SOUND SOURCE SEPARATION USING INTER-MICROPHONE DELAY, CROSS-CORRELATION, AND CROSS-COVARIANCE

Chanwoo Kim[1]*, Anjali Menon*[2]*, and Richard M. Stern*[2]

[1] Samsung Research
[2] Carnegie Mellon University
chanw.com@samsung.com, {anjalim, rms}@cmu.edu

ABSTRACT

In this paper, we present a two-microphone algorithm for sound source separation referred to as Minimum Normalized Cross-Correlation and Minimum Normalized Cross-coVariance (MNCC-MNCV). The MNCC-MNCV algorithm automatically selects the "Inter-Microphone Delay (IMD) threshold," which separates time-frequency segments that are believed to belong to the target from components that are dominated by noise sources. In this work, we develop an algorithm which considers both the normalized correlation and the correlation coefficient of the target and interfering signal signals, after a compressive nonlinearity. We demonstrate that the use of the MNCC-MNCV method of estimating IMD thresholds provides significantly better recognition accuracy in the presence of interfering speech sources, omni-directional noise, and reverberation than several other approaches including ZCAE [1], PDCW [2], and our previous automatic IMD threshold selection approach [3].

Index Terms: Robust speech recognition, signal separation, phase difference analysis, cross-correlation, normalized correlation

1. INTRODUCTION

Recently, improvement in deep learning technology [4, 5, 6, 7, 8, 9] has helped to commercialize voice assistant speakers [10, 11]. In far-field speech recognition environments, the impact of noise and reverberation is much larger than near-field cases. To enhance robustness of speech recognition systems, data augmentation techniques have been widely used with notable success [12, 13, 11, 14, 15, 16, 17]. Multi-microphone approaches have been also frequently used to select the target sound source while suppressing noisy sources [18, 19, 20, 21]. Various variations of Minimum Variance Distortionless Response (MVDR) beam-formers have been used with success [22, 23, 24, 25]. Many algorithms have been proposed that segregate signals on the basis of differences in arrival time (*e.g.* [26, 27, 28, 1, 29]). It is well documented that the human binaural system has a remarkable ability to separate speech that arrives from different azimuths (*e.g.* [30, 31]) using various types of acoustical cues are used to segregate the target signal from interfering sources. Motivated by these observations, a number of models and algorithms have been developed that separate signals according to Inter-Microphone Delays (IMDs) (*e.g.* [28, 1]), inter-microphone intensity difference (IIDs), inter-microphone phase differences (IPDs) (*e.g.* [32], [33]), as well as other cues. IMDs can be estimated using either phase differences (*e.g.* [2]), cross-correlation, or zero-crossings (*e.g.* [1]). Typically these algorithms compare IMDs or IPDs in each spectro-temporal segment of an input to a fixed threshold, in order to select those IMDs or IPDs that are consistent with the azimuth of the desired target speaker. In [3], we proposed an algorithm that selects the IMD threshold automatically by minimizing the correlation coefficient of power signals after the peripheral auditory nonlinearities. While this algorithm was found to be more robust than algorithms using fixed thresholds (*e.g.* [2]), this algorithm is not very effective for practical environments that include multiple noisy sources. In the present work, by considering both the cross-correlation and cross-covariance between the masked and unmasked channels, with additional Channel Weighting (CW), we obtain significant improvements for various conditions.

Thanks to Samsung Electronics for funding this research. The authors are thankful to Executive Vice President Seunghwan Cho and speech processing Lab. members at Samsung Research.

