SOUND SOURCE SEPARATION USING PHASE DIFFERENCE AND RELIABLE MASK SELECTION SELECTION

Chanwoo Kim[1†]*, Anjali Menon*[3]*, Michiel Bacchiani*[2]*, Richard Stern*[3]

[1]Samsung Research, [2]Google Speech, [3]Carnegie Mellon University

[1]chanw.com@samsung.com, [2]michiel@google.com [3]{anjalim, rms}@cs.cmu.edu

ABSTRACT

In this paper, we present an algorithm called Reliable Mask Selection-Phase Difference Channel Weighting (RMS-PDCW) which selects the target source masked by a noise source using the Angle of Arrival (AoA) information calculated using the phase difference information. The RMS-PDCW algorithm selects masks to apply using the information about the localized sound source and the onset detection of speech. We demonstrate that this algorithm shows relatively 5.3 percent improvement over the baseline acoustic model, which was multistyle-trained using 22 million utterances on the simulated test set consisting of real-world and interfering-speaker noise with reverberation time distribution between 0 *ms* and 900 *ms* and SNR distribution between 0 *dB* up to clean.

Index Terms— Far-field Speech Recognition, Sound source separation, phase difference, Time-frequency bin masking

1. INTRODUCTION

After the advancement of deep learning technology [1, 2, 3, 4, 5, 6], speech recognition accuracy has improved dramatically. Now, speech recognition systems are used not only in portable devices but also in standalone devices for far-field speech recognition. Examples include voice assistant systems such as Amazon Alexa and Google Home [7, 8]. In far-field speech recognition, the impact of noise and reverberation is much larger than near-field cases. Traditional approaches to far-field speech recognition include noise robust feature extraction algorithms [9, 10], on-set enhancement algorithms [11, 12]. Recently, we observed that training using noisy data generated using "room simulator" [7] improves speech recognition accuracy dramatically This system has been successfully employed for training acoustic models for Google Home or Google voice search.

However, as will be seen in Sec. 3, for highly non-stationary noise like interfering speaker noise, this Multistyle TRaining (MTR) or data augmentation approach is not sufficient. In this case, various multi-microphone processing may be employed to further enhance robustness [13, 14, 15, 16, 17]. It has been known that the Inter-microphone Time Delay (ITD) or Phase Difference (PD) between two microphones may be used to identify the Angle of Arrival (AoA) [18, 19]. The Inter-microphone Intensity Difference (IID) may also serve as a cue for determining the AoA [20, 21].

Using the ITD information, we proposed approaches such as Phase Difference Channel Weighting [18] or PAINT. Even though these approaches show good improvement for interfering speakers, it turns out that they may show degradation when the noise type is not highly stationary and reverberation is rather strong. This happens when the estimated mask information is not reliable enough. To tackle this problem, we developed an algorithm referred to as Reliable Masking Selection Phase Difference Channel Weighting (RMS-PDCW). In RMS-PDCW approach, we apply mask only when the mask is estimated more reliably. To test mask reliabiltiy, we use the two criteria: source concentration criterion and the onset criterion.

[†]Work performed while at Google.

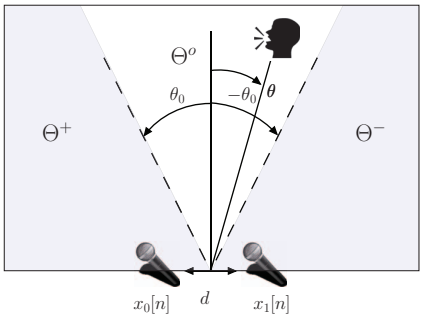


Fig. 1: Two microphones and the target sound source. The space inside a room is divided into three regions depending on the azimuth angle θ: Θ^+, Θ^o, and Θ^-. We use θ_0 of 15^o.

2. THE STRUCTURE OF THE RELIABLE MASK SELECTION PHASE DIFFERENCE CHANNEL WEIGHTING (RMS-PDCW) ALGORITHM

2.1. Review on estimation of the Angle of Arrival (AoA) from phase difference

In this section, we review the procedure for estimating the Angle of Arrival (AoA) of a sound source using two microphone signals [18, 22]. Suppose that we have a pair of microphones and a sound source in Fig. 1. The sound source is located in the direction of the azimuth angle θ.

