRSV estimator requires a-priori information of T_{60} and DRR or clarity index at least in full-bandmode, which in practice can be estimated by [18][19].

4. Dereverberation by Multi-microphone Beamforming and Post-Filtering

When multiple microphones are available, beamforming and post-filtering techniques can be used for the purpose of dereverberation [20][21][22][23]. The MVDR beamformer which performs best in diffuse interference fields is used here. This beamformer minimizes the output power while keeping a unity gain in the desired direction and its coefficients can be derived as,

$$\mathbf{W}[\ell] = \mathbf{\Gamma}_{\mathbf{vv}}^{-1}[\ell]\mathbf{d}/(\mathbf{d}^\mathrm{H}\mathbf{\Gamma}_{\mathbf{vv}}^{-1}\mathbf{d}), \qquad (8)$$

where $(\cdot)^\mathrm{H}$ is the Hermitian transpose and $\mathbf{d}$ is the steering vector. In order to vary the beamformer used, the coherence matrix of the interfering signals $\mathbf{\Gamma}_{\mathbf{vv}}$ is replaced by a diffuse interference field $\mathbf{\Gamma}_\mathrm{diff}$ for the superdirective (SD) beamformer [24] or by the identity matrix $\mathbf{I}$ for the delay and sum (DS) beamformer.

A post-filtering approach is used to calculate the coefficients in system (IV) [11] which can be expressed as an MVDR beamformer with a post-filter $\mathbf{W}^{(\mathrm{IV})}[\ell] = \mathrm{P}[\ell]\mathbf{W}[\ell]$ with the post filter transfer function [12] expressed by

$$P[\ell] = \max\left(\frac{\frac{2}{M(M-1)}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\tilde{\phi}_{x_e x_e}^{(ij)}[\ell]}{\frac{1}{M}\sum_{i=1}^{M}\tilde{\phi}_{xx}^{(ij)}[\ell]}, P_\mathrm{min}\right), \qquad (9)$$

with $\tilde{\phi}_{xx}^{(ij)}$being the auto-correlation of the speech signal in microphone channel i. In order to alleviate speech distortions in ASR systems, a lower bound P_min is introduced. The early reverberation variance (the cross correlation term) can be estimated as [11]

$$\tilde{\phi}_{x_e x_e}^{(ij)}[\ell] = \frac{\Re\{\tilde{\phi}_{xx}^{(ij)}[\ell]\} - \frac{1}{2}\Re\{\Gamma_{x_l x_l}^{(ij)}\}\left(\tilde{\phi}_{xx}^{(ii)}[\ell] + \tilde{\phi}_{xx}^{(jj)}[\ell]\right)}{1 - \Re\{\Gamma_{x_l x_l}^{(ij)}\}}, \qquad (10)$$

where $\Re\{\cdot\}$ calculates the real part of a complex signal. A time alignment is required for X when calculating (10) [11], which can beachieved by the steering vector $\mathbf{d}$ as seen in Fig. 1 (IV). In (10) a first-order recursive update of the auto- and cross-correlation calculationsis applied and a maximal threshold is introduced to avoid the denominator being non-positive. The LRSV coherence matrix in (8) can also be replaced by the LRSV coherence $\mathbf{\Gamma}_{x_l x_l}$ estimated from the M received microphone signals.

5. Experimental Results

The WSJCAM0 British English corpus [25] was used as database of clean (anechoic) speech utterances. It contains 7861 utterances for training and another 742 for testing at a sampling rate of 16 kHz. 18 real-world RIRs recorded by a circular microphone array (M = 8) with 20 cm diameter from the REVERB Challenge [7] were used for multi-condition training mode and another 6 RIRs [7] for generating various test sets (denoted by T1-T6 in the following) with different T_{60} and DRR values, as listed in Tab. 1. The STFT was computed using a Hanning window with two different analysis window lengths, 32 ms with 1/8 overlap (short term) and 96 ms with 1/2 overlap (long term). A white noise gain constraint of 10 dB was selected for the MVDR beamformer in (8). The weighting factor of 0.5 was used in the dd approach. η_att was chosen as 0.7η in (7). $\mathrm{G}_{\min}$ in (2) was set to -10 dB as a good value to the ASR performance.

$P_{\min}$ in (9) was chosen as 0.1 and the smoothing factor of the first-order recursive filter was 0.875 in (10). Directly from the RIRs in full-band mode, DRR or C_4/C_{48} was calculated accordingly and T_{60} was determined by using Schroeder's method [26].

Test	T1	T2	T3	T4	T5	T6
Room	Small	Small	Medium	Medium	Large	Large
Position	Near	Far	Near	Far	Near	Far
DRR (dB)	**17.73**	**4.56**	**11.42**	**0.25**	**10.40**	**-1.70**
T_{60} (ms)	**218.29**	**229.91**	**500.06**	**519.44**	**719.32**	**747.38**

Tab.1 Characteristics of all test sets T1-T6 with mean DRR and T_{60} values (from all 8 channels).