2. STRUCTURE OF THE MNCC-MNCV ALGORITHM

In this section, we explain the structure of our binaural sound source separation system. While the detailed description below assumes a sampling rate of 16 kHz, this algorithm is easily modified to accommodate other sampling frequencies. Our approach is based on Phase Difference Channel Weighting (PDCW) [2]. If the automatic threshold selection algorithm described in Sec. 2.2 is employed to obtain the target IMD threshold, we refer to the entire system as Minimum Normalized Cross-Correlation and Minimum Normalized Cross-coVariance (MNCC-MNCV). In this paper, we consider three different ways of selecting the IMD threshold: the Minimum Normalized Cross-coVariance (MNCV) which had been introduced in [3], a Minimum Normalized Cross-Correlation (MNCC), and a combination of these two approaches (MNCC-MNCV). These approaches will be explained in detail in Sec. 2.2. A block diagram of the MNCC-MNCV system is shown in Fig. 1. The system first performs a short-time Fourier transform (STFT) which decomposes the two input signals in time and in frequency. We use Hamming windows of duration 75 ms with 37.5 ms between frames, and a DFT size of 2048. The IMD is estimated indirectly by comparing the phase information from the two microphones at each frequency. The time-frequency mask identifying the subset of IMDs that are "close" to the IMD of the target speaker is obtained using the IMD threshold selection algorithm described in Sec. 2.2. To obtain better speech recognition accuracy in noisy environments, we apply a gammatone channel-weighting approach as described in [2] instead of directly applying the binary mask. This step is another improvement over our previous approach in [3]. Finally, the time domain signal is obtained using the IFFT and the OverLap-Add (OLA) method.

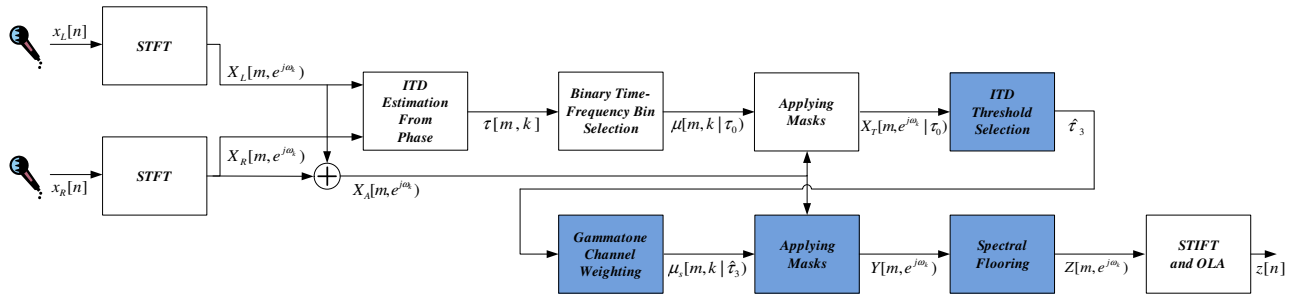


Fig. 1. *Block diagram of a sound source separation system using the MNCC-MNCV algorithm.*

2.1. Source Separation Using IMDs

In binaural sound source separation systems we usually assume that we have *a priori* knowledge about the target location. In this paper we assume that the target is located along the perpendicular bisector to the line connecting two microphones. Hence, the decision criteria can be expressed as follows:

$$\begin{cases} \text{considered to be a target:} & |\tau[m,k]| < \tau_0 \\ \text{considered to be a noise source:} & |\tau[m,k]| \geq \tau_0 \end{cases} \quad (1)$$

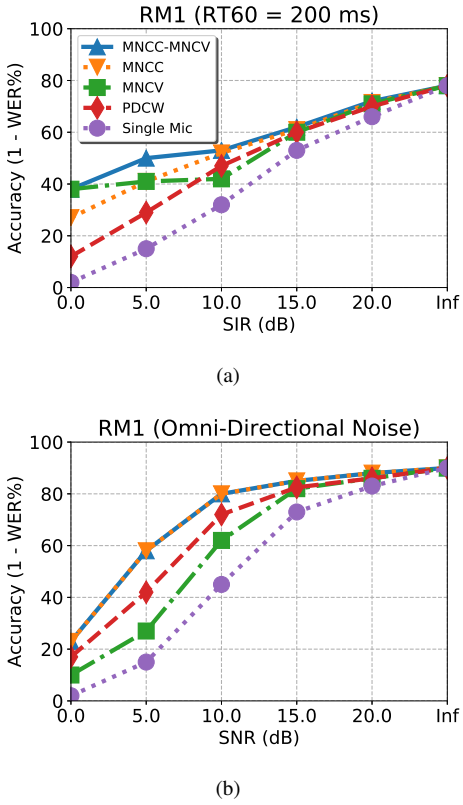
where m is the frame index, k is the frequency index, $\tau[m,k]$ is the estimated IMD for a time-frequency bin represented by $[m,k]$, and τ_0 is the threshold IMD. Thus, if we obtain a suitable IMD threshold using (1), we can make a binary decision. In this sound-source separation system the IMD is obtained for each time-frequency bin using phase information according to (1). In our previous work (*e.g.* [3]), it has been shown that the IMD $\tau[m,k]$ for each time-frequency bin can be obtained by:

$$|\tau[m,k]| \approx \frac{1}{|w_k|} \min_r \Big| \angle X_R[m,e^{-jw_k}] - \angle X_L[m,e^{-jw_k}] - 2\pi r \Big|, \; 0 \leq k \leq \frac{K}{2} \quad (2)$$

where m is the frame index, k is the frequency index, $X_L[m,e^{-jw_k}]$ and $X_R[m,e^{-jw_k}]$ are STFTs of the signals from the left and right microphones, receptively. ω_k is the discrete-time frequency defined to be $\omega_k = \frac{2\pi k}{K}$, and K is the DFT size. Using an IMD threshold τ_0, we construct a binary mask as in [3]. The procedure of obtaining an automatic threshold is described in Sec. 2.2.

2.2. Optimal IMD threshold selection using complementary masks

To find the optimal IMD threshold, computation is performed in discrete fashion, considering a set $\mathcal{T}$ of a finite number of possible IMD threshold candidates. The set $\mathcal{T}$ is defined by the following minimum and maximum values of the IMD threshold: $\tau_{min} = d\sin(\theta_{0,min})f_s/c$ and $\tau_{max} = d\sin(\theta_{0,max})f_s/c$, where d is the distance between two microphones, c is the speed of sound, f_s is the sampling frequency, and $\theta_{0,min}$ and $\theta_{0,max}$ are the minimum and the maximum values of the threshold angle. In the present implementation, we use values of $\theta_{0,min} = 5^\circ$ and $\theta_{0,max} = 45^\circ$. We use a set of candidate IMD thresholds $\mathcal{T}$ that consist of the 20 linearly-spaced values of θ_0 between $\theta_{0,min}$ and $\theta_{0,max}$. We determine which element of this set is the most appropriate IMD threshold by performing an exhaustive search over the set $\mathcal{T}$. Let us consider one element of this set, $\tau_0 = d\sin(\theta_0)f_s/c$. Using



(a)

(b)

Fig. 2. *Comparison of recognition accuracy for the DARPA RM database (a) corrupted by an interfering speaker located at 30 degrees under the reverberation of $T_{60} = 200ms$, and (b) in the presence of natural real-world noise.*

the procedure described in Sec. 2.1, we obtain the target spectrum $X_T[m,e^{j\omega_k}|\tau_0], \quad 0 \leq k \leq \frac{K}{2}$ as shown below:

$$X_T[m,e^{j\omega_k}|\tau_0] = X_A[m,e^{j\omega_k}]\mu[m,k|\tau_0] \quad (3)$$

where $X_A[m,e^{j\omega_k}]$ is the averaged signal spectrogram from the two microphones defined by

$$X_A[m,e^{j\omega_k}] = (X_L[m,e^{j\omega_k}] + X_R[m,e^{j\omega_k}])/2. \quad (4)$$

$\mu[m,k|\tau_0]$ is the binary mask obtained by (1). In the above equation we explicitly include τ_0 to show that the masked spectrum depends on the IMD threshold. Using this spectrum $X_T[m,e^{j\omega_k}]$, we obtain the target power and the power of the interfering sources. Since

everything which is not the target is considered to be an interfering source, the power associated with the target and interfering sources can be obtained by the following equations:

$$P_T[m|\tau_0] = \sum_{k=0}^{K-1} \left| X_T[m, e^{j\omega_k}|\tau_0] \right|^2 \tag{5a}$$

$$P_I[m|\tau_0] = \sum_{k=0}^{K-1} \left| X_A[m, e^{j\omega_k}] \right|^2 - P_T[m|\tau_0] \tag{5b}$$

As in [3, 34, 35], a compressive nonlinearity is invoked:

$$R_T[m|\tau_0] = P_T[m|\tau_0]^{a_0} \qquad R_I[m|\tau_0] = P_I[m|\tau_0]^{a_0} \tag{6a}$$

where $a_0 = 1/15$ is the power coefficient as in [3, 34].

