Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2008) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero-crossings

Hyung-Min Park [a,b,*], Richard M. Stern [b]

[a] Department of Electronic Engineering, Sogang University, Seoul 121-742, Republic of Korea
[b] Language Technologies Institute and Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Received 23 January 2008; received in revised form 22 May 2008; accepted 27 May 2008

Abstract

This paper describes an algorithm called zero-crossing-based amplitude estimation (ZCAE) that enhances speech by reconstructing the desired signal from a mixture of two signals using continuously-variable weighting factors, based on pre-processing that is motivated by the well-known ability of the human auditory system to resolve spatially-separated signals. Although most conventional methods of signal separation have been based on interaural time differences (ITDs) derived from cross-correlation information, the ZCAE approach provides sound segregation based on estimates of ITD from comparisons of zero-crossings [Kim, Y.-I., An, S.J., Kil, R.M., Park, H.-M., 2005. Sound segregation based on binaural zero-crossings. In: Proc. European Conf. on Speech Communication and Technology (INTERSPEECH-2005), Lisbon, Portugal, pp. 2325–2328]. These ITD estimates are used to determine the relative contribution of the desired source in a mixture and subsequently to reconstruct a closer approximation to the desired signal. The estimation of relative target intensity in a given time-frequency segment is accomplished by analytically deriving a monotonic function that maps the estimated ITD in each time-frequency segment to the putative relative intensity of each source. The ZCAE method is evaluated by comparing the sample standard deviation of ITD estimates derived using cross-correlation and using zero-crossing information, by comparing the speech recognition accuracy that is obtained by applying the proposed methods to speech in the presence of interfering speech sources, and by comparing recognition accuracy obtained using a continuous weighting versus a binary weighting of the target and masker. It is found that better results are obtained when ITDs are estimated using zero-crossing information rather than cross-correlation information, and when continuous weighting functions are used in place of binary weighting of the target and masker in each time-frequency segment.

Keywords: Automatic speech recognition; Noise robustness; Speech enhancement; Sound source segregation; Interaural time differences; Zero-crossings

1. Introduction

Noise robustness remains a very important issue in the field of automatic speech recognition (ASR). While ASR systems can achieve high recognition accuracy in controlled and noise-free acoustic environments, the performance of these systems is seriously degraded in more realistic environments which may be corrupted by noise and subjected to other types of distortion. This degradation is mainly due to differences between training and testing environments. Many algorithms have been proposed to compensate for these mismatches (e.g. Juang, 1991; Singh et al., 2002a,b). While these approaches can provide useful improvements in recognition accuracy under many circumstances, they frequently fail to obtain high recognition accuracy in dynamically changing environments with transient sources of disturbance, or in the presence of background speech or background music (e.g. Raj et al., 1997).

* Corresponding author. Address: Department of Electronic Engineering, Sogang University, Seoul 121-742, Republic of Korea. Tel.: +82 2 705 8916; fax: +82 2 706 4216.
E-mail address: hpark@sogang.ac.kr (H.-M. Park).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.05.012

In contrast, humans can understand speech even in the presence of competing speech or other noise sources (e.g. Assmann and Summerfield, 2004). This observation has motivated the development of many types of signal processing approaches based on aspects of human auditory perception (e.g. Hermansky, 1998; Wang and Brown, 2006). In his treatise on auditory scene analysis (ASA), Bregman (1990) identified various cues that are believed to be used by the human auditory system to segregate a target sound from interfering sources. These cues include fundamental frequency (F0), harmonics, onset or offset times, and source location. While most of the monaural cues such as F0 can be used as the basis for separation in computational ASA systems, their performance is critically dependent on characteristics of the input signal such as the presence or absence of voicing (Brown and Cooke, 1994). Localization cues which exploit small differences in the signals to the two ears, however, can identify the azimuth of sound sources regardless of signal content. Localization plays an important role in the human auditory system's ability to select a particular sound source and track the sound originating from that source. The primary acoustical cues for human sound localization are interaural time differences (ITDs) and interaural intensity differences (IIDs). ITDs serve as the localization cue primarily at frequencies below 1.5 kHz (Strutt, 1907), although the ITDs of the low-frequency envelopes of higher-frequency components of a sound can also be useful for sound localization (e.g. Henning, 1974; Nuetzel and Hafter, 1981). Information based on IIDs is primarily useful at higher frequencies.

Jeffress (1948) proposed a simple and intuitive mechanism that describes the estimation of ITDs based on interaural coincidences of hypothetical neural activity. Jeffress's hypothesis has motivated many computational models that describe and predict binaural processing (e.g. Braasch, 2005; Colburn and Kulkarni, 2005; Stern and Trahiotis, 1996). Most of these include a model of peripheral auditory processing which includes frequency analysis and subsequent nonlinear operations (e.g. Meddis and Hewitt, 1991), a mechanism for estimating the interaural cross-correlation function on a frequency-by-frequency basis, and a mechanism to disambiguate the temporal analysis, typically exploiting the IID of the signal or consistency over frequency. These models have been incorporated into several systems that perform ASR (e.g. Bodden, 1993; Bodden and Anderson, 1995; Tessier et al., 1999; Roman et al., 2003; Palomäki et al., 2004). Recently, Kim et al. estimated ITDs by measuring the time difference between zero-crossings, and showed that with this measure sound sources could be localized or segregated more robustly and accurately than using cross-correlation-based ITD estimation (Kim and Kil, 2004, 2005; Kim and Kil, 2007). In this paper we consider the use of zero-crossings of bandpass-filtered speech as an alternate way of estimating ITD information.

