Sound Source Separation Algorithm Using Phase Difference and Angle Distribution Modeling Near the Target

Chanwoo Kim[1]*, Kean K. Chin*[2]

Google, Mountain View CA 94043 USA
{chanwcom,kkchin}@google.com

Abstract

In this paper we present a novel two-microphone sound source separation algorithm, which selects the signal from the target direction while suppressing signals from other directions. In this algorithm, which is referred to as Power Angle Information Near Target (PAINT), we first calculate phase difference for each time-frequency bin. From the phase difference, the angle of a sound source is estimated. For each frame, we represent the source angle distribution near the expected target location as a mixture of a Gaussian and a uniform distributions and obtain binary masks using hypothesis testing. Continuous masks are calculated from the binary masks using the Channel Weighting (CW) technique, and processed speech is synthesized using IFFT and the OverLap-Add (OLA) method. We demonstrate that the algorithm described in this paper shows better speech recognition accuracy compared to conventional approaches and our previous approaches.

Index Terms: Robust speech recognition, signal separation, interaural time difference, binaural hearing, phase difference

1. Introduction

In spite of recent successes especially with close-talking applications like smart phones, speech recognition has been less successful for far-field applications such as car navigation systems and home appliances. The major problem is that under noisy or reverberant environments, speech recognition accuracy is seriously degraded. Thus, researchers have proposed various kinds of algorithms to address this problem [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Nevertheless, for highly non-stationary noise or reverberant speech, improvement has been still limited. For these environments, recent work motivated by human auditory processing seems to be more promising [10, 11, 12, 13, 14, 15]. Some other approaches are based on multi-microphone processing.

The algorithm we are describing in this paper is referred to as Power Angle Information Near Target (PAINT), which is an improvement over our previous Phase Difference Channel Weighting (PDCW) approach [16]. The PDCW algorithm is based on Interaural Time Delay (ITD) estimation for each time-frequency bin to obtain masks. In most ITD-based approaches [16, 17, 18, 19], we usually assume that we have prior knowledge about the expected target source. In the PAINT algorithm, we still need to know expected target location, but unlike our previous PDCW [16], the location does not need to be accurate as long as it is contained in the angle margin.

The most important difference between the PAINT algorithm and our previous Statistical Modeling of Angle Distributions-Channel Weighting (SMAD-CW) [19] is that we model the power-weighted source angle only near the expected target location. In SMAD-CW approach, we tried to model the source angle distribution for all directions, but the problem is that we usually do not have any prior knowledge about noise sources, thus modeling the noise distribution using a single von Mises distribution [20] might not be a good approach in some cases. Von Mises distribution is a simplified model of a wrapped Gaussian distribution, which results from wrapping of the Gaussian distribution around the unit circle. However, it still requires solving equations including Bessel functions for parameter estimation for each frame. Additionally, for certain frames, there might be multiple noise sources, and in this case, a single von Mises distribution is not a good model. On the contrary, in the PAINT algorithm, we build a statistical model only near the target, and assume that the power-weighted angle distribution of the target is represented by a Gaussian distribution. Since there should be only one target, usually it is a reasonable assumption. For noise sources, we do not model the distribution for all directions as done with SMAD-CW. Instead, we just assume that the noise distribution is uniform near the target. If the noise source is not very close to the target, this is a valid assumption.

2. Structure of PAINT processing

Fig. 1 shows the structure of the PAINT algorithm which we introduce in this paper. A Short-Time Fourier Transform (STFT) is performed using Hamming windows of duration 75 ms with 10 ms between frames, using a DFT of 2048. The location of the source angle is estimated from the phase difference of the left and right spectra. The power and source location estimates are used to build a statistical distribution model for each frame to create binary masks. Using these binary masks, channel weighting coefficients are obtained and speech spectrum is enhanced by channel weighting [16, 11, 21] Finally, enhanced signal is obtained by IFFT and OverLap-Add (OLA). Since the window length we are using (75 ms) is significantly longer than the window length for feature extraction (25 ms), we resynthesize speech and apply a conventional feature extraction process.

