Automatic Selection of Thresholds for Signal Separation Algorithms Based on Interaural Delay

Chanwoo Kim [1], *Richard M. Stern*[2], Kiwan Eom[3], and Jaewon Lee[4]

[2] Department of Electrical and Computer Engineering,
and [1,2]Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213 USA,
[3,4]Samsung Electronics, Suwon, Korea

chanwook@cs.cmu.edu, rms@cs.cmu.edu, kiwan.eom@samsung.com, jwonlee@samsung.com

Abstract

In this paper we describe a system that separates signals by comparing the interaural time delays (ITDs) of their time-frequency components to a fixed threshold ITD. While in previous algorithms the fixed threshold ITD had been obtained empirically from training data in a specific environment, in real environments the characteristics that affect the optimal value of this threshold are unknown and possibly time varying. If these configurations are different from the environment under which the ITD threshold had been pre-computed, the performance of the source separation system is degraded. In this paper, we present an algorithm which chooses a threshold ITD that minimizes the cross-correlation of the target and interfering signals, after a compressive nonlinearity. We demonstrate that the algorithm described in this paper provides speech recognition accuracy that is much more robust to changes in environment than would be obtained using a fixed threshold ITD.

Index Terms: Robust speech recognition, speech enhancement, signal separation, time delay analysis, phase difference analysis, cross correlation

1. Introduction

Following the introduction of statistical approaches such as HMMs (hidden Markov models) and SLMs (statistical language models) (*e.g.* [1] [2]), speech recognition accuracy has significantly improved. Nevertheless, maintaining good error rates in noisy conditions remains a problem that must be effectively resolved for speech recognition systems to be used in real consumer products in difficult acoustical environments.

It is well known that the human binaural system is remarkable in its ability to separate sound sources even in a very difficult environment (*e.g.* [3]). Motivated by these observations, many models and algorithms have been developed using interaural time differences (ITDs), interaural intensity difference (IIDs), interaural phase differences (IPDs), and other cues (*e.g.* [4, 5, 6, 7]). IPD and ITD have been extensively used in binaural processing because this information can be easily obtained by spectral analysis (*e.g.* [4] [8] [9]).

In many of the algorithms above either a binary or continuous "mask" is developed that indicates which time-frequency bins are believed to be 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). In both cases performance depends on how the ITD threshold is selected, and the optimal threshold depends on the configuration of the noise sources including their locations and strength. If the optimal ITD from a particular environment is applied to a somewhat different environment, the system performance will be degraded. In addition, the characteristics of the environment typically vary with time.

When target identification is obtained by a binary mask based on an ITD threshold, the value of that threshold is typically estimated from development test data. As noted above, the optimal ITD threshold itself will depend on the number of noise sources and their locations, both of which may be time-varying. If the azimuth of the noise source is very different from that of the target, an ITD threshold that is relatively far from that of the target may be helpful. On the other hand, if an interfering noise source is very close to the target and we use a similar ITD threshold, the system will also classify many components of the interfering signal as part of the target signal. If there is more than one noise source, or if the noise sources are moving, the problem becomes even more complicated.

In our approach, which is summarized in Fig. 1, we construct two complementary masks using a binary threshold. Using these two complementary masks, we obtain two different spectra: one for the target and the other for everything except for the target. From these spectra, we obtain the short-time power for the target and the interference. These power sequences are passed through a compressive nonlinearity. We compute the cross-correlation coefficient for the two resulting power sequences, and we obtain the ITD threshold by minimizing the correlation coefficient.

In Sec. 2 we review how to obtain the ITD from phase difference information. In Sec. 3, we explain how to construct the complementary masks and how to obtain the ITD threshold based on the minimum correlation criterion. We present experimental results in Sec. 4.

2. Obtaining ITDs from interaural phase differences

In this section we review the procedure for obtaining ITD from phase information (*e.g.* [9]). Let $x_L[n]$ and $x_R[n]$ be the signals from the left and right microphones, respectively. We assume that we know where the target source is located and, without loss of generality, we assume that it is placed at the perpendicular bisector of the line between two microphones.

