Robust speech recognition using temporal masking and thresholding algorithm

Chanwoo Kim[1]*, Kean K. Chin*[1]*, Michiel Bacchiani*[1]*, Richard M. Stern*[2]

[1] Google, Mountain View CA 94043 USA
[2] Carnegie Mellon University, Pittsburgh PA 15213 USA
{chanwcom, kkchin, michiel}@google.com, rms@cs.cmu.edu

Abstract

In this paper, we present a new dereverberation algorithm called Temporal Masking and Thresholding (TMT) to enhance the temporal spectra of spectral features for robust speech recognition in reverberant environments. This algorithm is motivated by the precedence effect and temporal masking of human auditory perception. This work is an improvement of our previous dereverberation work called Suppression of Slowly-varying components and the falling edge of the power envelope (SSF). The TMT algorithm uses a different mathematical model to characterize temporal masking and thresholding compared to the model that had been used to characterize the SSF algorithm. Specifically, the nonlinear highpass filtering used in the SSF algorithm has been replaced by a masking mechanism based on a combination of peak detection and dynamic thresholding. Speech recognition results show that the TMT algorithm provides superior recognition accuracy compared to other algorithms such as LTLSS, VTS, or SSF in reverberant environments.

Index Terms: Robust speech recognition, speech enhancement, reverberation, temporal masking, precedence effect

1. Introduction

In recent years, advances in machine learning techniques such as Deep Neural Network (DNN) [1], which exploits enhanced computational power [2] have greatly improved the performance of speech recognition systems, especially in clean environments. Nevertheless, the performance under noisy environments still needs to be significantly improved to be useful for far-field speech recognition applications.

Thus far, many researchers have proposed various kinds of algorithms to address this problem [3, 4, 5, 6, 7, 8]. To some degree, these efforts have been successful for the case of near-field additive noise, however, for far-field reverberant speech, the same algorithms usually have not shown the same amount of improvement. For such environments, we have frequently observed that algorithms motivated by auditory processing [9, 10, 11] and/or multi microphones [12, 13, 14] are more promising than traditional approaches.

Many hearing researchers believe that human perception in reverberation is facilitated by the "precedence effect" [15], which refers to an emphasis that appears to be given to the first-arriving wave-front of a complex signal in sound localization and possibly speech perception. To detect the first wave-front, we can either measure the envelope of the signal or the energy in the frame [16, 17, 18].

Motivated by this, we introduced in previous work an algorithm called Suppression of Slowly varying-components and the Falling edge of the power envelope (SSF) to enhance speech recognition accuracy under reverberant environments [19]. This algorithm has been especially successful for reverberation, but the processing introduces distortion in the resynthesized speech. The nonlinear high-pass filtering in [19] is an effective model to detect the first-arriving wavefront, but it might not be very close to how actual human beings perceive sound.

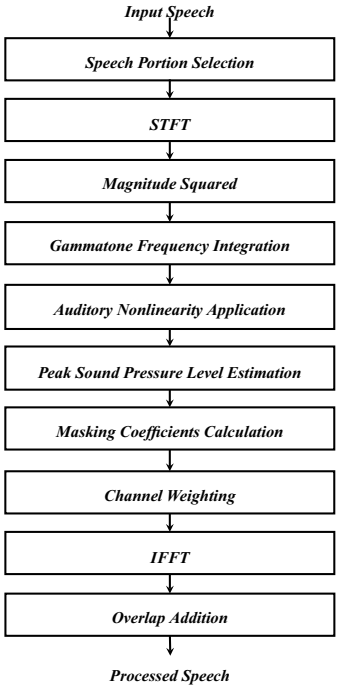

Figure 1: The structure of the TMT algorithm to obtain the normalized speech from the original input speech.

In this paper, we introduce a new algorithm named Temporal Masking and Thresholding (TMT). In this algorithm, temporal masks are constructed to suppress reflected wave files under reverberant environments. We estimate the perceived peak sound level after applying a power-law nonlinearity, and apply a temporal masking based on this. We also apply thresholding based on the peak power.

2. Structure of TMT processing

Figure 1 shows the entire structure of TMT processing. While in the discussion below, we assume that the sampling rate of the speech signal is 16 kHz, this algorithm may be applied for other sampling rates as well. We observe that with the TMT processing presented in this paper, it is better to not apply the algorithm to the silence portion. For this reason, it is better to

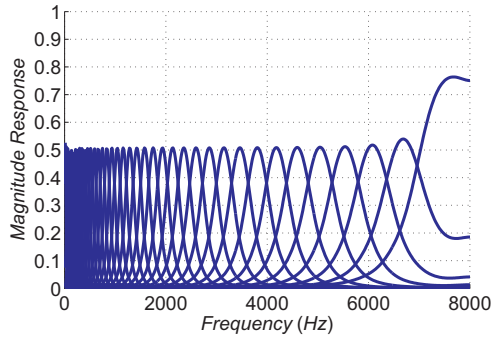

Figure 2: Frequency responses of a gammatone filberbank which is normalized using using (4).

apply a Voice Activity Detector (VAD) before processing and applying the TMT processing only for the speech portions of the waveform.