2.1. Obtaining the sound source location from two microphone signals

In this section, we review the procedure for estimating the angle of the sound source the source using two microphone signals [16, 17, 19]. Let us define the phase difference $\phi[m, \omega_k]$ at the time-frequency bin $[m, \omega_k]$ by the following equation [16]:

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

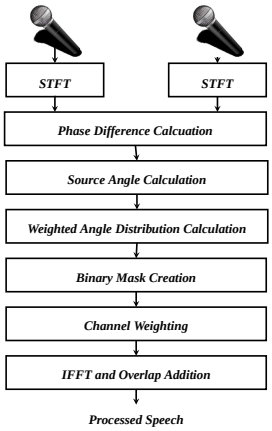


Figure 1: *The structure of PAINT processing*

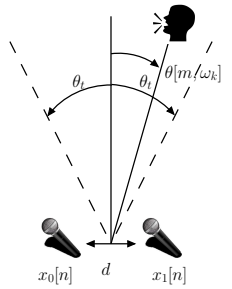


Figure 2: *Two microphones and a sound source.*

where $X_0[m, e^{j\omega_k}]$ and $X_1[m, e^{j\omega_k}]$ are STFT of the signals $x[0]$ and $x[1]$ received from each microphone as shown in Fig. 2. In (1) 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. From geometric consideration, $\theta[m, \omega_k]$ is estimated from $\Delta\phi[m, \omega_k]$ using the following equation [19]:

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

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.

2.2. Estimation of the angle distribution near target

In this section, we discuss how to model the distribution of angle θ near the expected target location for each frame. For notational simplicity, we will drop the frame index m in symbols in this section. In PAINT processing, we construct the statistical model near the location of the expected target as shown below:

$$\Theta_t = \{\theta | -\theta_t \le \theta \le \theta_t\}. \tag{3}$$

For θ_t value, we find $\frac{20}{180}\pi$ (corresponding to 20 degrees) is appropriate. This Θ_t region is depicted in Fig. 2. From the angle values $\theta[\omega_k]$, we define the set of DFT frequency indices K_t which is associated with Θ_t:

$$K_t = \{k | \theta[\omega_k] \in \Theta_t, 0 \le k \le K/2\}. \tag{4}$$

For $\theta[\omega_k], k \in K_t$, we assume that the angle θ distribution of noise and the target is given by the following equations:

$$f_0(\theta) = \frac{1}{2\theta_t}: \quad \text{noise} \tag{5a}$$

$$f_1(\theta) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}\right): \quad \text{target} \tag{5b}$$

In (5b), we use a fixed standard deviation of $\sigma_1 = 2$ degrees based on our observation of real target distributions. Thus, the Gaussian distribution in (5b) has only one parameter to be estimated, which is μ_1. For the m-th frame, the probability density function $f(\theta)$ may be represented as a mixture of two probability density functions $f_0(\theta)$ and $f_1(\theta)$ as follows:

$$f(\theta) = c_0 f_0(\theta) + c_1 f_1(\theta) \tag{6}$$

where c_0 and c_1 are mixture coefficients with $c_0 + c_1 = 1$. We use the symbol $\mathcal{M}$ to represent the parameters for (6) at the frame index m as shown below:

$$\mathcal{M} = \{c_1, \mu_1\}. \tag{7}$$

In the mixture probability density function represented by (5) and (6), we define a random variable $z[k]$ which defines whether the current sample $\theta[\omega_k]$ is from the noise mixture component in (5a) or from the target mixture component (5b). By definition, $z[k] = 0$, if $\theta[\omega_k]$ is originated from the noise mixture component (the 0-th component) and $z[k] = 1$, if $\theta[\omega_k]$ is originated from the target speech component (the 1-st component).

We estimate the parameters in (7) using the Expectation-Maximization (EM) algorithm due to this hidden variable $z[k]$ [22]. We represent the estimated model parameters after t iterations by the superscript $^{(t)}$. We define the following sets of angles, the corresponding power, and the corresponding mixture indices which satisfy the "near target criterion" defined in (3):

$$\Theta_t = \{\theta[\omega_k] | k \in K_t\}, \tag{8a}$$

$$\mathcal{P}_t = \{p[\omega_k] | k \in K_t\}, \tag{8b}$$

$$Z_t = \{z[k] | k \in K_t\}. \tag{8c}$$

Power $p[\omega_k]$ in (8b) is calculated by squaring the average of the left and the right spectra.

