BINAURAL SOUND SOURCE SEPARATION MOTIVATED BY AUDITORY PROCESSING

Chanwoo Kim [1]*, Kshitiz Kumar* [2]*, and Richard M. Stern* [1,2]

[1]Language Technologies Institute
and [2]Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh PA 15213 USA
{chanwook, kshitizk, rms}@cs.cmu.edu

ABSTRACT

In this paper we present a new method of signal processing for robust speech recognition using two microphones. The method, loosely based on the human binaural hearing system, consists of passing the speech signals detected by two microphones through bandpass filtering. We develop a spatial masking function based on normalized cross-correlation, which provides rejection of off-axis interfering signals. To obtain improvements in reverberant environments, a temporal masking component, which is closely related to our previously-described de-reverberation technique known as SSF. We demonstrate that this approach provides substantially better recognition accuracy than conventional binaural sound-source separation algorithms.

Index Terms— Robust speech recognition, signal separation, interaural time difference, cross-correlation, auditory processing, binaural hearing

1. INTRODUCTION

In recent decades, speech recognition accuracy in clean environments has significantly improved. Nevertheless, it is frequently observed that the performance of speech recognizers is significantly degraded under noisy or mismatched environments. These environmental mismatch might be due to additive noise, channel distortion, reverberation, and so on. Maintaining good error rates in noisy conditions remains a problem that must be effectively resolved for speech recognition systems to be useful for real consumer products. Many algorithms have been developed to enhance speech recognition accuracy under noisy environments (*e.g.* [1, 2]).

It is well known that the human binaural system is very effective in its ability to separate sound sources even in difficult and cluttered acoustical environments (*e.g.* [3]). Motivated by these observations, many theoretical models (*e.g.* [4]) and computational algorithms (*e.g.* [4, 5, 6, 7]) have been developed using interaural time differences (ITDs), interaural intensity difference (IIDs), interaural phase differences (IPDs), and other cues. Combination of binaural information has also been employed, such as IPD and ITD (*e.g.* [7, 8, 9]), ITD and interaural level difference (ILD) combined with missing-feature recovery techniques (*e.g.* [10]), and ITD combined with reverberation masking (*e.g.* [11]).

In many of the algorithms above, either binary or continuous "masks" are developed to indicate which time-frequency bins are dominated by the target source. Typically this is done by sorting the time-frequency bins according to ITD (either calculated directly or inferred from estimated IPD). Spatial masks using ITD have been shown to be very useful for separating sound sources (*e.g.* [9]), but their effectiveness is reduced in reverberant environments. In [11], they incorporated reverberation masks, but this approach does not show improvement in purely reverberant environments (reverberation without noise) compared to the baseline system.

In this study we combine the use of a newly-developed form of single-microphone temporal masking that has proved to be very effective in reverberant environments with a new type of spatial masking that is both simple to implement and effective in noise. We evaluate the effectiveness of this combination of spatial and temporal masking (STM) in a variety of degraded acoustical environments.

2. SIGNAL SEPARATION USING SPATIAL AND TEMPORAL MASKS

2.1. Structure of the STM system

The structure of our sound source separation system, which crudely models some of the processing in the peripheral auditory system and brainstem, is shown in Fig. 1. Signals from the two microphones are processed by a bank of 40 modified gammatone filters [12] with the center frequencies of the filters linearly spaced according to Equivalent Rectangular Bandwidth (ERB) [13] between 100 Hz and 8000 Hz, using the implementation in Slaney's Auditory Toolbox [14]. As we have done previously (*e.g.* [15]), we convert the gammatone filters to a zero-phase form in order to impose identical group delay on each channel. The impulse responses of these filters $h_l(t)$ are obtained by computing the autocorrelation function of the original filter response:

$$h_l(t) = h_{g,l}(t) * h_{g,l}(-t) \tag{1}$$

where l is the channel index and $h_{g,l}(t)$ is the original gammatone response. While this approach compensates for the difference in group delay from channel to channel, it also causes the magnitude response to become squared, which results in bandwidth reduction. To compensate approximately for this, we intentionally double the bandwidths of the original gammatone filters at the outset. Even though doubling the bandwidth is not the perfect compensation, we observe that it is sufficient for practical purposes. We obtain binary spatial masks by calculating the normalized cross-correlation coefficient and comparing its value to a pre-determined threshold value, as described in detail in Sec. 2.2. Along with the spatial masks, we also generate binary temporal masks. This is accomplished by calculating the short-time power for each time-frequency bin and comparing this value to a short-time average value that had been obtained by IIR