Let us define the phase difference $\Delta\phi[m, \omega_k]$ for each time-frequency bin $[m, \omega_k]$ [18]:

$$\Delta\phi[m,\omega_k] \triangleq \text{Arg}\left(X_1[m,e^{j\omega_k}]\right) - \text{Arg}\left(X_0[m,e^{j\omega_k}]\right) \mod [-\pi,\pi), \quad 0 \le k \le \frac{K}{2}, \quad (1)$$

where m is the frame index and ω_k is the discrete frequency index defined by $\omega_k = \frac{2\pi k}{K}, 0 \le k \le K/2$ where K is the DFT size. $X_0[m, e^{j\omega_k}]$ and $X_1[m, e^{j\omega_k}]$ are Short-Time Fourier Transform (STFT) of the signals from each microphone. We use a Hamming window of length 100 *ms*.

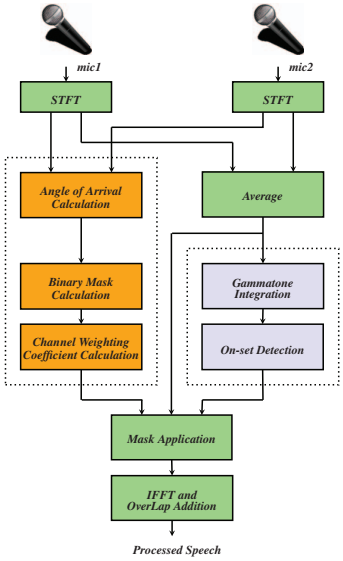


Fig. 2: A block diagram showing the structure of the Reliable Mask Selection - Phase Difference Channel Weighting (RMS-PDCW) algorithm.

From geometric consideration, the AoA $\theta[m,\omega_k]$ is estimated using the following equation [22]:

$$\theta[m,\omega_k] = \arcsin\left(\frac{c_{air}\Delta\phi[m,\omega_k]}{f_s\omega_k d}\right), \quad 0 \le k \le \frac{K}{2}, \tag{2}$$

where f_s is the sampling rate of the signal, and c_{air} is the speed of sound in air, d is the distance between two microphones. In obtaining results in Sec. 3, we use $f_s = 16,000$ *Hz*, $c_{air} = 343$ *m/s*, and $d = 0.04$ *m*.

2.2. Reliable Binary Mask Selection

In this section, we describe the Reliable Binary Mask Selection (RBMS) approach used in the RMS-PDCW algorithm. In the original PDCW [18], we obtain binary mask by examining whether the estimated AoA corresponds to the region Θ^o in Fig. 1.

$$\mu[m,k] = \begin{cases} 1: & \text{if } \left|\theta[m,\omega_k]\right| < \theta_0. \\ 0: & \text{if } \left|\theta[m,\omega_k]\right| > \theta_0. \end{cases} \tag{3}$$

In our experiments in Sec. 3, we use θ_0 value of 15^o. Under reverberation, the estimated AoA $\theta[m,\omega_k]$ in (2) will have an error, which degrades the reliability of the estimated mask $\mu[m,k]$ in (3). When mask estimation is inaccurate, masking may degrade signals rather than enhance them. To test the reliability of masking, we first test whether a target or noisy sound source is likely to be present in a speech frame. We divide the discrete frequency range $0 \le k \le \frac{K}{2}$ into three subsets corresponding to spatial regions Θ^+, Θ^o, and Θ^- in Fig. 1 depending on the Angle of Arrival (AoA) $\theta[m,\omega_k]$ in (2).

$$\mathcal{K}^+[m] = \{k \big| \theta[m,\omega_k] \in \Theta^+ \;\; 0 \le k \le K/2\}, \tag{4a}$$
$$\mathcal{K}^o[m] = \{k \big| \theta[m,\omega_k] \in \Theta^o, \;\; 0 \le k \le K/2\}, \tag{4b}$$
$$\mathcal{K}^-[m] = \{k \big| \theta[m,\omega_k] \in \Theta^-, \;\; 0 \le k \le K/2\}. \tag{4c}$$