In general, the optimal IMD threshold is determined by identifying the value of τ_0 that minimizes the cross-correlation between the signals $R_T[m|\tau_0]$ and $R_I[m|\tau_0]$ from (6), but there are several plausible ways of computing this cross-correlation. The first method considered, which was used in an earlier paper [3], is based on the normalized cross-covariance of the signals in (6):

$$\rho_{T,I}(\tau_0) = \frac{\frac{1}{K}\sum_{m=1}^{M} R_T[m|\tau_0]R_I[m|\tau_0] - \mu_{R_T}\mu_{R_I}}{\sigma_{R_T}\sigma_{R_I}} \tag{7}$$

where μ_{R_1} and μ_{R_2}, and σ_{R_T} and σ_{R_I}, are the means and standard deviations of $R_T[m|\tau_0]$ and $R_I[m|\tau_0]$, respectively. (This statistic is also known as the Pearson product-moment correlation.)

The optimal IMD threshold τ_0 is selected to minimize the absolute value of the cross-covariance:

$$\hat{\tau}_1 = \arg\min_{\tau_0} |\rho_{T,I}(\tau_0)| \tag{8}$$

Again, we refer to this approach as the MNCV statistic, and it has provided good speech recognition accuracy as shown in Fig. 2, especially at low SNRs such as 0 or 5 dB. Nevertheless, at moderate SNRs such 10 or 15 dB, the speech recognition accuracies obtained using MNCV processing are worse than those obtained using the PDCW algorithm. MNCC-MNCV processing using the MNCV statistic also provides poor recognition accuracy in the presence of omnidirectional natural noise, as shown in Fig. 2(b). We have also found in pilot studies that the MNCV statistic in (7) is not a helpful measure in situations where there is a single interfering source with power that is comparable to that of the target, or where there are multiple interfering sources.

To address this problem, we consider a second related statistic, the normalized cross-correlation:

$$r_{T,I}(\tau_0) = \frac{\frac{1}{K}\sum_{m=1}^{M} R_T[m|\tau_0]R_I[m|\tau_0]}{\sigma_{R_T}\sigma_{R_I}} \tag{9}$$

$$\hat{\tau}_2 = \arg\min_{\tau_0} |r_{T,I}(\tau_0)| \tag{10}$$

We refer to implementations using $\hat{\tau}_2$ as Minimum Normalized Cross-correlation (MNCC) systems.

The final IMD threshold $\hat{\tau}_3$ is obtained easily by calculating the minimum of $\hat{\tau}_1$ and $\hat{\tau}_2$ as shown below:

$$\hat{\tau}_3 = \min(\hat{\tau}_1, \hat{\tau}_2) \tag{11}$$

We refer to implementations using $\hat{\tau}_3$ as MNCC-MNCV systems. As can be seen in Figs. 2(a) and 2(b), systems using the MNCC-MNCV stastistic consistently provide recognition accuracy that is similar to or better than that obtained using either the MNCV or MNCC approaches. For these reasons we adopt MNCC-MNCV processing as the default approach, unless some other approach is stated explicitly. Instead of directly applying the mask $\mu[m, k|\hat{\tau}_3]$, we obtain a smoothed mask $\mu_s[m, k|\hat{\tau}_3]$ using the Channel Weighting (CW) technique in [2]. By applying $\mu_s[m, k|\hat{\tau}_3]$ to the averaged spectrum $X_A[m, e^{j\omega_k}]$, we obtain the enhanced spectrum $Y[m, e^{j\omega_k}]$ as follows:

$$Y[m, e^{j\omega_k}] = X_A[m, e^{j\omega_k}]\mu_s[m, k|\hat{\tau}_3] \tag{12}$$

2.3. Spectral flooring

In our previous work (*e.g.* [34]), it has been frequently observed that an appropriate flooring helps in improving noise robustness. For this reason we also apply a flooring level to the spectrum, which is described by the equation:

$$Y_f = \delta_f \sqrt{\frac{1}{N_f K} \sum_{m=0}^{N_f-1} \sum_{k=0}^{K-1} \left| Y[m, e^{j\omega_k}] \right|^2} \tag{13}$$

where δ_f is the flooring coefficient, N_f is the number of frames in the utterance, K is the FFT size, and Y_f is the obtained threshold. We use a value of 0.01 for the flooring coefficient δ_f. Using the flooring level Y_f, the floored spectrum $Z[m, e^{j\omega k}], 0 \leq k \leq K$ is obtained as follows:

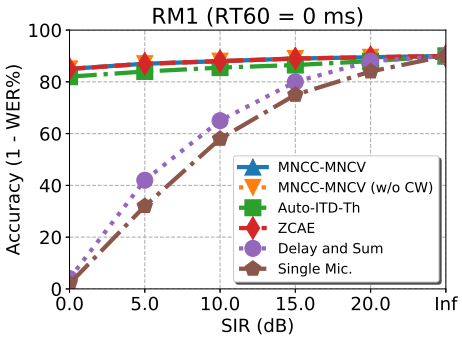
$$Z[m, e^{j\omega k}] = \max\left[\left| Y[m, e^{j\omega_k}] \right|, Y_f \right] e^{j\angle Y[m, e^{j\omega_k}]} \tag{14}$$

Using $Z[m, e^{j\omega k}]$, speech is resynthesized using IFFT and the overlap-add method (OLA). While the improved threshold optimization criteria in Sec. 2.2 is the most important improvement over our previous approach in [3], Channel Weighting (CW) and this spectral flooring are also additional improvements.

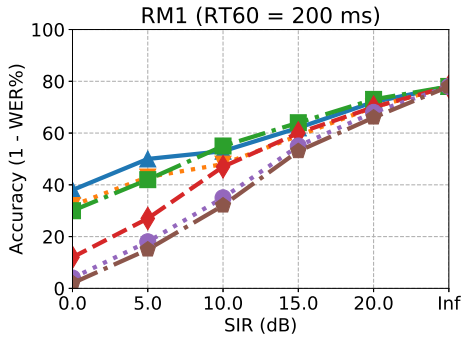
3. EXPERIMENTAL RESULTS

In this section, we present experimental results using the MNCC-MNCV algorithms described in this paper, focussing on the performance of the new methods that were developed to estimate blindly the IMD threshold parameter. To evaluate the effectiveness of this MNCC-MNCV algorithm, we performed comparison with the source separation algorithm developed in [3], which is referred to as "Auto-ITD-Th" in this section. To see the additional benefit of Channel Weighting (CW), we also compared the performance when the Channel Weighting (CW) approach is not employed in MNCC-MCVN.

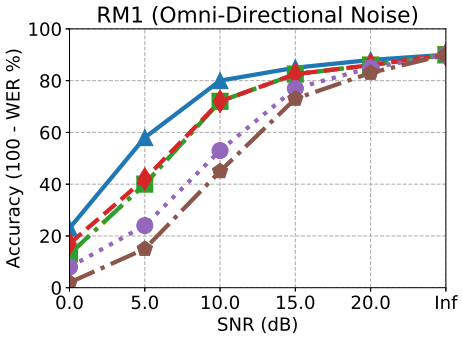
We also compare our approach with an earlier technique, the ZCAE algorithm described in [1] using binary masking as in [3]. In all speech recognition experiments described in this paper we perform feature extraction using the version of MFCC processing implemented in `sphinx_fe` in `sphinxbase 0.4.1`. For acoustic model training, we used `SphinxTrain 1.0`, and decoding was performed using the `CMU Sphinx 3.8`, all of which are readily available in Open Source form. We used subsets of 1600 utterances and 600 utterances, respectively, from the DARPA Resource Management (RM1) database for training and testing. A bigram language model was used in all experiments. In all experiments, we used feature vectors of length of 39 including delta and delta-delta features. We assumed that the distance between two microphones is 4 cm.