This research was supported by the National Science Foundation (Grant IIS-I0916918). The authors are grateful to Prof. Bhiksha Raj for many useful discussions.

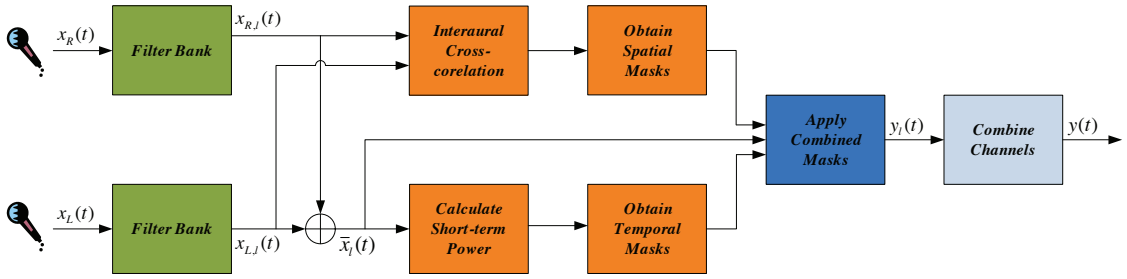

Fig. 1. The block diagram of the sound source separation system using spatial and temporal masks (STM).

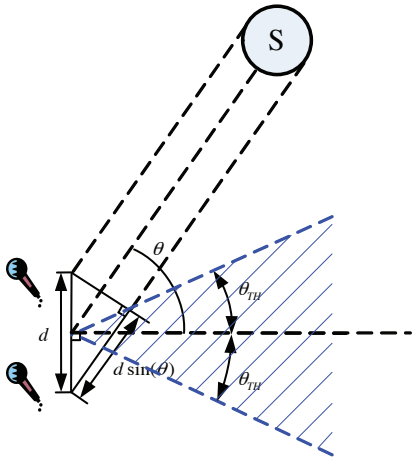

Fig. 2. Selection region for a binaural sound source separation system: if the location of the sound source is determined to be inside the shaded region, we assume that the signal is from the target.

lowpass filtering, as described in detail in Sec. 2.3. We obtain the final masks by combining these temporal masks and spatial masks as described in Sec. 2.4. To resynthesize speech, we combine the signals from each channel:

$$y(t) = \sum_{l=0}^{L-1} y_l(t) \tag{2}$$

where L is the number of channels (40 at present) and $y_l(t)$ is the signal from in each channel l after applying the masks. The final output of the system is $y(t)$.

2.2. Spatial mask generation using normalized cross-correlation

In this section, we describe the construction of the binary masks using normalized cross-correlation. In our previous research (*e.g.* [9] [16]), we have frequently observed that an analysis window that is longer than the conventional window of about 25 ms typically used for speech recognition is more effective in noise-robustness algorithms. Hence, we use a window length of 50 ms with 10 ms between analysis frames as in [17] for the present study. We define the normalized correlation $\rho(t_0, l)$ for the time-frequency segment that begins at $t = t_0$ and belongs to frequency bin l to be

$$\rho(t_0, l) = \frac{\frac{1}{T_0}\int_{T_0} x_{R,l}(t;t_0)x_{L,l}(t;t_0)dt}{\sqrt{\frac{1}{T_0}\int_{T_0}(x_{R,l}(t;t_0))^2dt}\sqrt{\frac{1}{T_0}\int_{T_0}(x_{L,l}(t;t_0))^2dt}} \tag{3}$$

where l is the channel index, $x_{R,l}(t;t_0)$ and $x_{L,l}(t;t_0)$ are the short-time signals from the left and right microphones after Hamming windowing, and t_0 refers to the time when each frame begins. If $x_{R,l}(t;t_0) = x_{L,l}(t;t_0)$, then $\rho(t_0, l) = 1$ in Eq. (3). $|\rho(t_0, l)|$ is less than one otherwise. We note that this statistic is widely used in models of binaural processing (*e.g.* [18]), although typically for different reasons.