Suppose that the total number of interfering sources is S. Each source $s, 1 \leq s \leq S$ has an ITD of $d_s[m,k]$ where m is the frame index and k is the frequency index. Note that both S and $d_s[m,k]$ are unknown. We assume that $x_0[n]$ represents the

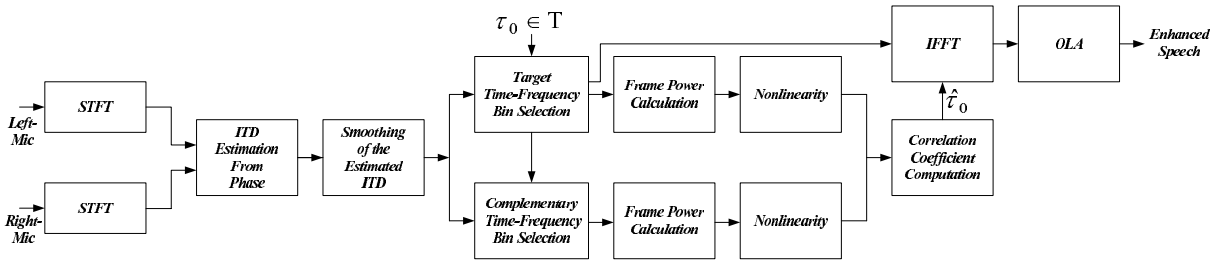

Figure 1: *Block diagram of the optimal ITD selection algorithm for sound source separation.*

target signal and that the notation $x_s[n], 1 \leq s \leq S$, represents signals from each interfering source received from the "left" microphone. In the case of signals from the "right" microphone, the target signal is still $x_0[n]$, but the interfering signals are delayed by $d_s[m,k]$. These assumptions imply that for the target signal $x_0[n]$, $d_0[m,k] = 0$ for all m and k.

To perform spectral analysis, we obtain the following short-time signals by multiplication with a Hamming window $w[n]$:

$$x_L[n;m] = x_L[n - mL_{fp}]w[n] \quad (1a)$$
$$x_R[n;m] = x_R[n - mL_{fp}]w[n] \quad (1b)$$
$$\text{for } 0 \leq n \leq L_{fl} - 1$$

where m is the frame index, L_{fp} is the number of samples between frames, and L_{fl} is the frame length. The window $w[n]$ is a Hamming window with a length of L_{fl}. We use a 75-ms window length based on previous results [9]. The short-time Fourier transforms of (1) can be represented as

$$X_L[m, e^{j\omega_k}) = \sum_{s=0}^{S} X_s[m, e^{j\omega_k}) \quad (2a)$$
$$X_R[m, e^{j\omega_k}) = \sum_{s=0}^{S} e^{-jw_k d_s[m,k]} X_s[m, e^{j\omega_k}) \quad (2b)$$

where $w_k = 2\pi k/N$ and N is the FFT size. We represent the strongest sound source for a specific time-frequency bin $[m,k]$ as $s^*[m,k]$. This leads to the following approximation:

$$X_L[m, e^{j\omega_k}) \approx X_{s^*[m,k]}[m, e^{-jw_k}) \quad (3a)$$
$$X_R[m, e^{j\omega_k}) \approx e^{-jw_k d_{s^*[m,k]}[m,k]} \times X_{s^*[m,k]}[m, e^{-jw_k}) \quad (3b)$$

Note that $s^*[m,k]$ may be either 0 (the target source) or $1 \leq s \leq S$ (any of the interfering sources). From (3), The ITD for a particular time-frequency bin $[m,k]$ is given by:

$$|d_{s^*[m,k]}[m,k]| \approx \frac{1}{|w_k|} \min_r \left| \angle X_R[m, e^{-jw_k}) - \angle X_L[m, e^{-jw_k}) - 2\pi r \right| \quad (4)$$

Thus, based on whether the obtained ITD from (4) is within a certain range of the target ITD (which is zero), we can make a simple binary decision concerning whether the time-frequency bin $[m,k]$ is likely to belong to the target speaker or not.

3. Optimal ITD threshold selection from complementary masks

This algorithm is based on two complementary binary masks, one that identifies time-frequency components that are believed to belong to the target signal and the other that identifies the components that are believed to belong to the interfering signals ($i.e.$ everything except the target signal). These masks are used to construct two different spectra corresponding to the power sequences representing the target and the interfering sources. We apply a compressive nonlinearity to these power sequences, and define the optimal ITD threshold to be the threshold that minimizes the cross-correlation between these two output sequences (after the nonlinearity).