Initial parameter estimation: The initial parameter estimation does not need to be accurate. However, we observe that reasonably good estimation of $c_1^{(0)}$ is important for good performance. We divide the near target angle defined by (3) into two portions and assume that the outer portion is primarily dominated by the noise source, which follows the uniform distribution. The set of frequency indices associated with the inner portion $K_{t,i}$ and the outer portion $K_{t,o}$ are defined by the following equations:

$$K_{i,o} = \{k | |\theta[\omega_k]| \le \theta_t/2, \quad 0 \le k \le K/2\}. \tag{9a}$$

$$K_{t,o} = \{k | \theta_t/2 \le |\theta[\omega_k]| \le \theta_t, \quad 0 \le k \le K/2\}. \tag{9b}$$

For the initial parameter $\mathcal{M}^{(0)}$ estimation, we assume that the outer target region $\Theta_{t,o}$ is dominated by the noise distribution which follows the uniform distribution in (3). Under this assumption, $c_0^{(0)}$ is obtained by the following equation:

$$c_0^{(0)} = \min\left(2\frac{\sum_{k\in K_{t,o}} p[\omega_k]}{\sum_{k\in K_t} p[\omega_k])}, 1\right). \tag{10}$$

$c_1^{(0)}$ is obtained by $1 - c_0^{(0)}$. $\mu_1^{(0)}$ is obtained from the weighted average of $\theta[\omega_k]$ belonging to the inner target portion:

$$\mu_1^{(0)} = \frac{\sum_{k \in K_{i,o}} p[\omega_k]\theta[\omega_k]}{\sum_{k \in K_{i,o}} p[\omega_k]} \tag{11}$$

Assuming that the target speaker does not move rapidly with respect to the microphone, we apply the following smoothing in each parameter update step to the estimate mean to improve performance:

$$\hat{\mu}_1^{(0)} = \lambda \mu_{1,prev} + (1 - \lambda)\mu_1^{(0)} \tag{12}$$

where $\mu_{1,prev}$ is the estimated μ_1 value in the previous frame, λ is a forgetting factor equal to 0.95. In the next parameter update step, we use $\hat{\mu}_1^{(0)}$ instead of the original $\mu_1^{(0)}$.

Parameter update: Given the model parameter at the t-th iteration $\mathcal{M}^{(t)} = \left\{c_1^{(t)}, \mu^{(t)}\right\}$ and the observed data sets Θ_t in (8), the E-step of the EM algorithm is given as follows:

$$Q(\mathcal{M}, \mathcal{M}^{(t)}) = E_{Z_t|\Theta_t,\mathcal{M}^{(t)}}\left[\ln\left(f(\Theta_t, Z_t|\mathcal{M})\right]\right. \tag{13}$$

where Θ_t and Z_t are the set of estimated angles and the set of mixture indices defined in (8), and Z_t is the set of mixture indices. The equation (13) is expressed in terms of $\theta[\omega_k]$ and $p[\omega_k]$ as follows:

$$\begin{aligned} &Q(\mathcal{M}, \mathcal{M}^{(t)}) \\ &= E_{Z_t|\Theta_t,\mathcal{M}^{(t)}}\left[\sum_{k \in K_t} p[\omega_k] \ln\left(f\left(\theta[\omega_k], z_i|\mathcal{M}\right)\right)\right]. \end{aligned} \tag{14}$$

We define the conditional probability of $z[k]$ by $\gamma_{j,k}, j = 0, 1$ as follows:

$$\gamma_{j,k}^{(t)} = P\left(z[k] = j|\theta[\omega_k], \mathcal{M}^{(t)}\right), \tag{15}$$

which may be expressed as follows, using the Bayes theorem:

$$\gamma_{j,k}^{(t)} = \frac{c_j^{(t)} f_j^{(t)}(\theta[\omega_k])}{c_0^{(t)} f_j^{(t)}(\theta[\omega_k]) + c_1^{(t)} f_j({}^{(t)}\theta[\omega_k])}, \quad j = 0, 1. \tag{16}$$

