Signal Separation for Robust Speech Recognition Based on Phase Difference Information Obtained in the Frequency Domain

Chanwoo Kim, Kshitiz Kumar, Bhiksha Raj, and Richard M. Stern

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

Abstract

In this paper, we present a new two-microphone approach that improves speech recognition accuracy when speech is masked by other speech. The algorithm improves on previous systems that have been successful in separating signals based on differences in arrival time of signal components from two microphones. The present algorithm differs from these efforts in that the signal selection takes place in the frequency domain. We observe that additional smoothing of the phase estimates over time and frequency is needed to support adequate speech recognition performance. We demonstrate that the algorithm described in this paper provides better recognition accuracy than time-domain-based signal separation algorithms, and at less than 10 percent of the computation cost.

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

1. Introduction

Speech recognition systems have significantly improved in the past decades but noise robustness and computational complexity remain critical issues. A number of algorithms have shown improvements for stationary noise (*e.g.* [1, 2]). Nevertheless, improvement in non-stationary noise remains a difficult issue (*e.g.* [3]). In these environments, auditory processing [4] and missing-feature-based approaches [5] are promising. An alternative approach is signal separation based on analysis of differences in arrival time (*e.g.* [6, 7, 8]). It is well documented that the human binaural system bears remarkable ability in speech separation (*e.g.* [8]). Many models have been developed that describe various binaural phenomena (*e.g.* [9, 10]), typically based on interaural time difference (ITD), interaural phase difference (IPD), interaural intensity difference (IID), or changes of interaural correlation.

The Zero Crossing Amplitude Estimation (ZCAE) algorithm was recently introduced by Park [7] which is similar in some respects to work by Srinivasan *et al.* [6]. These algorithms (and similar ones by other researchers) typically analyze incoming speech in bandpass channels and attempt to identify the subset of time-frequency components for which the ITD is close to the nominal ITD of the desired sound source (which is presumed to be known *a priori*). The signal to be recognized is reconstructed from only the subset of "good" time-frequency components. This selection of "good" components is frequently treated in the computational auditory scene analysis (CASA) literature as a multiplication of all components by a binary mask that is nonzero for only the desired signal components. Although ZCAE provides impressive performance even at low SNRs, it is very computationally intensive, which makes it unsuitable for hand-held devices.

The goals of this work are twofold. First, we would like to obtain improvements in word error rate (WER) for speech recognition systems that operate in real world environments that include noise and reverberation. We also would like to develop a computationally efficient algorithm than can run in real time in embedded systems. In the present ZCAE algorithm much of the computation is taken up in the bandpass filtering operations. We found that computational cost could be significantly reduced by estimating the ITD through examination of the phase difference between the two sensors in the frequency domain. We describe in the sections below how the binary mask is obtained using frequency information. We also discuss the duration and shape of the analysis windows, which can contribute to further improvements in WER.

The rest of the paper is organized as follows: Sec. 3 describes our algorithm at a general level. We propose our time-frequency weighting scheme in Sec. 3. Experimental results are discussed in Sec.4, and we summarize our work in Sec. 5.

2. Phase-difference-based binary time-frequency mask estimation

Our work on signal separation is motivated by binaural speech processing. Sound sources are localized and separated by the human binaural system primarily through the use of ITD information at low frequencies and IID information at higher frequencies, with the crossover point between these two mechanisms considered to be based on the physical distance between the two ears and the need to avoid spatial aliasing (which would occur when the ITD between two signals exceeds half a wavelength). In our work we focus on the use of ITD cues and avoid spatial aliasing by placing the two microphones closer together than occurs anatomically. When multiple sound sources are presented, it is generally assumed that humans attend to the desired signal by attending only to information at the ITD corresponding to the desired sound source.