Computation is performed in discrete fashion, considering a set T of a finite number of possible ITD candidates. We determine which element of this set is the most appropriate ITD threshold by performing an exhaustive search over the set T. Let us consider one element of this set, τ_0. We obtain the target mask and the complementary mask for τ_0 for $0 \leq k \leq N/2$:

$$\mu_T[m,k] = \begin{cases} 1, & \text{if } |d[m,k]| \leq \tau_0 \\ \delta, & \text{otherwise} \end{cases} \quad (5a)$$
$$\mu_I[m,k] = \begin{cases} \delta, & \text{if } |d[m,k]| > \tau_0 \\ 1, & \text{otherwise} \end{cases} \quad (5b)$$

For $N/2 \leq k \leq N-1$, we use the following symmetry condition:

$$\mu_I[m,k] = \mu_T[m, N-k], \quad N/2 \leq k \leq N-1 \quad (6a)$$
$$\mu_I[m,k] = \mu_I[m, N-k], \quad N/2 \leq k \leq N-1 \quad (6b)$$

In other words, we assume that time-frequency bins for which $|d(m,k)| < \tau_0$ are presumed to belong to the target speaker, and that time-frequency bins for which $|d[m,k]| > \tau_0$ belong to the noise source. We are presently using a value of 0.01 for the floor constant δ. The masks $\mu_T[m,k]$ and $\mu_I[m,k]$ in (5) are applied to $\bar{X}[m, e^{j\omega_k})$, the averaged signal spectrogram from the two microphones:

$$\bar{X}[m, e^{j\omega_k}) = \frac{1}{2}\{X_L[m, e^{j\omega_k}) + X_R[m, e^{j\omega_k})\} \quad (7)$$

Using this procedure, we obtain the target spectra $X_T[m, e^{j\omega_k}|\tau_0)$ and the interference spectra $X_I[m, e^{j\omega_k}|\tau_0)$ as shown below:

$$X_T[m, e^{j\omega_k}|\tau_0) = \bar{X}[m, e^{j\omega_k})\mu_T[m, e^{j\omega_k}) \quad (8a)$$
$$X_I[m, e^{j\omega_k}|\tau_0) = \bar{X}[m, e^{j\omega_k})\mu_I[m, e^{j\omega_k}) \quad (8b)$$

In the above equation, we explicitly include τ_0 to show that the masked spectrum will depend on the ITD threshold. Using the

spectra $X_T[m, e^{j\omega_k})$ and $X_I[m, e^{j\omega_k})$, we obtain the power:

$$P_T[m|\tau_0) = \sum_{k=0}^{N-1} \left| X_T[m, e^{j\omega_k}) \right|^2 \quad (9a)$$

$$P_I[m|\tau_0) = \sum_{k=0}^{N-1} \left| X_I[m, e^{j\omega_k}) \right|^2 \quad (9b)$$

Because the power signals in (9) have a very dynamic range, it is not very helpful to obtain the cross-correlation directly from them. A reasonable way of reducing dynamic range is by applying a compressive nonlinearity, which may be considered to represent the perceived loudness of the sound. While many nonlinearities have been proposed to characterize the relationship between signal intensity and perceived loudness we chose the following power-law nonlinearity motivated by previous work [10, 11]):

$$R_T[m|\tau_0) = P_T[m|\tau_0)^{a_0} \quad (10a)$$

$$R_I[m|\tau_0) = P_I[m|\tau_0)^{a_0} \quad (10b)$$

where $a_0 = 1/15$ as in [11].

The cross-correlation coefficient of the signals in (10) is obtained as follows:

$$\rho_{T,I}(\tau_0) = \frac{\frac{1}{N}\sum_{m=1}^{M} R_T[m|\tau_0) R_I[m|\tau_0) - \mu_{R_T}\mu_{R_I}}{\sigma_{R_T}\sigma_{R_I}} \quad (11)$$

where μ_{R_1} and μ_{R_2} and σ_{R_T} and σ_{R_I} are the means and standard deviations of $R_T[m|\tau_0)$ and $R_I[m|\tau_0)$, respectively.