Let us consider the case where the sound source is located at an angle θ as shown in Fig. 2. We assume that the desired signal is along the perpendicular bisector of the line between the two mics. This leads to a decision criterion in which a component is accepted if the putative location of the sound source for a particular time-frequency segment is within the shaded region (*i.e.* $|\theta| < \theta_{TH}$), and rejected otherwise. If the bandwidth of a filter is sufficiently narrow, then the signal after filtering can be approximated by the sinusoidal function [6]:

$$x_{R,l}(t;t_0) = A\sin(\omega_0 t) \tag{4a}$$
$$x_{L,l}(t;t_0) = A\sin(\omega_0(t-\tau)) \tag{4b}$$

where ω_0 is the center frequency of channel l. By inserting (4) into (3), we obtain the following simple relation:

$$\rho(t_0, l) = \cos(\omega_0\tau) = \cos\left(\omega_0 d\sin(\theta)\right) \tag{5}$$

As long as the microphone distance is small enough to avoid spatial aliasing, Eq. (5) implies that $\rho(t_0, l)$ decreases monotonically as $|\theta|$ increases. Thus, we can retain a given time-frequency bin if $\rho(t_0, l) \geq \rho_{TH}$ and reject it if $\rho(t_0, l) < \rho_{TH}$, where for each channel ρ_{TH} is given by $\rho_{TH} = \cos\left(\omega_0 d\sin(\theta_{TH})\right)$.

2.3. Temporal mask generation using modified SSF processing

Our temporal masking generation approach is based on a modification of the SSF approach introduced in [19]. First, we obtain the short-time power for each time-frequency bin:

$$P_l[m] = \int_{T_0}^{T_0+T_f} \left(\bar{x}_l(t;t_0)\right)^2 dt \tag{6}$$

where $\bar{x}(t;t_0)$ is the short-time average of $x_{L,l}(t;t_0)$ and $x_{R,l}(t;t_0)$, which are the Hamming-windowed signals at time t_0 in Channel l from the two microphones. The index of the frame that begins at

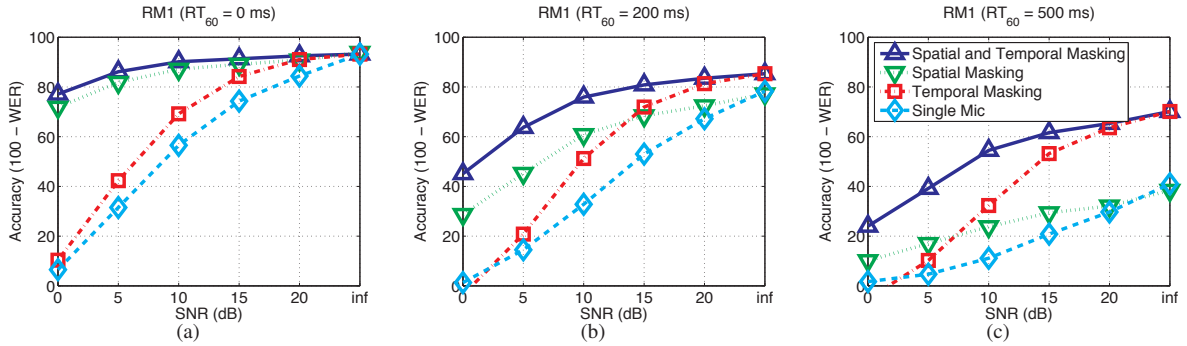

Fig. 3. Dependence of recognition accuracy on the type of mask used (spatial vs temporal) for speech from the DARPA RM corpus corrupted by an interfering speaker located at 30 degrees, using various simulated reverberation times: (a) 0 ms (b) 200 ms (c) 500 ms. We used a threshold angle of 15 degrees with STM, PDCW, and ZCAE algorithms.