Our processing approach, which we refer to as Phase Difference Channel Weighting (PDCW), crudely emulates human binaural processing, and is summarized in Fig. 1. Briefly, the system first performs a short-time Fourier transform (STFT) which decomposes the two input signals in time and in frequency. ITD is estimated indirectly by comparing the phase information from the two microphones at each frequency, and the time-frequency mask identifying the subset of ITDs that are "close" to the ITD of the target speaker is identified. A set of channels is developed by weighting this subset of time-

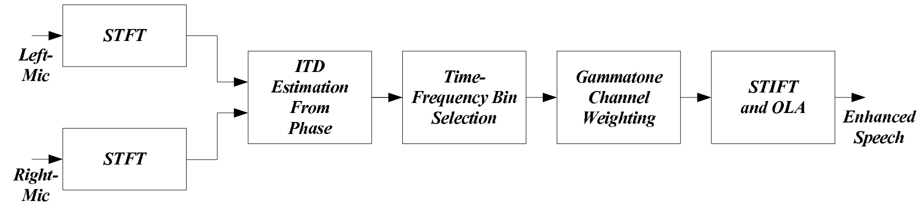
Left-Mic
STFT
Right-Mic
STFT
ITD Estimation From Phase
Time-Frequency Bin Selection
Gammatone Channel Weighting
STIFT and OLA
Enhanced Speech

Figure 1: *The block diagram of the Phase Difference Channel Weighting (PDCW)) algorithm*

frequency components using a series of Gammatone functions, and the time domain signal is obtained by the overlap-add method. As noted above, the principal novel feature in this paper is the use of interaural phase information in the frequency domain rather than ITD, IPD, or IID information in the time domain to obtain the binary mask.

Consider the two signals that are input to the system which we refer to as $x_L[n]$ and $x_R[n]$. We assume that the location of the desired target signal is known *a priori* and without loss of generality we assume its ITD to be equal to zero. For mathematical convenience, we refer to the number of interfering sources as L, with $\delta(l)$ being their respective ITDs. Note that both L and $\delta(l)$ are unknown. With the above formulations, the signals are the microphones are

$$x_L[n] = \sum_{l=0}^{L} x_l[n] \ , \quad x_R[n] = \sum_{l=0}^{L} x_l[n - \delta(l)] \qquad (1)$$

with $x_0[n]$ representing the target signal, $x_l(l \neq 0)$ representing interfering signals, x_L and x_R, respectively, representing the signals at the left and right microphones. The corresponding short-time Fourier transforms can be represented as

$$X(k,m) = \sum_{n=-\infty}^{\infty} x[n]w[m-n]e^{-j2\pi kn/N} \qquad (2)$$

$$X_L(k,m) = \sum_{i=0}^{L} X_i(k,m) \qquad (3)$$

$$X_R(k,m) = \sum_{i=0}^{L} e^{-jw_k d_i(k,m)} X_i(k,m) \qquad (4)$$

where $w[n]$ is a finite-duration Hamming window, k indicates one of N frequency bins, with positive frequency samples corresponding to $w_k = 2\pi k/N$ for $0 \leq w_k \leq N/2 - 1$. In our work N equals 512 for 26.5-ms windows and 2048 for 75-ms windows. Note that even though (1) indicates that signals at the microphones are identical except for a time delay, it is more appropriate that we consider the time delays associated with each frequency component of the signal. Correspondingly, we replace the frequency-independent ITD parameter δ in (1) by the frequency-dependent ITD parameter $d(k,m)$ in (4). Next, we assume that a specific time-frequency bin (k_0, m_0), is dominated by a single sound source l. This leads to

$$X_L(k_0;m_0) \approx X_{l^*}(k_0,m_0) \qquad (5)$$

$$X_R(k_0;m_0) \approx e^{-jw_{k_0} d(k_0,m_0)} X_{l^*}(k_0,m_0) \qquad (6)$$

where the source l^* dominates the time-frequency bin (k_0, m_0). This leads to a simple binary decision concerning whether the time-frequency bin (k_0, m_0) belongs to the target speaker or not. The frequency-dependent ITD $d(k,m)$ for a particular time-frequency bin (k_0, m_0) is

$$|d(k_0,m_0)| \approx \qquad (7)$$

$$\frac{1}{|w_{k_0}|} \min_r |\angle X_R(k_0,m_0) - \angle X_L(k_0,m_0) - 2\pi r|$$

for positive values of w_n of positive value, as discussed above, from which we derive the binary masking criterion

$$\mu(k_0,m_0) = \begin{cases} 1, \text{if } |d(k_0,m_0)| \leq \tau \\ \eta, \text{otherwise} \end{cases} \qquad (8)$$