The threshold τ_0 is then selected to minimize the absolute value of the cross-correlation

$$\hat{\tau}_0 = \arg\min_{\tau_0} |\rho_{T,I}(\tau_0)| \quad (12)$$

4. Experimental results

In this section we present experimental results using the ITD threshold selection algorithm described in this paper. We compare an IPD-based signal separation system using binary masks, and using the automatically-determined ITD threshold as described above, to a similar system that uses a fixed ITD threshold. In the experiments below we assume a room of dimensions 5 x 4 x 3 m, with microphones that are located at the center of the room. The target is 2 m away from the microphone along the perpendicular bisector of the line between two microphones. The target and noise signals are digitally added after simulating reverberation effects using the `Room Impulse Response (RIR)` software [12]. The two microphones were placed 4 cm apart from one another. 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 subsets of 1600 utterances and 600 utterances, respectively, from the DARPA Resource Management (RM1) database for training and testing.

For the fixed-ITD threshold system, we obtained the optimal threshold by conducting an experiment in a specific environment: we located the interfering speaker along a 45-degree line to the side of the perpendicular bisector of the line between two microphones, and the interfering speaker generating a speech noise at 0-dB signal-to-interference ratio (SIR). We further assumed that there was no reverberation in this room.

We conducted two different sets of experiments. In the first set of the experiments, we kept the geometrical configuration

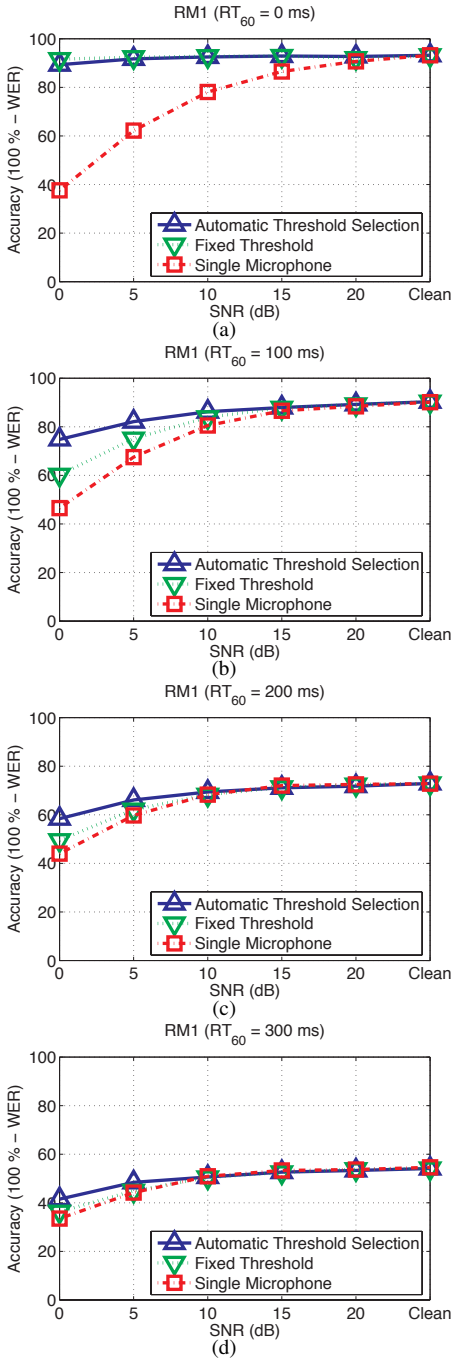

Figure 2: *Comparison of recognition accuracy for the DARPA RM database corrupted by an interference speaker located at 45 degrees at different reverberation times (a) 0 ms (b) 100 ms (c) 200 ms (d) 300 ms.*

the same as the above, changing only the SIR and reverberation time. As shown in Fig. 2, in the absence of reverberation at 0-dB SIR, both the fixed ITD system and the automatic-ITD system show comparable performance. If reverberation is present, however, the automatic-ITD system provides substantially better performance than the fixed-ITD signal separation system.