The framework for the ASR experiments was implemented based on the Hidden Markov Model Toolkit (HTK) [27]. Overlapping speech segments of 25 ms duration and 10 ms shift were used for the calculation of mel-frequency cepstral coefficients with delta and double-delta coefficients as well as cepstral mean and variance normalization. Context-dependent triphone hidden Markov models with 3 states per model were applied together with 12 Gaussian mixture models per state and a language scaling factor of 14.0 for the 5k-word-bigram language model.

Fig. 2 shows the word error rate (WER) results of our systems under test with 32 ms STFT analysis window length. The baseline results come from the multi-condition training with the original reverberant speech signal from the first microphone, i.e. i=1. For single-microphone scenarios, the MMSE estimator is applied to the first channel. For comparison, the results with beamformers alone (either with SD or DS) are presented. The rest are the system combination outputs (cf. Fig. 1). Both the MMSE and the beamformer alone scenarios improve the ASR output, with the SD beamformer showing lower WERs of approximately 4% compared to the single microphone MMSE estimator.

The performances of system (I) and the MVDR-only system show very similar results. For test sets T1, T3 and T5, i.e. the near-position tests, higher WERs can be observed when compared to the SD beamformer alone. This may be caused by the distortions of the diffuse field because of the front MMSE estimators. Compared to system (I), 1% WER improvement can be obtained by system (II). A more accurate LRSV estimate is employed in system (III) which results in a slightly better performance than system (II). This indicates that, the spatial correlation introduced by the beamformer blurs the MVDR-filtered RIR in system (II) so that it cannot exactly extract the true late reverberation.

Overall, average WER reduction of 6.17% is obtained by system (III) with the SD beamformer compared to the baseline. Such improvements become more obvious for the far-position testsets such as T4 and T6 than the near-position test sets such as T3. A similar trend can be observed for system (IV), for which the SD beamformer still performs best. It can also be observed that the results degrade when the beamformer in (1) uses the LRSV coherence matrix. A possible explanation is that the late reverberation behaves non-stationary and its coherence actually does not match the diffuse property, especially for the near-position test sets T1, T3 and T5 as discussed in [28].

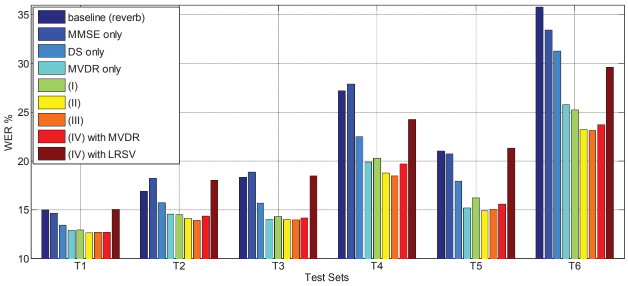

Fig.2 WERs of the dereverberation strategies with 32 ms STFT analysis window length and 1/8 hop size.

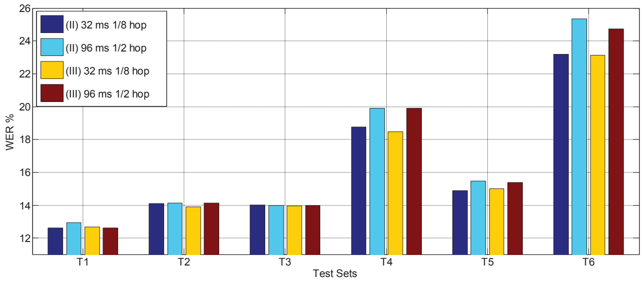

Fig.3 WERs of system (II) and (III) with two different frame lengths, i.e. short and long term STFT analysis window length.

Fig. 3 compares the WERs of the two best performing systems (II) and (III) with two different STFT analysis window sizes. Using shorter frame sizes benefits in reducing WERs. The results shows that the 32 ms with 1/8 hop window improves the WER by 4.18% and 3.57% compared to the 96 ms window with 1/2 overlap window with systems (II) and (III) respectively.

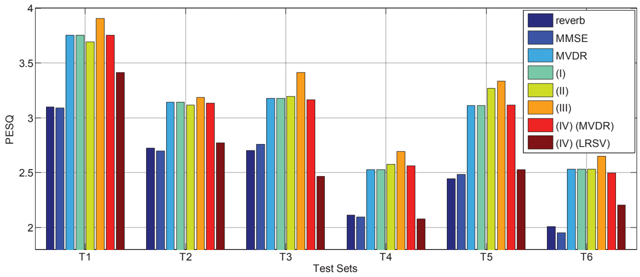

Fig.4 PESQ scores from the output of different systems; a male utterance from the test data is employed and the respective clean (anechoic) speech is used as the reference signal.

A perceptual evaluation of speech quality (PESQ) [29] has also been conducted. Fig. 4 shows the PESQ scores of the different proposed systems with one male test utterance. The performance of multi-microphone dereverberation strategies in PESQ tests is much better than that of single-microphone approaches. Here system (III) shows the best results compared to all other scenarios, which is in consilience with the WERs results in Fig. 2.