Speech is segmented into 50-ms frames with 10-ms intervals between adjacent frames. The use of this medium-duration window is motivated by our previous research [20, 21]. A Hamming window is applied for each frame, and a short-time Fourier transform (STFT) is performed. Spectral power in 40 analysis bands is obtained. Temporal masking and thresholding is performed in each channel, and the speech spectrum is reshaped based on these processing. Finally, the output speech is resynthesized using the IFFT and the OverLap Addition (OLA) method. The following subsections describe each stage in more detail.

2.1. Gammatone frequency integration and auditory nonlinearity

As shown in Fig. 1, the first step of TMT processing is performing a short-time Fourier transform (STFT) using Hamming windows of duration 50 ms. We use this medium-duration window which is longer than those used in ordinary speech processing, since it has been frequently observed that medium-duration windows are more appropriate for noise suppression [20, 21]. As in [22], the gammatone spectral integration is performed by the following equation:

$$P[m,l] = \sum_{k=0}^{K/2} \left| X[m,e^{j\omega_k}] H_l[e^{j\omega_k}] \right|^2 \tag{1}$$

where K is the DFT size, m and l represent the frame and channel indices respectively. ω_k is the discrete-time frequency defined by $\omega_k = \frac{2\pi k}{K}$, and $H_l(e^{j\omega_k})$ is the gammatone response for the l^{th} channel. $P[m,l]$ is the power obtained for the time-frequency bin $[m,l]$. When processing signals in the frequency domain, we only consider the lower half of the spectrum ($0 \le k \le \frac{K}{2}$, since the Fourier Transform of real signals satisfies the complex conjugate property:

$$X[m,e^{j\omega_k}] = X^*[m,e^{j\omega_{K-k}}]. \tag{2}$$

The gammatone responses $H_l[e^{j\omega_k}]$ are slightly different from those used in our previous research in [22, 6]. The frequency responses are modified to satisfy the following constraint:

$$\sum_{l=0}^{L-1} H[e^{j\omega_k}] = 1, \qquad 0 \le k \le \frac{K}{2}. \tag{3}$$

where L is the number of the gammatone channels. The reason for this constraint will be explained in Sec. 2.3. Even though frequency responses $Q_l[e^{j\omega_k}]$ of an ordinary filter bank usually do not satisfy (3), we may normalize the filter responses to make them satisfy (3) as follows:

$$H_l[e^{j\omega_k}] = \frac{\left|Q_l[e^{j\omega_k}]\right|}{\sum_{l=0}^{L-1}\left|Q_l[e^{j\omega_k}]\right|}, \qquad 0 \le k \le \frac{K}{2}. \tag{4}$$

For $Q_l[e^{j\omega_k}]$, we use the implementation described in [23]. Fig. 2 shows the magnitude response obtained using (4).

Since the power $P[m,l]$ in (1) is not directly related to how human beings perceive the sound level, we apply an auditory nonlinearity based on the power function [22, 24, 13].

$$S[m,l] = P[m,l]^{a_0} \tag{5}$$

We use a value of $a_0 = \frac{1}{15}$ for the power coefficient, as in [22, 13, 10].

2.2. Peak sound level estimation and binary mask generation

From $S[m,l]$, we obtain the peak sound level for each channel l. The peak sound level is the upper envelope of the $S[m,l]$ as shown in Fig. 3. We use the following simple mathematical model.

$$T[m,l] = \max(\lambda T[m-1,l], S[m,l]) \tag{6}$$

For the time constant λ in (6), we use the value of $\lambda = 0.99$. Using the peak sound level $T[m,l]$, the binary mask $\mu[m,l]$ is constructed using the following criterion:

$$\mu[m,l] = \begin{cases} 1, & \text{if } S[m,l] \ge T[m,l] \\ 0, & \text{if } S[m,l] < T[m,l]. \end{cases} \tag{7}$$

One issue with the procedure described in (6) and (7) is that the peak sound level detection method in (6) does not consider the absolute intensity of the peak of $T[m,l]$. If $T[m,l]$ itself is too small for human listeners to perceive, then this onset should not mask the falling portion following this onset. Thus, we should not apply the TMT technique for silence portion of the utterance. One easy way to achieve this objective is to apply a VAD to remove silence portions of the input utterance before performing the TMT processing. Fig. 5 shows the speech recognition with VAD and without VAD using the TMT processing on the Wall Street Journal 0 (WSJ0) 5k test set. The experimental configuration is described in Sec. 3. As shown in this Fig. 5, to obtain better speech recognition accuracy, we need to apply the TMT processing only to the speech portions of the waveform. For VAD processing, we used a very simple approach based on the threshold of frame energy and smoothing using a state machine.