In the second set of the experiments, we changed the location of the interfering speaker while maintaining the SIR at 0 dB. As shown in Fig. 3(a), even if the SIR is the

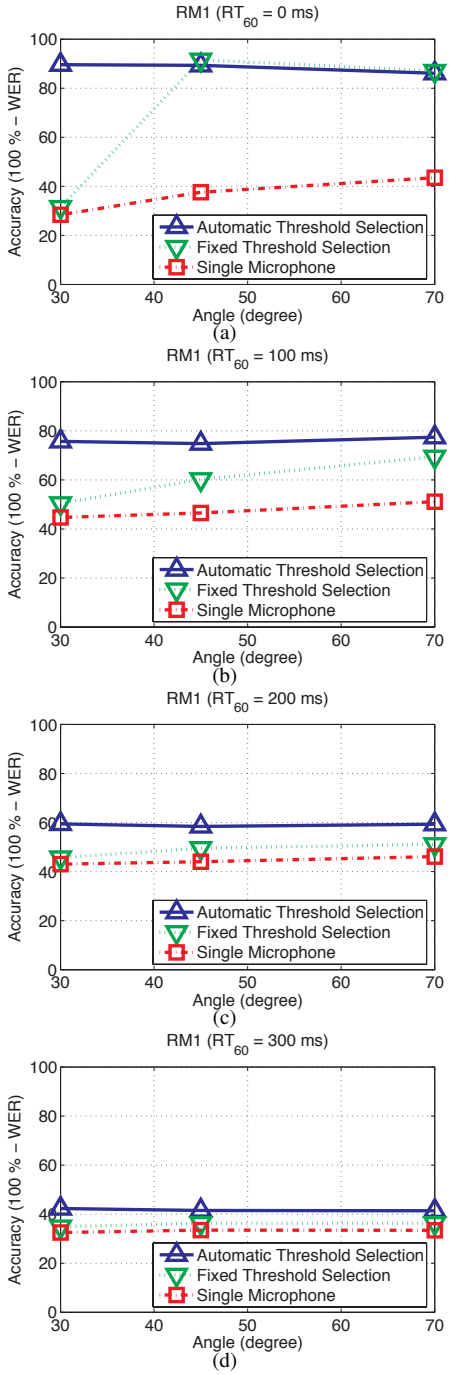

Figure 3: *Comparison of recognition accuracy for the DARPA RM database corrupted by an interference speaker located at different locations at different reverberation time (a) 0 ms (b) 100 ms (c) 200 ms (d) 300 ms.*

same as in the calibration environment, the fixed-ITD threshold system produces substantially degraded performance if the actual interfering speaker location is different from the location used in the calibration environment. The automatic-ITD-threshold selection system provides recognition results that are much more robust with respect to the locations of the interfering sources. In this figure we observe that as the interfering speaker moves toward the target, the fixed-ITD threshold PD system shows increased word error rate. We repeated the same experiment by changing the simulated reverberation time. As shown in Figs. 3(b) to 3(d), the automatic-threshold-selection algorithm provides consistently better recognition accuracy than the fixed threshold system, as expected. MATLAB code for this algorithm may be found at `http://www.cs.cmu.edu/~robust/archive/algorithms/AUTOITD_IS2010/`.

5. Conclusions

In this paper we present a new algorithm which selects an ITD threshold by minimizing the correlation of nonlinearity power from the masked and non-masked spectral regions. Experimental results show while the conventional fixed ITD threshold system shows degraded performance in unmatched conditions, this automatic ITD threshold selection algorithm makes the binary mask system much more reliable.

6. Acknowledgements

This research was supported by Samsung Electronics. The authors are grateful to Professor Bhiksha Raj and Kshitiz Kumar for many helpful discussions. A US Patent has been filed for [13] based on the work described in this paper.

7. References

[1] J. K Baker, "The dragon system - An overview," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 23, no. 1, pp. 24–29, Feb. 1975.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of IEEE*, vol. 64, no. 4, pp. 532–556, Apr. 1976.

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

[4] 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.

[5] S. Srinivasan, M. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Comm.*, vol. 48, no. 11, pp. 1486–1501, Nov. 2006.

[6] 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.

[7] 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.

[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. 149 – 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] 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.

[11] ——, "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.

[12] S. G. McGovern, "A model for room acoustics," http://2pi.us/rir.html.

[13] C. Kim, R. M. Stern, K. Eom, and J. Lee, "Automatic interaural time delay threshold selection method for sound source separation," *United States Patent (Filed)*, 2010.