(a)



(b)



(c)

Fig. 3. *Comparison of recognition accuracy for the DARPA RM database (a) corrupted by an interfering speaker located at 30 degrees, (b) corrupted by an interfering speaker located at 30 degrees under the reverberation of $T_{60} = 200$ ms, and (c) in the presence of natural real-world noise.*

We conducted three different sets of experiments in this section. The first two sets of experiments involve simulated reverberant environments in which the target speaker is masked by a single interfering speaker. The reverberation simulations were accomplished using the *Room Impulse Response* open source software package [36] based on the image method. In these experiments, we assume a room of dimensions 5 m x 4 m x 3 m, with microphones that are located at the center of the room. Both the target and interfering sources are 1.5 m away from the microphone. For the ZCAE algorithm, we used a threshold angle of $\theta_{TH} = 20^\circ$.

In the first set of experiments, we assume that the target is located along the perpendicular bisector of the line between two microphones, which means $\theta_T = 0^\circ$. We assume that the interfering source is located at $\theta_I = 30^\circ$. We repeated the experiments by changing the Signal-to-Interference Ratio (SIR) and reverberation time. As shown in Fig. 3(a), in the absence of reverberation at 0-dB SIR, both the fixed-IMD system (ZCAE) and the automatic-IMD systems (MNCC-MCNV, Auto-ITD-Th) provide comparable performance. If reverberation is present, however, MNCC-MNCV provides substantially better performance than the ZCAE signal separation system. In the second set of the experiments we changed the location of the interfering speaker while maintaining the SIR at 0 dB. As shown in Fig. 4, even if the SIR is the same as in the calibration environment, the performance of the fixed-IMD threshold system (ZCAE) becomes significantly degraded if the actual interfering speaker location is different from the location used in the calibration environment. The MNCC-MNCV selection system provides recognition results that are much more robust with respect to the locations of the interfering sources compared to ZCAE and the basic MNCC-MNCV/Auto-ITD-Th algorithm described in [3].

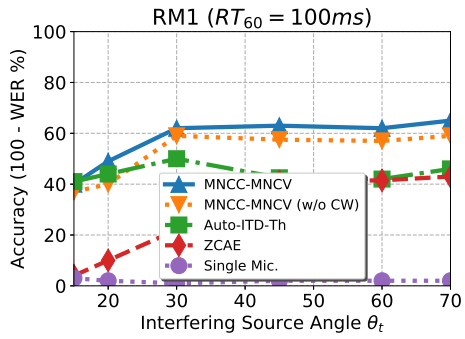


Fig. 4. *Comparison of recognition accuracy for the DARPA RM database corrupted by an interfering speaker at different locations in a simulated room with 100 ms different reverberation time. (SIR) is fixed at 0 dB.*

In the third set of experiments we added noise recorded with two microphones in real environments such as a public market, a food court, a city street and a bus stop. These real noise sources surround the two microphones, and the signals from these recordings are digitally added to clean speech from the test set of the RM database. Fig. 3(c) shows speech recognition accuracy for this configuration. Again we observe that MNCC-MNCV provides the best performance by a significant margin, while the Auto-ITD-Th, MNCC-MCNV without CW, and ZCAE show similar performance to each other.

4. SUMMARY

In this paper, we present a two-microphone algorithm for sound source separation referred to as Minimum Normalized Cross-Correlation and Minimum Normalized Cross-coVariance (MNCC-MNCV). We demonstrate that the use of the MNCC-MNCV method of estimating IMD thresholds provides significantly better recognition accuracy in the presence of interfering speech sources, omni-directional noise, and reverberation than several other approaches including ZCAE [1], PDCW [2], and our previous automatic IMD threshold selection approach [3].

5. REFERENCES

[1] H. Park, and R. M. Stern, "Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero crossings," *Speech Communication*, vol. 51, no. 1, pp. 15–25, Jan. 2009.

[2] C. Kim, K. Kumar, B. Raj, and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain," in *INTERSPEECH-2009*, Sept. 2009, pp. 2495–2498.

[3] C. Kim, K. Eom, J. Lee, and R. M. Stern, "Automatic selection of thresholds for signal separation algorithms based on interaural delay," in *INTERSPEECH-2010*, Sept. 2010, pp. 729–732.