In our previous SSF algorithm [19], we used a first-order IIR lowpass filter output for a similar purpose, but in this work we use a model more closely related to human perception. In binary masking, it has been frequently observed that a suitable flooring is necessary [20, 12]. In many masking approaches, fixed multiplier values like 0.01, or 0.001 have been frequently used for masked time-frequency bins to prevent them from having zero power [20]. In the TMT algorithm, instead of using such scaling constants, we use a threshold power level $\rho[m,l]$ motivated by auditory masking level, which depends on the peak sound level $T[m,l]$ for each time-frequency bin:

$$\rho[m,l] = \rho_0 T[m,l]^{\frac{1}{a_0}} \tag{8}$$

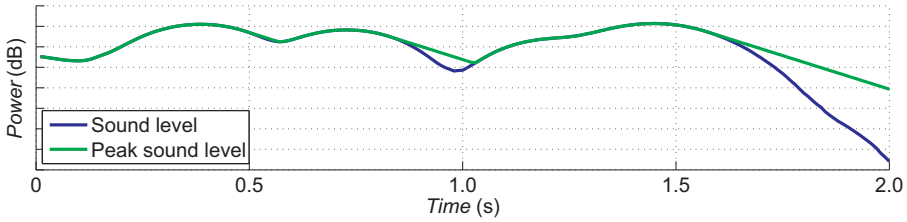

Figure 3: Comparisons of sound level $S[m, l]$ in (5) and peak sound level $T[m, l]$ in (6)

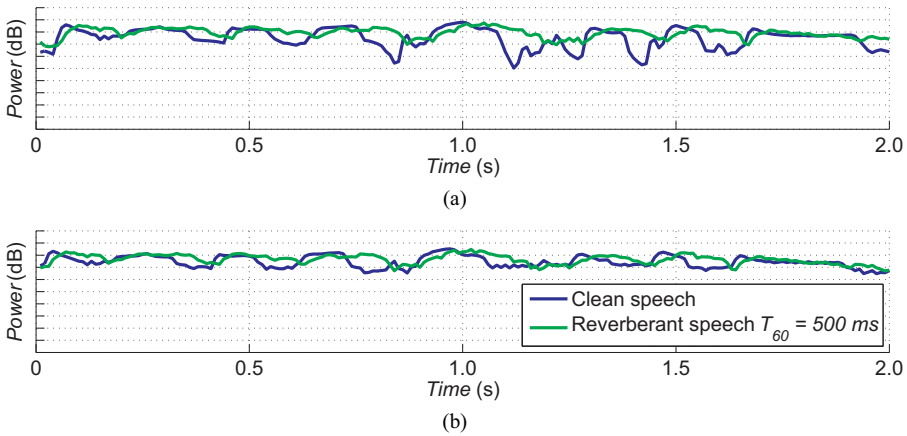

Figure 4: Comparisons of power contours: (4a) power contour $P[m, l]$ of unprocessed speech for clean and reverberant speech ($T_{60} = 500$ *ms*). (4b) power contour of TMT-processed speech for clean and reverberant speech ($T_{60} = 500$ *ms*). For processed speech, we obtained the power contour from $Y[m, e^{j\omega_k}]$.

where a_0 is the power coefficient for the compressive nonlinearty in (5). Since the compressive nonlineary is expanded in (8), it is evident that the threshold power level $\rho[m, l]$ is 20 dB below the time-varying peak power. This thresholding scheme is also motivated by the human auditory masking effect. We believe this thresholding approach is closer to the actual human perception rather than just using some fixed constants like 0.01.

The final masking coefficients $\mu_f[m, l]$ are obtained using the threshold level $\rho[m, l]$ as follows:

$$\mu_f[m, l] = \max\left(\mu[m, l], \frac{\rho[m, l]}{P[m, l]}\right). \tag{9}$$

where $P[m, l]$ is the power in the time-frequency bin $[m, l]$ in (1).