In other words, only time-frequency bins for which $|d(k_0,m_0)| < \tau$ are presumed to belong to the target speaker. We are presently using a value of 0.01 for the floor constant η. The mask $\mu(k,m)$ in (8) is applied to $\bar{X}(k,m)$, the averaged signal spectrogram from the two channels, and speech is reconstructed from the $\tilde{X}(k,m)$ where

$$\bar{X}(k,m) = \frac{1}{2}\{X_L(k,m) + X_R(k,m)\} \qquad (9)$$

$$\tilde{X}(k,m) = \mu(k,m)\,\bar{X}(k,m) \qquad (10)$$

In Figure 2 we plot typical example of spectra from a signal that is corrupted by an interfering speaker with a signal-to-interference ratio (SIR) of 5 dB. We discuss two extensions to the basic PDCW algorithm in the next section.

3. Smoothed phase-difference-based binary mask estimation

While the basic procedure described in Sec. 2 provides signals that are audibly separated, the phase estimates are generally too noisy to provide useful speech recognition accuracy. In this section we discuss the implementation of two methods that smooth the estimates over frequency and time.

3.1. Gammatone channel weighting

As noted above, the estimates produced by Eq. (8) are generally noisy and must be smoothed. To achieve smoothing along frequency, we use a gammatone weighting that functions in a similar fashion to that of the familiar triangular weighting in MFCC features. Specifically, we obtain the gammatone channel weighting coefficients $w(i,m)$ according to the equation

$$w(i,m) = \frac{\sum_{k=0}^{\frac{N}{2}-1} \mu(k,m) \left|\bar{X}(k;m) H_i(k)\right|}{\sum_{k=0}^{\frac{N}{2}-1} \left|\bar{X}(k;m) H_i(k)\right|} \qquad (11)$$

where $\mu(k,m)$ is the original binary mask that is obtained using (8). With this weghting we effectively map the ITD for each of the 256 original frequencies to an ITD for what we refer to as one of $I = 40$ channels. Each of these channels is associated with H_i, the frequency response of one of a set of gammatone filters with center frequencies distributed according to the Equivalent Rectangular Bandwidth (ERB) scale [11]. The final spectrum weighting is obtained using the gammatone mask μ_g

$$\mu_g(k,m) = \frac{\sum_{i=0}^{I-1} w(i,m) \left|H_i(k)\right|}{\sum_{i=0}^{I-1} \left|H_i(k)\right|} \qquad (12)$$

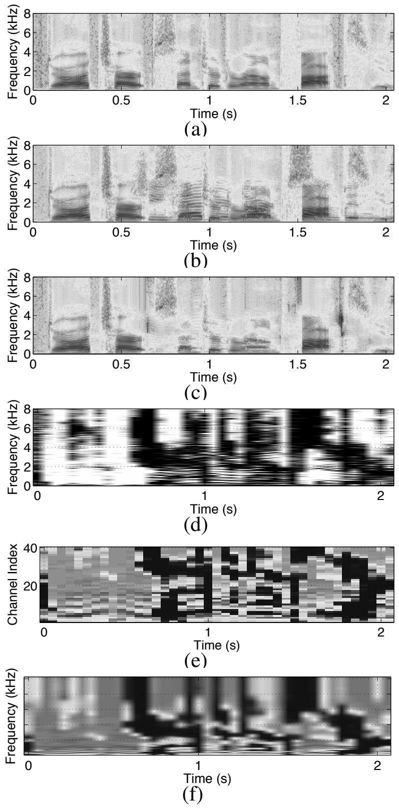

Figure 2: *Sample spectrograms illustrating the effects of PDCW processing. (a) original clean speech, (b) noise-corrupted speech, (c) reconstructed (enhanced) speech (d) the time-frequency mask obtained with (8) (e) gammatone channel weighting obtained from the time-frequency mask in (11) (e) final frequency weighting shown in (12) (f) enhanced speech spectrogram using the entire PDCW algorithm*

Examples of $w(i,m)$ and $\mu_g(k,m)$ are shown shown for a typical spectrum in Fig. 2(e) and Fig. 2(f), respectively, with an SIR of 5 dB as before. The reconstructed spectrum is given by:

$$\tilde{X}\left(k,m\right) = \max\left(\mu_g(k,m),\eta\right)\ \bar{X}\left(k;m\right) \qquad (13)$$

where again we use $\eta = 0.01$ as in (8).