In RBMS , we apply binary mask only when a localized source is identified within each of the spatial regions Θ^+, Θ^o, and Θ^- at a specific frame m. For this decision, we calculate the mean and the standard deviation of the estimated AoA $\theta[m,\omega_k]$ for each of these spatial regions. For calculation, we use the magnitude squared spectrum as weighting. For Θ^o, the mean $\mu_\theta^o[m]$ and the standard deviation $\sigma^o[m]$ are calculated using the following equations:

$$\mu_\theta^o[m] = \frac{\sum_{k\in\mathcal{K}^o[m]} p[m,k]\theta[m,\omega_k]}{\sum_{k\in\mathcal{K}^o[m]} p[m,k]}, \tag{5a}$$

$$\sigma_\theta^o[m] = \sqrt{\frac{\sum_{k\in\mathcal{K}^o[m]} p[m,k]\theta[m,\omega_k]^2}{\sum_{k\in\mathcal{K}^o[m]} p[m,k]} - (\mu_\theta^o[m])^2}. \tag{5b}$$

where $p[m,\omega_k]$ is the magnitude squared spectrum defined by:

$$p[m,\omega_k] = \left|\left(X_1[m,\omega_k] + X_2[m,\omega_k]\right)/2\right|^2. \tag{6}$$

$\left(\mu_\theta^+[m], \sigma_\theta^+[m]\right)$ and $\left(\mu_\theta^-[m], \sigma_\theta^-[m]\right)$ are calculated using the same equation just by replacing $\mathcal{K}^o$ with an appropriate subset in (4). For target source presence test, we use the following two criteria:

$$\sigma_\theta^o[m] < \sigma_T, \tag{7a}$$
$$-\theta_0 + \sigma_\theta^o[m] < \mu_\theta[m] < \theta_0 - \sigma_\theta^o[m]. \tag{7b}$$

σ_T is a constant and we use a value of 10^o. This test checks whether the power distribution is sufficiently concentrated in (7a), and checks whether the mean $\mu_\theta[m]$ is separated from the AoA threshold θ_0 by more than the standard deviation $\sigma_\theta^o[m]$ in (7b). If the target source presence test in (2.2) fails, we assume that mask calculation in (3) is unreliable, and use $\mu[m,k] = 1$ for the entire $0 \le k \le K/2$ regardless of the following two noise source presence tests in (8) and (9). To test the noise source presence in Θ^-, we use the following two criteria:

$$\sigma_\theta^-[m] < \sigma_N, \tag{8a}$$
$$\mu_\theta^-[m] < -\theta_0 - \sigma_\theta^-[m]. \tag{8b}$$

The intention of this test for Θ^- is is the same as the target source presence test. σ_N is a constant and we use a value of 20^o. If the test in (8) fails, then $\mu[m,k] = 1$ for $k \in \mathcal{K}^-[m]$, otherwise $\mu[m,k] = 0$ for $k \in \mathcal{K}^-[m]$. The noise source presence test for Θ^+ is performed in the same way as the above noise source presence test for Θ^-:

$$\sigma_\theta^+[m] < \sigma_N, \tag{9a}$$
$$\mu_\theta^+[m] > \theta_0 + \sigma_\theta^+[m]. \tag{9b}$$

If the test in (9) fails, then $\mu[m,k] = 1$ for $k \in \mathcal{K}^+[m]$, otherwise $\mu[m,k] = 0$ for $k \in \mathcal{K}^+[m]$. In sum, by using tests in , (8), and (9), we mask time-frequency bins corresponding to spatial region Θ^+ and Θ^- when both the target and noise sources are likely to exist.

2.3. Reliable Channel Mask Selection (RCMS)

In this section, we describe Reliable Channel Mask Selection (RCMS). Channel masking is accomplished using the Channel Weighting (CW) approach described in [23, 22]. To select more reliable channel masks, we develop a simple onset detection algorithm based on [11]. This is motivated by the fact that the onset portion of speech is less affected by reverberation [24]. In our previous work

[23, 22], we observed that the applying ratio masks for each channel gives better result than applying the binary masks $\mu[m,k]$ in (3) for each DFT index.