Using (15), (13) may be written as:

$$\begin{aligned} Q(\mathcal{M}, \mathcal{M}^{(t)}) = \sum_{k \in K_t} \Bigg[&\gamma_{0,k}^{(t)} p[\omega_k] \ln\left((1 - c_1^{(t)})\frac{1}{2\theta_t}\right) \\ &+ \gamma_{1,k}^{(t)} p[\omega_k] \Bigg(\ln(c_1^{(t)}) - \frac{1}{2}\ln\left(\sigma_1^2\right) \\ &- \frac{\left(\theta[\omega_k] - \mu_1^{(t)}\right)^2}{2\sigma_1^2} - \frac{1}{2}\ln(2\pi)\Bigg)\Bigg] \end{aligned} \tag{17}$$

Differentiating (13) with respect to $c_1^{(t)}$, and $\mu_1^{(t)}$ and using the conditional probability in (15), it can be shown that the following update equations maximize (13):

$$c_1{}^{(t+1)} = \frac{\sum_{k \in K_t} \gamma_{1,k}^{(t)} p[\omega_k]}{\sum_{k \in K_t} p[\omega_k]}, \tag{18}$$

$$\mu_1^{(t+1)} = \frac{\sum_{kinK_t} \gamma_{1,k}^{(t)} p[\omega_k]\theta[\omega_k]}{\sum_{k \in K_t} \gamma_{1,k}^{(t)} p[\omega_k]} \tag{19}$$

As in the case of the initial parameter update, we apply a smoothing to $\mu_1^{(t+1)}$ in the form of (12) after each parameter update stage.

2.3. Binary mask creation

We set up two hypotheses for each time-frequency bin $[m, k]$:

$$\begin{cases} H_0[m,k] & : \text{The signal is from a noise source or silence.} \\ H_1[m,k] & : \text{The signal is from a target source.} \end{cases} \tag{20}$$

The hypothesis testing is performed for each time-frequency point $\mathcal{P}[m, k]$ using the probability density functions obtained in Sec. 2.2. The testing is based on Maximum A Posteriori (MAP) criterion as shown below:

$$c_0[m] f_0(\theta) \underset{H_1}{\overset{H_0}{\gtrless}} c_1[m] f_1(\theta) \tag{21}$$

Note that $f_0(\theta)$ and $f_1(\theta)$ in (21) are probability density functions in (5) and (6). The binary mask $\mu[m, k]$ is obtained by the following equation:

$$\mu[m,k] = \begin{cases} 0: & \text{if } H_0[m,k] \text{ is chosen.} \\ 1: & \text{if } H_1[m,k] \text{ is chosen.} \end{cases} \tag{22}$$

2.4. Channel weighting coefficients from binary masks

Channel Weighting (CW) is a technique of applying a continuous masking (or weighting) coefficient rather than directly applying a binary mask, which has shown to be helpful for better speech recognition accuracy [16, 19]. For the l-th filterbank channel, the original power is given by:

$$P_i[m,l] = \sum_{k=0}^{K/2} \left| X_a[m, e^{j\omega_k}] H_l[e^{\omega_k}] \right|^2 \tag{23}$$

where $X_a[m, e^{j\omega_k}]$ is the average of the spectra from the left and the right microphones, and $H_l[e^{\omega_k}]$ is the frequency response of the l-th channel. In our implementation, we use zero-phase rectangular shaped frequency responses, whose center frequencies are uniformly spaced along the Equivalent Rectangular Bandwidth (ERB) scale. Using the binary mask $\mu[m, k]$ in (22), the power for the same time-frequency bin after mask application is given by:

$$P_o[m,l] = \sum_{k=0}^{K/2} \mu[m,k] \left| X_a[m, e^{j\omega_k}] H_l[e^{j\omega_k}] \right|^2. \tag{24}$$