[4] M. Seltzer, D. Yu, and Y.-Q. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Int. Conf. Acoust. Speech, and Signal Processing*, 2013, pp. 7398–7402.

[5] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition tasks," in *Proceedings of the International Conference on Learning Representations*, 2013.

[6] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on CPUs," in *Deep Learning and Unsupervised Feature Learning NIPS Workshop*, 2011.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. 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, Nov. 2012.

[8] T. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Feb. 2017.

[9] ——, "Raw Multichannel Processing Using Deep Neural Networks," in *New Era for Robust Speech Recognition: Exploiting Deep Learning*, S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, Ed. Springer, Oct. 2017.

[10] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K-C Sim, R. Weiss, K. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon, "Acoustic modeling for Google Home," in *INTERSPEECH-2017*, Aug. 2017, pp. 399–403.

[11] C. Kim, A. Misra, K.K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *INTERSPEECH-2017*, Aug. 2017, pp. 379–383.

[12] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, Apr 1987, pp. 705–708.

[13] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *Interspeech 2016*, 2016, pp. 2378–2382. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1386

[14] C. Kim, E. Variani, A. Narayanan, and M. Bacchiani, "Efficient implementation of the room simulator for training deep neural network acoustic models," 2018, pp. 3028–3032. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2566

[15] C. Kim, T. Sainath, A. Narayanan, A. Misra, R. Nongpiur, and M. Bacchiani, "Spectral distortion model for training phase-sensitive deep-neural networks for far-field speech recognition," in *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, April 2018 (submitted).

[16] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Int. Conf. Mach. Learn. (ICML) Workshop Deep Learn. Audio, Speech, Lang. Process.*, 2013.

[17] C. Kim, K. Kumar and R. M. Stern, "Robust speech recognition using small power boosting algorithm," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2009, pp. 243–248.

[18] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2017, pp. 286–290.

[19] T. Higuchi and N. Ito and T. Yoshioka and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2016, pp. 5210–5214.

[20] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, J. Roux, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *INTERSPEECH-2016*, Sept 2016, pp. 1981–1985.

[21] C. Kim and K. K. Chin, "Sound source separation algorithm using phase difference and angle distribution modeling near the target," in *INTERSPEECH-2015*, Sept. 2015, pp. 751–755.

[22] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the mvdr beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, Jan 2010.

[23] P. Chevalier, J. Delmas, and A. Oukaci, "Optimal widely linear mvdr beamforming for noncircular signals," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3573–3576.

[24] P. Chevalier and A. Blin, "Widely linear mvdr beamformers for the reception of an unknown signal corrupted by noncircular interferences," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5323–5336, Nov 2007.

[25] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, "Microphone subset selection for mvdr beamformer based noise reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 550–563, March 2018.

[26] Y. Zhang, W. Li, P. Zhang, and Y. Yan, "Improving multichannel speech recognition with generalized cross correlation inputs and multitask learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5704–5708.

[27] K. J. Palomaki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, pp. 361–378, 2004.

[28] S. Srinivasan and M. Roman and DeL. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Comm.*, vol. 48, pp. 1486–1501, 2006.

[29] C. Kim, A. Menon, M. Bacchiani, and R. M. Stern, "Sound source separation using phase difference and reliable mask selection," in *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, April 2018 (submitted).

[30] W. Grantham, "Spatial hearing and related phenomena," in *Hearing*, B. C. J. Moore, Ed. Academic, 1995, pp. 297–345.

[31] R. M. Stern, D. Wang, and G. J. Brown, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. Wang and G. J. Brown, Eds. Wiley-IEEE Press, 2006.

[32] P. Arabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Tran. Systems, Man, and Cybernetics-Part B:*, vol. 34, no. 4, pp. 1763–1773, Aug. 2004.

[33] D. Halupka, S. A. Rabi, P. Aarabi, and A. Sheikholeslami, "Real-time dual-microphone speech enhancement using field programmable gate arrays," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2005, pp. v/149 – v/152.

[34] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 1315–1329, July 2016.

[35] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2012, pp. 4101–4104.

[36] S. G. McGovern, "a model for room acoustics," http://www.sgm-audio.com/research/rir/rir.html.