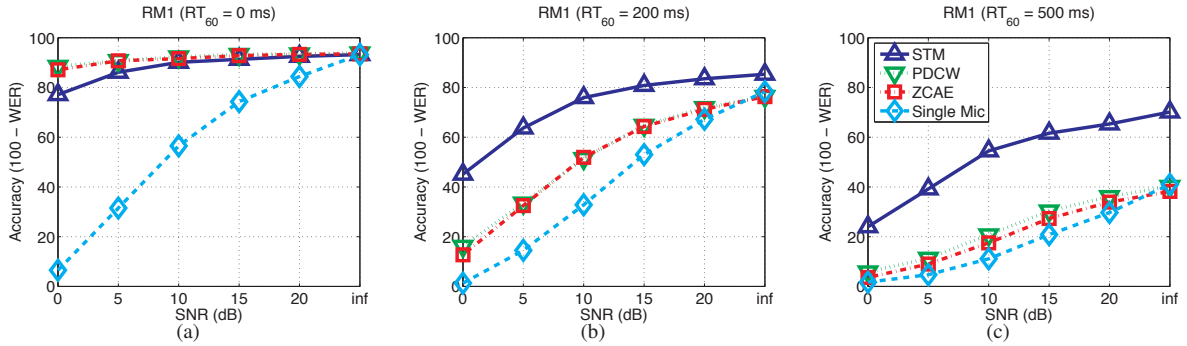

Fig. 4. Comparison of recognition accuracy using the STM, PDCW, and ZCAE algorithms for the DARPA RM database corrupted by an interfering speaker located at 30 degrees, using various simulated reverberation times: (a) 0 ms (b) 200 ms (c) 500 ms.

$t = t_0$ is m, and T_f is the window length. As in [19], we obtain a first-order IIR lowpassed output:

$$Q_l[m] = \lambda Q_l[m-1] + (1-\lambda)P_l[m] \tag{7}$$

where λ is the forgetting factor which determines the bandwidth of the filter. Based on a pilot study in [19], we use the value $\lambda = 0.04$. If the power in a specific time-frequency bin is less than the lowpassed output developed in Eq. (7), we assume that it is masked by temporal masking, so we accept a time-frequency segment if $P_l[m] \geq Q_l[m]$ and reject it if $P_l[m] < Q_l[m]$.

2.4. Application of spatial and temporal masks

If a specific time-frequency bin must be accepted by both the spatial and temporal masking processes described Secs. 2.2 and 2.3, then this time-frequency bin is selected; otherwise it is rejected. Binary masking is applied according to the following equation:

$$\begin{cases} y_l(t,t_0) = \bar{x}_l(t,t_0) & \texttt{if selected} \\ y_l(t,t_0) = \mu \bar{x}_l(t,t_0) & \texttt{if rejected} \end{cases} \tag{8}$$

where μ is a scaling factor that suppresses (but does not annihilate) the signal in the rejected time-frequency bin. The signal $y_l(t,t_0)$ is the short-time signal in each time-freqeuency bin after applying the mask, and $\bar{x}_l(t,t_0)$ is the average of the left and right short-time signals starting at time t_0 in the l^{th} channel.

In previous work (*e.g.* [20]), we have observed that power flooring (*i.e.* the imposition of a minimum power) is very important for robust speech recognition. In this study as in others the choice of the power flooring coefficient μ is important to prevent power from approaching zero too closely. In pilot work we have found the following scaling factor to be useful:

$$\mu = \sqrt{\frac{\delta\left(\frac{1}{T}\int_0^T \bar{x}_l^2(t)dt\right)}{\frac{1}{T_f}\int_0^{T_f} \bar{x}_l^2(t;t_0)dt}} \tag{9}$$

In the above equation, $\bar{x}_l(t)$ is the average of the left and right signals for this l^{th} channel for this utterance, T is the length of the entire utterance, and T_f is the frame length (which is 50 ms in our implementation). The above equation means that the input power of time-frequency bins that are rejected is reduced to δ times the average power $\left(\frac{1}{T}\int_0^T \bar{x}_l^2(t)dt\right)$ in this channel. We have found that $\delta = 0.01$ is a suitable coefficient.