Once ITDs and IIDs are obtained at each frequency of interest, sound segregation or speech recognition can be performed on the basis of their values. Many contemporary algorithms based on ASA have used binary masks that specify which time-frequency components belong to a particular sound source on an "all-or-none" basis (e.g. Roman et al., 2003; Palomäki et al., 2004; Kim et al., 2005). This is clearly an oversimplified description of how sound sources are combined since each sound source contributes to the mixture to varying extents. For speech recognition applications, this oversimplification may be mitigated by the use of "soft masks" such as those proposed by Barker et al. (2000) and Morris et al. (2001). Over the years several research groups have described systems in which speech signals are reconstructed using spectro-temporal components that are implicitly weighted according to the likelihood that the observed ITD (and in some cases IID) would be appropriate for the desired source location (e.g. Bodden, 1993; Bodden and Anderson, 1995; Tessier et al., 1999). Recently, Srinivasan et al. (2004, 2006) introduced a way to obtain masks that provide continuously-variable estimates of the energy ratio between the desired components of a signal and the total signal. These ratio masks were obtained by first obtaining an empirical characterization of the dependence of ITDs and IIDs on energy ratios, and using this characterization to develop a function that estimates energy ratios from observed ITDs and IIDs.

In contrast to the empirically-derived ratio-masks that have been described by other groups, we derive in this paper an analytical relationship between the observed ITD at each frequency and the relative extent to which a desired sound source contributes to a particular time-frequency segment. From this information we develop an estimate of the desired signal in isolation by combining time-frequency segments of the total waveform in proportion to the estimated power of the target signal. As had been demonstrated previously by Srinivasan et al. (2004, 2006), the use of a continuously-variable weighting function rather than a binary mask provides smoother transitions between segments that contain greater and lesser components of the desired target signal, which improves speech recognition accuracy.

In addition to introducing an analytical approach to ratio-mask estimation, we also compare the performance of zero-crossing-based ITD estimation with cross-correlation-based ITD estimation. We will show that the zero-crossing method provides superior performance, both in terms of the sample standard deviation of the ITD estimates and in the resulting speech recognition accuracy.

The remainder of the paper is organized as follows: Section 2 describes the procedure for estimating the desired signal from observations by using continuously-variable masks derived from zero-crossing-based ITD estimation. In Section 3, the standard deviation of the resulting estimates obtained using this method is compared to the corresponding results obtained using cross-correlation-based methods. Comparisons of speech recognition accuracy using the algorithms considered are described in Section 4. Finally, our conclusions are summarized in Section 5.

Please cite this article in press as: Park, H.-M., Stern, R.M., Spatial separation of speech signals using amplitude estimation ..., Speech Comm. (2008), doi:10.1016/j.specom.2008.05.012

2. Algorithm description

Fig. 1 illustrates the overall procedure of the proposed algorithm, which we refer to as zero-crossing-based amplitude estimation (ZCAE). The signals which comprise the inputs to the two sensors of the system are mixtures of target and interfering sources from spatially-separated locations. The input signals are first subjected to frequency analysis, typically accomplished by passing each input through a bank of Gammatone filters which simulates cochlear filtering, and zero-crossings are detected from each filter output. The ITD is estimated from the time differences between the zero-crossings of the signals from the filter outputs at each center frequency, and these estimated ITDs are used to estimate the amplitude ratio which describes the corresponding relative contribution of the desired signal to each component. Finally, an estimate of the desired signal is obtained using a procedure based on the methods of (Weintraub, 1986; Brown and Cooke, 1994) that compensates for the phase distortion introduced by the Gammatone filters. Specifically, a time-reversed version of the amplitude-weighted signal from each frequency band is convolved with the corresponding Gammatone filter, and the results of each convolution are time-reversed again and summed across frequency to estimate the final output signal. These procedures are described in greater detail below.

2.1. ITD estimation based on zero-crossings

It is well-known that the auditory system develops estimates of ITD from the synchronous response of low-frequency auditory-nerve fibers to the fine structure of a sound source, and the exploitation of zero-crossing information is one possible way in which this processing could be achieved. The ITD in each frequency channel is estimated by first identifying the sample points at which the filtered input signal changed from a negative value to a positive value, or vice versa. The exact zero-crossing time (which generally falls between the sample times) is then estimated by linear interpolation based on the actual amplitudes of the signals at the sample points that straddle the zero-crossings. The corresponding zero-crossing time in the second sensor is assumed to be the zero-crossing that was closest in time, and the estimated ITD is defined to be the difference between these two zero-crossing times. Estimates of ITD that are greater than the time needed for a sound wave to travel from one sensor to the other are discarded.