Let us first review the Channel Weighting [23]. The filter bank energy of the l-th channel at the frame index m is given by the following equation:

$$P[m,l] = \sum_{k=0}^{K/2} \left| X_a[m, e^{j\omega_k}] H_l[e^{\omega_k}] \right|^2 \quad (10)$$

where $X_a[m, e^{j\omega_k}]$ is the average spectrum given by $X_a[m, e^{j\omega_k}] = \left(X_1[m, e^{j\omega_k}] + X_2[m, e^{j\omega_k}] \right)/2$. After applying the binary mask $\mu[m,k]$ in (3), the filter bank energy for the same l-th channel is given by:

$$P_\mu[m,l] = \sum_{k=0}^{K/2} \mu[m,k] \left| X_a[m, e^{j\omega_k}] H_l[e^{\omega_k}] \right|^2 \quad (11)$$

The channel mask coefficient $w[m,l]$ is the square root of the ratio of $P_\mu[m,l]$ in (11) to $P[m,l]$ in (10):

$$w[m,l] = \sqrt{\frac{P_\mu[m,l]}{P[m,l]}}. \quad (12)$$

Onset detection algorithm we use is motivated by our onset enhancement algorithm in [11]. From the filter bank energy $P[m,l]$ in (10), the low-passed signal is given by:

$$M[m,l] = \lambda M[m-1,l] + (1-\lambda) P[m,l] \quad (13)$$

In our implementation, we use the forgetting factor $\lambda = 0.01$ when the period between successive frames is 50 ms. The onset detection is based on the following decision criterion:

$$\mu[m,k] = \begin{cases} \text{Onset :} & \text{if } P[m,l] > M[m,l], \\ \text{Non-Onset :} & \text{if } P[m,l] \le M[m,l]. \end{cases} \quad (14)$$

For non-onset portion, we do not update the channel mask coefficient:

$$w_{\text{onset}}[m,l] = \begin{cases} \sqrt{\frac{P_\mu[m,l]}{P[m,l]}}. & \text{if } P[m,l] > M[m,l], \\ w_{\text{onset}}[m-1,l] & \text{if } P[m,l] \le M[m,l]. \end{cases} \quad (15)$$

The enhanced spectrum is given by the following equation:

$$Y[m,\omega_k] = \sum_{l=0}^{L-1} w_{\text{onset}}[m,l] X_a[m,\omega_k] H_l[\omega_k] \quad (16)$$

The output time-domain waveform is synthesized using the Inverse Fast Fourier Transform (IFFT) and OverLap Addition (OLA).

2.4. Acoustic model training

Fig. 3 shows the structure of the acoustic model pipeline used for training the speech recognition system in our experiments. The pipeline is based on our work described in [8, 7] with some modification. The "room simulator" generates one-channel simulated utterance by randomly picking up a room configuration. The room configuration distribution , noise sources, SNR, and reverberation time distribution are exactly the same as what we described in [7]. One major difference is instead of generating two-channel simulated waveform, we generate one-channel waveform. After every epoch,

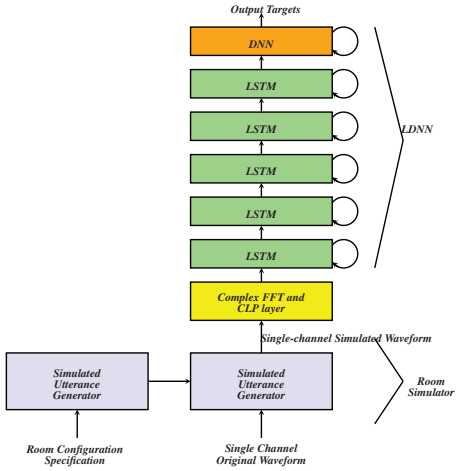


Fig. 3: The architecture for acoustic model training using the room simulator and LSTMs and a DNN (LDNN) [27, 7].