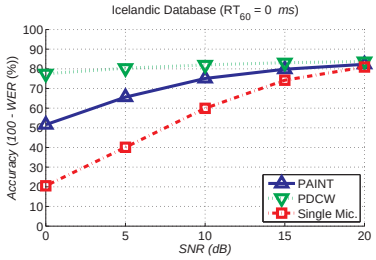
From (23) and (24), the power ratio is given by the following equation:

$$w[m,l] = \frac{\sum_{k=0}^{K/2} \mu[m,k] \left| X_a\left[m, e^{j\omega_k}\right] H_l\left[e^{j\omega_k}\right] \right|^2}{\sum_{k=0}^{K/2} \left| X_a\left[m, e^{j\omega_k}\right] H_l\left[e^{j\omega_k}\right] \right|^2}. \tag{25}$$

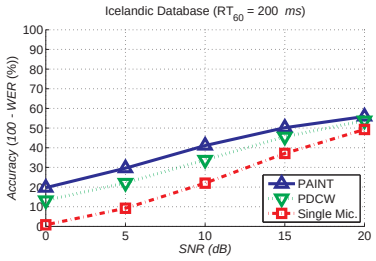
We refer this power ratio $w[m, l]$ in (25) to channel weighting coefficient [16]. Using the channel weighting coefficients $w[m, l]$ in (25), we obtain the enhanced spectrum $Y[m, l]$ using the channel weighting technique [9, 11]:

$$Y[m, e^{j\omega_k}] = \sum_{l=0}^{L-1} \left(\sqrt{w[m,l]} X[m, e^{j\omega_k}] H_l[e^{j\omega_k}] \right), \tag{26}$$

where L is the number of filter channels. After obtaining the enhanced spectrum $Y[m, l]$, the output speech is synthesized using the IFFT and the OverLap-Add (OLA).



(a)



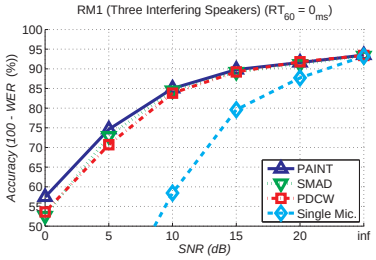
(b)

Figure 3: *Comparison of speech recognition accuracy using PAINT, PDCW, and the baseline for anechoic environment (Fig. 3a) and for reverberant environment (Fig. 3b) with* $T_{60} = 200ms$.

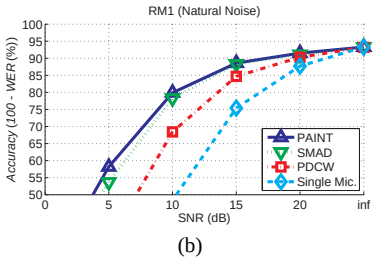
3. Experimental Results

In this section we describe experimental results in two different configurations. In the first set of experiments, we use anonymized Icelandic speech database. consisting of 92,851 training utterances for training and 9,792 evaluation utterances. For features, forty filter bank coefficients from twenty previous frames, the current frame, and five future frames are concatenated to create a feature vector. For acoustic model training and evaluation, we use a Hidden Markov Model(HMM) /Deep Neural Network(DNN) hybrid system. Reverberation simulations with this Icelandic database were accomplished using a `Room Simulator`, which is based on the image method [23]. The room size is assumed to be 5.0 x 4.0 x 3.0 meters, and the microphone array consisting of two microphones is placed at the center of the room. The distance between two microphones is 4 centimeters.

We compare our PAINT algorithm with our previous algorithm , Phase Difference Channel Weighting (PDCW) and the baseline single-microphone system. For PDCW processing, instead of using the original system in [16], we use this PAINT system with a fixed angle threshold of 20 degrees without performing hypothesis testing. The experimental results are shown in Fig. 4. For anechoic environment with one interfering speaker, as shown in Fig. 3a, the PDCW algorithm shows a remarkable result followed by this PAINT algorithm. However, we would like to emphasize that one interfering noise source at a fixed position without reverberation is a very unrealistic environment. In presence of reverberation with T_{60} of 200 ms, as can be seen in Fig. 3b, the PAINT algorithm shows a significantly better result than PDCW for all the SNR ranges. In the second set of experiments, we used subsets of 1600 utterances and 600 utterances, respectively, from the DARPA Resource Management (RM1) database for training and evaluation. In this experiment, we compare our PAINT algorithm with the SMAD algorithm in [19], the PDCW algorithm [16], and the baseline single microphone system. Fig. 4a shows experimental results when there are three interfering speakers inside a room.