6. Conclusion

This work explored possible combination architectures for dereverberation by (single-microphone) spectral enhancement schemes and (multi-microphone) beamforming with the aim of improving ASR performancein various reverberant environments. Results indicate that all the combined systems are able to provide benefits for ASR systemsand specifically, the system (III) combining the SD beamformer and the MMSE estimator with the LRSV refinement by the MVDR beamformer coefficients achieves nearly 30% average relative WER improvement compared to the baseline, as well as 15% average relative PESQ boost (from one example) compared to the first channel reverberant speech signal. As well, short STFT analysis window length provides better ASR performance than a longer window length.

7. Literature

[1] P.A. Naylor, N.D. Gaubitch, and E. Cross, "Speech Dereverberation,"*Noise Control Engineering Journal*, 59(2): 211–211, 2011.

[2] M. Wölfel and J. McDonough,"Distant Speech Recognition,"*John Wiley & Sons Ltd*, 2009.

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

[4] B.E.D. Kingsbury, N. Morgan, and S. Greenberg,"Robust Speech Recognition using the Modulation Spectrogram,"*Speech communication*, 25(1):117-132, 1998.

[5] M.L. Seltzer, B. Raj, and R.M. Stern,"Likelihood-Maximizing Beamforming for Robust Hands-Free Speech Recognition,"*Speech and Audio Processing, IEEE Transactions on*, 12(5):489–498, 2004.

[6] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann,"Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberationfor Automatic Speech Recognition,"*IEEE Signal Processing Magazine*, 29(6):114–126,2012.

[7] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj,"The Reverb Challenge: A common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* 2013.

[8] F. Xiong, N. Moritz, R. Rehr, J. Anemüller, B.T. Meyer, T. Gerkmann, S. Doclo, and S. Goetze,"Robust ASR in Reverberant Environments using Temporal Cepstrum Smoothing for Speech Enhancement and an Amplitude Modulation Filterbank for Feature Extraction,"*REVERB challenge*, Florence, 2014.

[9] E.A.P. Habets, "Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement,"*Ph.D. thesis*, University of Eindhoven, Eindhoven, The Netherlands, Jun. 2007.

[10] K.U. Simmer, J. Bitzer, and C. Marro, "Microphone Arrays", Chapter "Post-Filtering Techniques," pp. 39–60, *M. Brandstein and D. Ward (Eds.), Springer*, Berlin, Heidelberg, May 2001.

[11] I.A. McCowan and H. Bourlard, "Microphone Array Post-Filter based on Noise Field Coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–715, Nov. 2003.

[12] R. Zelinski, "A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *IEEE Int. Conf. on Acoustics,Speech, and Signal Processing (ICASSP)*, New York, NY, USA,Apr. 1988, vol. 5, pp. 2578–2581.

[13] P.C. Loizou,"Speech Enhancement: Theory and Practice,"*CRC press*, 2013.

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

[15] K. Lebart, J.M. Boucher, and P.N. Denbigh,"A New Method based on Spectral Subtraction for Speech,"*ActaAcusticaunited with Acustica*, 87(3):359–366, 2001.

[16] 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.

[17] 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.

[18] J. Eaton, N.D. Gaubitch, and P.A. Naylor, "Noise-Robust Reverberation Time Estimation using Spectral Decay Distributions with Reduced Computational Cost," *in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 161–165.

[19] F. Xiong, S. Goetze, and B.T. Meyer, "Blind Estimation of Reverberation Time based on Spectro-Temporal Modulation Filtering," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 443–447.

[20] J.B. Allen, D.A. Berkley, and J. Blauert, "Multi-microphone Signal Processing Technique to Remove Room Reverberation from Speech Signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[21] C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques based onMicrophone Arrays with Postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[22] M. Jeub and P. Vary, "Binaural Dereverberation based on A Dual-Channel Wiener Filter with Optimized Noise Field Coherence," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4710–4713.

[23] A.Westermann, J.M. Buchholz, and T. Dau, "Binaural Dereverberation based on Interaural Coherence Histograms," *J. Acoust. Soc.* Am., vol. 133, no. 5, pp. 2767–2777, 2013.

[24] J. Bitzer and K.U. Simmer, "*Microphone Arrays*, chapter Superdirective Microphone Arrays," pp. 19–38, *M. Brandstein and D. Ward (Eds.), Springer*, Berlin, Heidelberg, May 2001.

[25] 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, Michigan, USA, May 1995, pp. 81–84.

[26] M.R. Schroeder, "New Method of Measuring Reverberation Time," *J. Acoust.Soc. Amer.*, vol. 37, no. 3, pp. 409–412, 1965.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.4),"*Cambridge University Engineering Department,* Cambridge, 2009, http://htk.eng.cam.ac.uk/.

[28] F.Xiong, B.T. Meyer, and S. Goetze, " A Study on joint Beamforming and Spectral Enhancement for Robust Speech Recognition in Reverberant Environments," in *IEEE Int.Conf. on Acoustics, Speech, and Signal Processing (ICASSP),* Brisbane, Australia, Apr. 2015.

[29] 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.