we apply a different room configuration to the utterance so that each utterance may be regenerated in somewhat different configuration. As input, we use the 128 dimension log-mel feature whose window size is 32 *ms*. The interval between successive frame is 10 *ms*. The low and upper cutoff frequencies of the mel filterbank are 125 *Hz* and 7500 *Hz* respectively. Since it has been shown that long-duration features represented by overlapping features are helpful [25], four frames are stacked together and the input is downsampled by a factor of 3. Thus we use a context dependent feature consisting of 512 elements given by 128 (the size of the log-mel feature) x 4 (number of stacked frames). The feature is processed by a typical multi-layer LSTM acoustic model. We use 5-layer LSTMs with 768 units in each layer. The output of the final LSTM layer is passed to a 768 unit DNN, followed by a softmax layer. The softmax layer has 8192 nodes corresponding to the number of tied context-dependent phones in our ASR system. The output state label is delayed by five frames, since it was observed that the information about future frames improves the prediction of the current frame [26]. The acoustic model was trained using the Cross-Entropy (CE) minimization as the objective function after aligning each utterance. The Word Error Rates (WERs) are obtained after 120 million steps of acoustic model training.

3. EXPERIMENTAL RESULTS

In this section, we show experimental results obtained with the RMS-PDCW algorithm. For training, we used an anonymized 22-million English utterances (18,000-hr), which are hand-transcribed. The training set is the same as what we used in [8, 7]. For evaluation, we used around 15-hour of utterances (13,795 utterances) obtained from anonymized voice search data. We also generate noisy evaluation sets from this relatively clean voice search data. The "room simulator" in [7]. was used to generate noisy utterances assuming room configuration shown in Fig. 4. For noisy data, we use two different . types of noise. The first one is the DEMAND [28] noise, which contains various real-world noises from kitchens, rivers, hallways, buses, metro, cars, etc [28]. The noise in the DEMAND noise is relatively stationary. The second noise type we used is interfering speaker utterances which were obtained from the Wall Street Journal (WSJ) si-284 corpus. In the noisy set in Table. 1, we used

Table 1: Word Error Rates (WERs) obtained with multi-microphone approaches with Multistyle TRaining (MTR) using the room simulator [7].

	Clean	Simulated noisy set	Relative improvement over the baseline with MTR (%)
Baseline	11.3 %	51.7 %	-
Baseline with MTR	11.7 %	35.1 %	-
Delay and sum with MTR	11.7 %	34.9 %	0.6 %
PPDCW with MTR	11.8 %	34.4 %	3.2 %
PDCW + RCMS with MTR	11.8 %	33.6 %	4.2 %
PDCW + RBMS with MTR	11.8 %	33.3 %	5.0 %
RMS-PDCW with MTR	11.8 %	**33.2** %	**5.3** %

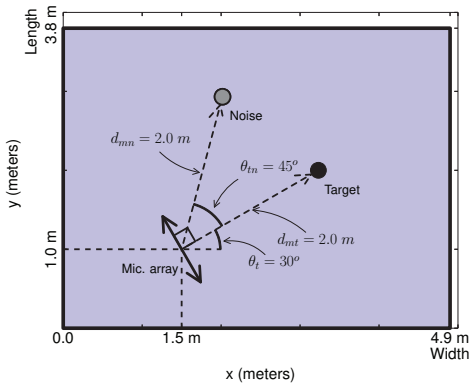
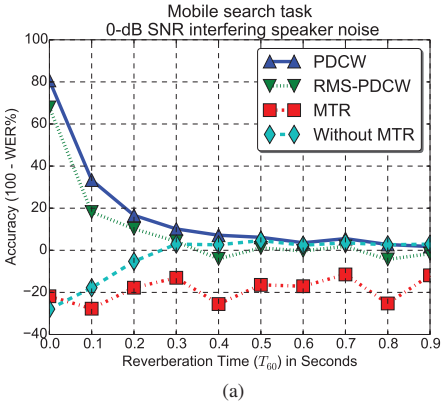
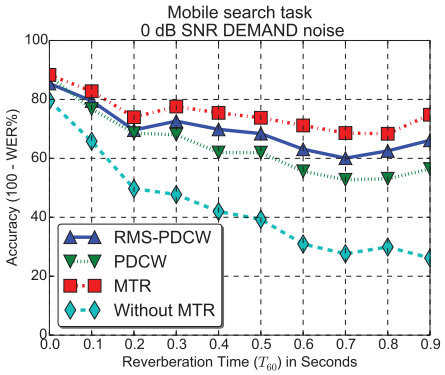


Fig. 4: Room configuration used in the experiment in Sec. 3



(a)



(b)

Fig. 5: Word Error Rates (WERs) for the voice search test set at different reverberation time corrupted by (a) an interfering speaker and (b) various noise in the DEMAND noise database.