2.3. Channel Weighting

Using the masking coefficients $\mu_f[m, l]$ obtained in (9), we obtain the enhanced spectrum $Y[m, e^{j\omega_k}]$ using the channel weighting technique [6, 19].

$$Y[m, e^{j\omega_k}] = \sum_{l=0}^{L-1}\left(\sqrt{\mu_f[m, l]}X[m, e^{j\omega_k}]H_l[e^{j\omega_k}]\right), \quad 0 \le k \le \frac{K}{2} \tag{10}$$

We obtained the square root of the floored masking coefficient $\mu_f[m, l]$ in the above equation, because, the masking coefficients in Sec. 2.2 is defined for power. For higher frequency components, $\frac{K}{2} \le k \le K - 1$, the spectrum is obtained by the symmetric property of real signals (2).

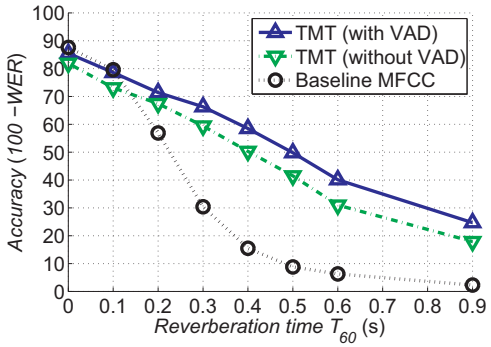

Figure 5: Comparison of speech recognition accuracy with and without the use of a VAD for excluding non-speech portions. The experiment was conducted using the Wall Street Journal (WSJ) SI-84 training and the WSJ0 5k test set.

Now, we are ready to discuss why the constraint of unity in (3) must be upheld for the frequency responses. In (10), if $\mu_f[m, l] = 1$ for all $0 \le l \le L - 1$ at a certain frame m, then we expect the output $Y[m, e^{j\omega_k}]$ to be the same as the input $X[m, e^{j\omega_k}]$. From this, it is obvious that the filter bank needs to satisfy the constraint (3). As before, m and l are the

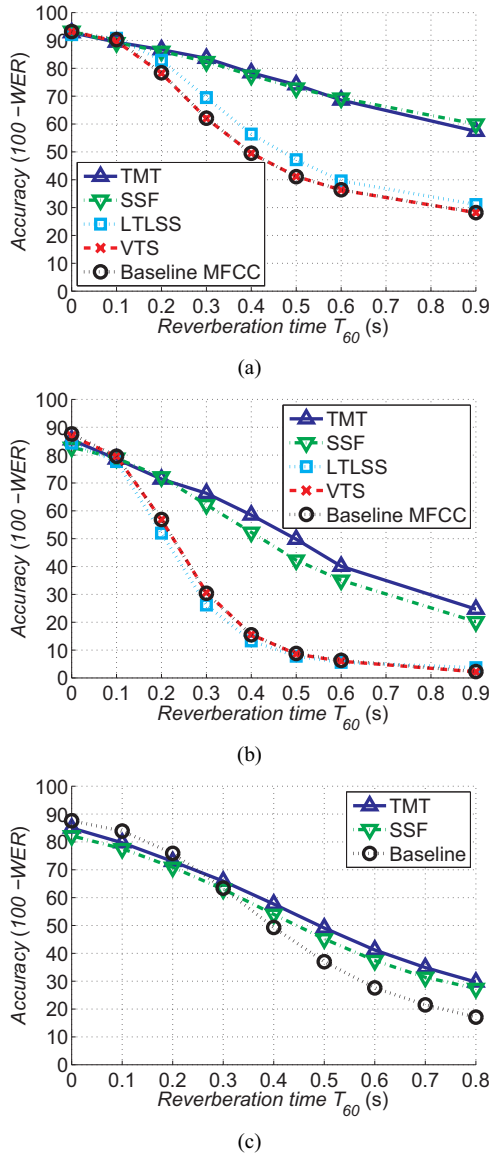

Figure 6: Comparison of speech recognition accuracy using TMT, SSF (Type-II), VTS, and the baseline MFCC: for (Fig. 6a) the Resource Management 1 (RM1) database and (Fig. 6b) the Wall Street Journal (WSJ) SI-84 training and the 5k test set. Fig. 6c shows speech recognition accuracy obtained from the Google Icelandic database using TMT, SSF (Type-II), and baseline processing.

frame and channel indices, and L is the number of channels. After obtaining the enhanced spectrum $Y[m, e^{j\omega_k}]$, the output speech is resynthesized using the IFFT and the overlap-addition (OLA) method. As shown in Fig. 4, after TMT processing, the distortion between two contours obtained from clean and reverberant speech becomes significantly smaller.