3.2. The effect of the window length

In conventional speech coding and speech recognition systems, we generally use a length of approximately 20 to 30 ms for the Hamming window $w[n]$ in order to capture effectively the temporal fluctuations of speech signals. Nevertheless, longer observation durations are usually better for estimating environmental parameters. Using the procedures described below in Sec. 4, we considered the effect of window length on recognition accuracy. These results obtained with PDCW described Subsection 3 and 3.1 are summarized in Fig. 3, which indicate that best performance is achieved with window length of about 75 ms. In the experiments described below we Hamming windows of duration 75 ms with 37.5 ms between successive frames.

4. Experimental Results

In this section, we present experimental results for two different environmental conditions. In the first condition, we simulate different reverberant environments, where the target is masked by an interfering speaker. We used the Room Impulse Response (RIR) software [12] for simulating the effects of room reverberation. We assumed a room of dimensions $5 \times 4 \times 3$ m, a distance between the microphone and the speaker of 2 m, with the microphone located at the center of the room. We assumed that the target source is located along the perpendicular bisector of the line between two microphones, and that the masker is 45 degrees to one side. The target and noise signals are digitally added after simulating the reverberation effects. The two microphones are 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.

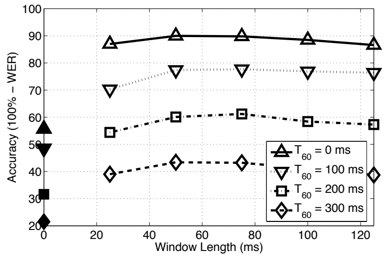

Figure 3: *The dependence of word recognition accuracy* ($100\% - WER$) *on the window length, using an SIR of 10 dB and various reverberation times. The filled symbols at 0 ms represent baseline results obtained with a single microphone.*

Fig. 4 compares word recognition accuracy for several of the algorithms discussed in the paper. ZCAE refers to the time-domain algorithm described in [7] with binary masking, as the better-performing continuous-masking does not work in environments with reverberation or more than one masking source. PD refers to the algorithm described in Secs. 2 and 3 of this paper with the 75-ms analysis window but without the gammatone frequency weighting, and PDCW refers to the complete algorithm including the gammatone channel weighting (CW) described in Sec. 3.1 with the 75-ms analysis window. To see the effects of the window length, we also present the PD results with the conventional 25-ms analysis window as well. As can be seen, the PDCW (and to a lesser extent the PD) algorithm provides lower WER than ZCAE, and the superiority of PDCW over ZCAE increases as the amount of reverberation increases.

In our second set of experiments, we still assume that the distance between the two microphones is the same, but 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. Fig. 4(d) illustrates these experimental results. Again we observe that PDCW (and to a lesser extent PD) provides much better performance than ZCAE for all conditions.

We also profiled the run times of implementations in C of the PDCW and ZCAE algorithms on two machines. The PDCW algorithms ran in only 9.03% of the time required to run the ZCAE algorithm on an 8-CPU Xeon E5450 3-GHz system, and in only 9.68% of the time to run the ZCAE algorithm on an embedded system with an ARM11 667-Mhz processor using a vector floating point unit. The major reason for the speedup is that in ZCAE the signal must be passed