3. EXPERIMENTAL RESULTS AND CONCLUSIONS

In this section we present experimental results using the STM algorithm described in this paper. We assume a room of dimensions 5 x 4 x 3 m, with two microphones located at the center of the room. The distance between two microphones is 4 cm. The target is located 1.5 m away from the microphones along the perpendicular bisector of the line connecting two microphones, and an interfering speaker is located at 30 degrees to one side and 1.5 m away from the microphones. The target and interfering signals are digitally added

after simulating reverberation effects using the RIR software package. We used `sphinx_fe` included in `sphinxbase 0.4.1` for speech feature extraction, `SphinxTrain_1.0` for speech recognition training, and `Sphinx3.8` for decoding, all of which are readily available in Open Source form. We used a subset of 1600 utterances from the DARPA Resource Management (RM1) training data for acoustic modeling and a subset of 600 utterances from the RM test data for evaluation.

Figure 3 describes the contributions of spatial masking and temporal masking in the environments considered. We note that while temporal masking scheme must be applied both to training and test data to avoid increased Word Error Rate (WER) due to environmental mismatch, the system performance is essential the same regardless of whether spatial masking is used in training or no. This is not surprising, as spatial masking should routinely accept all components of clean speech from the target location.

In the anechoic environment (Fig. 3(a)), we observe that improvement with the STM algorithm is mostly provided by spatial masking, with temporal masking providing only marginal improvement. If T_{60} is increased to 200 ms (Fig. 3(b)), or 500 ms (Fig. 3(c)), however, we observe that the contribution of temporal masking becomes quite substantial. When both noise and reverberation are present, the contributions of temporal and spatial maskings are complementary and synergistic.

Figure 4 compares speech recognition accuracy for several algorithms including the STM system described in this paper, the Phase Difference Channel Weighting (PDCW) [9], and the Zero Crossing Amplitude Estimation (ZCAE) in [6], all using binary masking. To compare the performance of these different systems in the same condition, we used a threshold angle of 15 degrees with all algorithms to obtain binary masks. In the anechoic condition (Fig. 4(a)), the STM approach provided slightly worse performance than the PDCW and ZCAE algorithms. In reverberant environments the STM system provides the best results by a very large margin, and the PDCW results were slightly better than the corresponding ZCAE results. In terms of computational cost, PDCW requires the least amount of computation due to its efficient frequency-domain implementation, while STM and ZCAE require much more computation due to time-domain filtering.

The MATLAB code for the STM algorithm can be found at `http://www.cs.cmu.edu/~robust/archive/algorithms/STM_ICASSP2011/`.

4. REFERENCES

[1] 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.

[2] 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.

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 58–67, Jan. 2006.

[11] K. J. Palomaki, G. J. Brown, and DeL. 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.

[12] P. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory and Perception*, Oxford, UK, 1992, pp. 429–446, Y. Cazals, L. Demany, and K. Horner, (Eds), Pergamon Press.

[13] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acustica - Acta Acustica*, vol. 82, pp. 335–345, 1996.

[14] M. Slaney, "Auditory toolbox version 2," *Interval Research Corporation Technical Report*, , no. 10, 1998.

[15] 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.

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

[18] C. Trahiotis, L. Bernstein, R. M. Stern, and T. N. Buell, "Interaural correlation as the basis of a working model of binaural processing," in *Sound Source Localization*, R. Fay and T. Popper, Eds., vol. 25 of *Springer Handbook of Auditory Research*, pp. 238–271. Springer-Verlag, 2005.

[19] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH-2009*, Sept. 2009, pp. 28–31.

[20] C. Kim and R. M. Stern, "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.