3. Experimental Results

In this section we describe experimental results obtained using the DARPA Resource Management 1 (RM1) database, Wall Street Journal 0 (WSJ0) database, and the Google proprietary Icelandic speech database. For the RM1 experiment, we used 1,600 utterances for training and 600 utterances for evaluation. For the WSJ0 experiment, We used 7,138 utterances for training (WSJ SI-84), and used 330 utterances from the WSJ0 5k test set for evaluation. For the Google Icelandic speech recognition experiment, we used 92,851 utterances for training and 9,792 utterances for evaluation.

For the RM1 and WSJ0 experiments, we used `sphinx_fe` included in `sphinx_base 0.4.1` to obtain the MFCC feature. `SphinxTrain 1.0` and `Sphinx 3.8` [25] were used for acoustic model training and decoding for these RM1 and WSJ0 experiments. For the Google Icelandic experiments, the filter coefficients from 20 previous frames, the current frame, and 5 future frames are concatenated to obtain the feature vector. For acoustic modeling and decoding for the Google Icelandic database, we used the proprietary `DistBelief` and `GRECO3`. The Sphinx speech recognition system is based on Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM), and the Google speech recognition system is using the HMM/DNN.

Reverberation simulations in RM1 and WSJ0 were accomplished using the *Room Impulse Response* algorithm [26] based on the image method [27]. We assume a room dimension of 5 x 4 x 3 meters, a distance between the microphone and the speaker of 1.5 meters, with microphone locations at the center of the room. Reverberation simulations with the Google Icelandic database were accomplished using the Google proprietary `Room Simulator`, which is also based on the image method. The room size is assumed to be 4.8 x 4.3 x 2.9 meters, and the microphone is located at the $(2.04, 1.46, 1.0)$-meter position with respect to one corner of the room with the distance from the speaker being 1.5 meters.

We compare our TMT algorithm with our previous algorithm SSF, Vector Taylor Series (VTS) [28] and baseline MFCC processing. The experimental results are shown in Fig. 6a and Fig. 6b. As shown in these two figures, the TMT algorithm has shown consistent performance improvement over SSF. For the smaller RM1 database, the performance difference between TMT and SSF is very small, but as the database size increases in Fig. 6b and Fig. 6c, the performance difference between TMT and SSF becomes larger. VTS provides almost the same results as baseline processing, and LTLSS provides slightly better performance than the baseline for the RM1 database, but slightly worse performance than the baseline for the WSJ0 database. Both LTLSS and VTS produce significantly worse performance than the TMT processing described in this paper. For both SSF and TMT processing, we trained the acoustic models using the same type of processing used in testing. Without such retraining, performance is significantly worse than what is shown in these figures.

4. Conclusions

In this paper, we describe a new dereverberation algorithm, TMT, that is based on temporal enhancement by estimating the peak sound level and applying the temporal masking. We have observed that even though the TMT algorithm is quite simple, it provides better speech recognition accuracy than existing algorithms such as LTLSS or VTS. MATLAB code for the TMT algorithm may be found at `http://www.cs.cmu.edu/~robust/archive/algorithms/tmt`.

5. Acknowledgements

This research was supported by Google.

6. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[2] 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.

[3] 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.

[4] R. Drullman, J. M. Festen and R. Plomp, "Effect of reducing slow temporal modulations on speech recognition," *J. Acoust. Soc. Am.*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[5] C. K. K. Kumar and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May 2011, pp. 4784–4787.

[6] 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.

[7] 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.

[8] C. Kim, K. Seo, and W. Sung, "A robust formant extraction algorithm combining spectral peak-picking and roots polishing," *Eurasip Journ. on Applied Signal Processing*, vol. 2006, pp. Article ID 67 960, 16 pages, 2006.

[9] 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.

[10] ——, "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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1997.

[17] Y. Park and H. Park, "Non-stationary sound source localization based on zero crossings with the detection of onset intervals," *IEICE Electronics Express*, vol. 5, no. 24, pp. 1054–1060, 2008.

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

[20] 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.

[21] 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.

[22] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, (accepted).

[23] M. Slaney, "Auditory Toolbox Version 2," *Interval Research Corporation Technical Report*, no. 10, 1998. [Online]. Available: http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/

[24] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA USA, Dec. 2010.

[25] CMU Sphinx Consortium Sphinx Consortium. CMU Sphinx Open Source Toolkit for Speech Recognition: Downloads. [Online]. Available: http://cmusphinx.sourceforge.net/wiki/download/

[26] S. G. McGovern, "A model for room acoustics," http://www.sgmaudio.com/research/rir/rir.html.

[27] 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.

[28] 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.