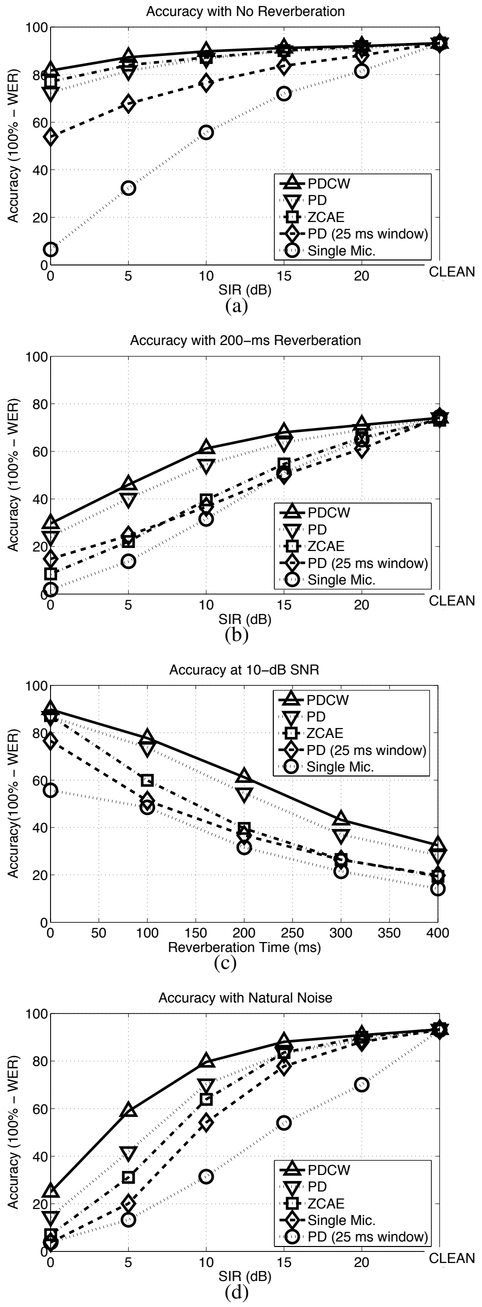

Figure 4: *Speech recognition accuracy using different algorithms (a) in the presence of an interfering speech source as a function of SNR in the absence of reverberation, (b,c) in the presence of reverberation and speech interference, as indicated, and (d) in the presence of natural real-world noise.*

through a bank of 40 filters while PDCW requires only two FFTs and one IFFT for each feature frame. A MATLAB version of PDCW with sample audio files is available at http://www.cs.cmu.edu/˜robust/archive/algorithms/PDCW_IS2009. The code in this directory was used to obtain the results described in this paper.

5. Conclusions

In this work, we present a speech separation algorithm, PDCW, based on ITD that is inferred from phase information. The algorithm uses gammatone weighting and longer analysis windows. This algorithm is quite computationally efficient and shows significant improvement in recognition accuracy under practical environmental conditions of noise and reverberation.

6. Acknowledgements

This study was supported by NSF Grant IIS-0420866. The authors are thankful to David-Huggins Daines, Hyung-Joon Lim, and Umpei Kurokawa for helpful discussions and for providing the noise data.

7. References

[1] R. Singh, R. M. Stern, and B. Raj, "Signal and feature compensation methods for robust speech recognition," in *Noise Reduction in Speech Applications*, G. M. Davis, Ed. CRC Press, 2002, pp. 219–244.

[2] R. Singh, B. Raj, and R. M. Stern, "Model compensation and matched condition methods for robust speech recognition," in *Noise Reduction in Speech Applications*, G. M. Davis, Ed. CRC Press, 2002, pp. 245–275.

[3] B. Raj, V. N. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 2, Apr. 1997, pp. 851–854.

[4] 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.

[5] B. Raj and R. M. Stern, "Missing-Feature Methods for Robust Automatic Speech Recognition," *Speech Communication*, vol. 22, no. 5, pp. 101–116, Sept. 2005.

[6] S. Srinivasan, M. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Comm.*, vol. 48, pp. 1486–1501, 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] 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 Arrys, 2008*, May. 2008.

[9] R. M. Stern and C. Trahiotis, "Models of binaural interaction," in *Hearing*, B. C. J. Moore, Ed. Academic Press, 2002, pp. 347–386.

[10] 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.

[11] B. .C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acustica - Acta Acustica*, vol. 82.

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