(a)



(b)

Figure 4: *Comparison of speech recognition accuracy using PAINT, SMAD-CW, PDCW, and the baseline in the presence of three interfering speakers randomly located inside a room (Fig. 4a) and in the presence of natural real-world noise. (Fig. 4b)*

The geomertic configuration of the room and the microphone array is the same as the configuration of the first experiment. The location of three interfering speakers is random inside this room. Thus, it is possible that some interfering speakers might located in a similar direction to the target speaker. As shown in Fig. 4a, the PAINT algorithm shows the best results followed by the SMAD and the PDCW algorithms. In the last experiment in Fig. 4b, we added noise recorded in real environments with real two-microphone hardware in locations such as a public market, a food court, a city street and a bus stop with background speech. In this experiment using natural noise, again the PAINT algorithm provides better performance than SMAD and PDCW.

4. Conclusion

In this paper, we presented a source separation algorithm, PAINT, based on source location angles calculated from phase difference. This algorithm is improvement over our previous PDCW algorithm. The sound source angle distribution near the expected target location is obtained for each frame assuming a mixture of a Gaussian and a uniform distribution. Statistical hypothesis testing is performed to make binary masks for each time-frequency bin, and ratio masks are obtained using the Channel Weighting technique. As shown in experimental results, as noise conditions become more realistic, the PAINT algorithm provides better performance over other algorithms such as PDCW and SMAD. The Matlab version of PAINT with sample audios is available at `http://www.cs.cmu.edu/~robust/archive/algorithms/PAINT_INTERSPEECH2015`.

5. Acknowledgements

This research was supported by Google. The authors are grateful to Ananya Misra and Prof. Richard Stern for helpful discussion.

6. References

[1] P. Pujol, D. Macho, and C. Nadeu, "On real-time mean-and-variance normalization of speech recognition features," in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1, May 2006, pp. 773–776.

[2] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Nov. 2001, pp. 21–24.

[3] H. Mirsa, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral entropy based feature for robust ASR," in *IEEE Int. Conf. Acoust. Speech, and Signal Processing*, May 2004, pp. 193–196.

[4] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Communication*, vol. 53, no. 5, pp. 736–752, May-June 2011.

[5] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *IEEE Int. Conf. Acoust., Speech and Signal Processing*, May. 1996, pp. 733–736.

[6] S. Ganapathy, S. Thomas, and H. Hermansky, "Robust spectro-temporal features based on autoregressive models of hilbert envelopes," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4286–4289.

[7] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Apr. 1997, pp. 33–42.

[8] C. Kim and R. M. Stern, "Power function-based power distribution normalization algorithm for robust speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2009, pp. 188–193.

[9] 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.

[10] 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.

[11] ——, "Nonlinear enhancement of onset for robust speech recognition," in *INTERSPEECH-2010*, Sept. 2010, pp. 2058–2061.

[12] ——, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4574–4577.

[13] ——, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH-2009*, Sept. 2009, pp. 28–31.

[14] C. Kim, Y.-H. Chiu, and R. M. Stern, "Physiologically-motivated synchrony-based processing for robust automatic speech recognition," in *INTERSPEECH-2006*, Sept. 2006, pp. 1975–1978.

[15] C. Kim and K. Seo, "Robust DTW-based recognition algorithm for hand-held consumer devices," *IEEE Trans. Consumer Electronics*, vol. 51, no. 2, pp. 699–709, May 2005.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Geoffrey W. Hill, "Algorithm 518: Incomplete Bessel Function I0. The Von Mises Distribution [S14]," *ACM Transactions on Mathematical Software*, vol. 3, no. 3, pp. 279–284, Sept. 1977.

[21] C. Kim, K. 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.

[22] M. D. Srinath, P. K. Rajasekaran, and R. Viswanathan, *Introduction to Statistical Signal Processing with Applications*. Upper Saddle River, NJ: Prentice-Hall, Inc. Upper Saddle River, NJ, 1996.

[23] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.