50 percent of noise from the DEMAND noise set and 50 percent from WSJ si-284 corpus. For reverberation time, we used a uniform distribution from 0 seconds to 900 *ms*. For the SNR distribution, we used 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, and the clean utterance in equal proportions.

Due to the page limitation, we cannot show results for each specific SNR level, noise type, and reverberation time. The relative improvement is not uniform for different conditions. For example as shown in Fig. 5a, PDCW and RMS-PDCW show very large improvement for interfering speaker noise at relatively small reverberation time. For example at 0 dB SNR and $T_{60} = 0$ *ms*, PDCW and RMS-PDCW show more than 80 % Word Error Rate (WER) reduction.

However, for strong stationary noise under high reverberation time, the MTR is quite effective. In such cases, PDCW and RMS-PDCW may show somewhat worse performance than the baseline MTR as shown in Fig. 5b. As a whole, The baseline system without MTR shows 51.7 % Word Error Rate (WER) as shown in Table 1. The baseline with the MTR using the room simulator in [7] reduces the WER down to 35.1 %. We observe that MTR is more effective for stationary noise rather than highly non-stationary noise such as the interfering speaker noise. The PDCW system shows 34.4 % WER, which is relatively 3.2 % WER reduction over the baseline with MTR. RBMS and RCMS described in 2.2 and 2.3 bring additional improvement over the standard PDCW. RMS-PDCW which includes both the RBMS and RCMS shows relatively 5.3 % WER reduction as shown in Table 1.

4. CONCLUSIONS

In this paper, we described the RMS-PDCW algorithm which selects more reliable masks and applies them to utterances corrupted by noise and reverberation. Our experimental results show that the this algorithm shows relatively 5.3 % WER reduction over the single-channel baseline trained using the room simulator .

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] ——, "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.

[7] 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.

[8] 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.

[9] 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.

[10] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp. 142–152, Feb. 2008.

[11] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *INTERSPEECH-2010*, Sept. 2010, pp. 2058–2061.

[12] C. Kim, K. Chin, M. Bacchiani, and R. M. Stern, "Robust speech recognition using temporal masking and thresholding algorithm," in *INTERSPEECH-2014*, Sept. 2014, pp. 2734–2738.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] R. M. Stern, E. Gouvea, C. Kim, K. Kumar, and H .Park, "Binaural and multiple-microphone signal processing motivated by auditory perception," in *Hands-Free Speech Communication and Microphone Arrays, 2008*, May. 2008, pp. 98–103.

[18] 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.

[19] C. Kim, K. Kumar, and R. M. Stern, "Binaural sound source separation motivated by auditory processing," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May 2011, pp. 5072–5075.

[20] H. S. Colburn and A. Kulkarni, "Models of sound localization," in *Sound Source Localization*, A. N. Popper and R. R. Fay, Eds. Springer-Verlag, 2005, pp. 272–316.

[21] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[22] C. Kim, C. Khawand, and R. M. Stern, "Two-microphone source separation algorithm based on statistical modeling of angle distributions," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2012, pp. 4629–4632.

[23] 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.

[24] P. M. Zurek, *The precedence effect*. New York, NY: Springer-Verlag, 1987, ch. 4, pp. 85–105.

[25] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition," in *INTERSPEECH-2015*, Sept. 2015, pp. 1468–1472.

[26] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, Apr. 2015, pp. 4580–4584.

[27] T. Sainath, R. Weiss, K. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2016, pp. 5075–5079.

[28] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proc. 21st Int. Congr. on Acoust.*, June 2013.