It should be noted that while we use terms such as ITD from the binaural hearing literature, we make use of two sensors that are spaced far more closely than human ears to avoid the effects of spatial aliasing for frequencies up to half the sampling frequency, so the largest possible delay between the sensors is always less than half a period over all frequencies of interest.

2.2. Analytical derivation of the relationship between the ITD and the relative signal strength

Since processing algorithms using ASA are based on the development of a "mask" that describes which time-frequency components of an input signal are likely to be useful in describing the desired target component, the development of accurate masks is of critical importance to the system's performance. With some exceptions as noted above and especially the work of (Srinivasan et al., 2004, 2006), most previous work has been based on binary masks which select the time-frequency segments where the estimated target energy is greater than the estimated interference energy. Nevertheless, binary masks have the drawback that they cannot describe small differences in the extent to which a desired source contributes to a mixture.

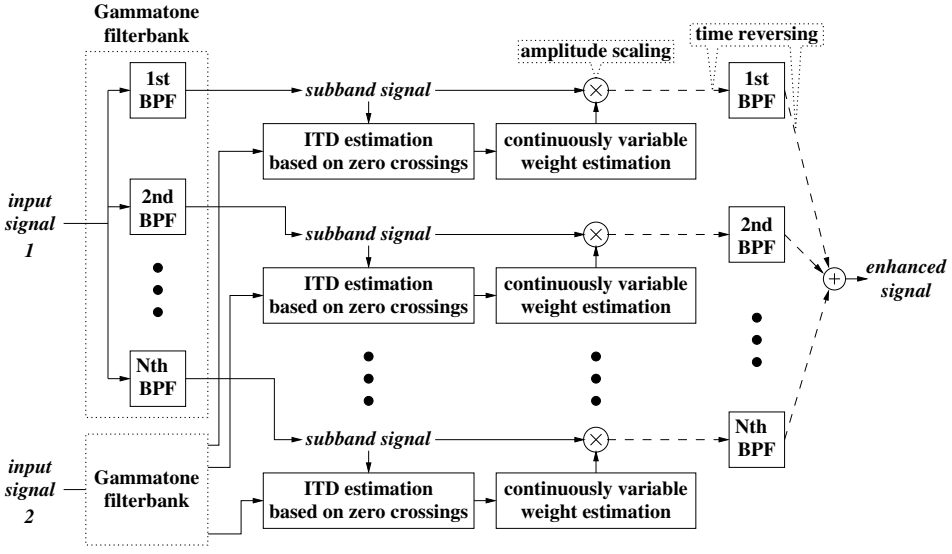


Fig. 1. Overall procedure of the ZCAE algorithm.

Please cite this article in press as: Park, H.-M., Stern, R.M., Spatial separation of speech signals using amplitude estimation ..., Speech Comm. (2008), doi:10.1016/j.specom.2008.05.012

In order to overcome this limitation of binary masks, we developed a reliable method that estimates continuously-variable masks by analytically deriving the relationship between the ITD and the relative contribution of the desired target. Because the two sensors are sufficiently close to avoid the effects of spatial aliasing, and because there is no object between them to cause acoustical shadowing, we assume zero IID between the signals arriving at these sensors. In addition, our derivation of the relationship between the ITD and the contribution of a desired signal is based on the assumption that there is only one target and one interfering source, and that their azimuth angles are known. Many researchers have developed methods that estimate the azimuth angles of sources composing observations (e.g. Stern and Colburn, 1978; Lyon, 1983; Lindemann, 1986; Bodden, 1993), and a zero-crossing-based method also can be used for estimating the azimuth and provided robust estimation especially for noisy mixtures (Kim and Kil, 2004; Kim and Kil, 2007). Using these methods, one may obtain azimuth angles of sources reliably, so the assumption on the azimuth angles is not critical.

We now describe a method by which the relative contribution of the target signal can be estimated from the estimates of ITD in each frequency band. Let us approximate the outputs of the Gammatone filters as pure tones, one from each source as follows:

$$x_1(t) = A_1\cos(\omega_1 t + \phi_1) + A_2\cos(\omega_2 t + \phi_2), \tag{1a}$$

$$x_2(t) = A_1\cos(\omega_1(t-d_1)+\phi_1) + A_2\cos(\omega_2(t-d_2)+\phi_2), \tag{1b}$$

where A_i, w_i, d_i, and ϕ_i denote amplitude, frequency, delay, and phase for the ith source, respectively. As noted in the previous paragraph, the amplitudes of the signals arriving at the two sensors are assumed to be equal.

Without loss of generality, we assume that a zero-crossing for x_1 occurs at time t_1, and that the nearest zero-crossing for x_2 occurs at $t_1+\tau$, producing an ITD of τ. We assume that $\omega_i(\tau - d_i)$ is small, which is especially valid at low frequencies, so $x_2(t_1+\tau)$ can be approximated by

$$x_2(t_1+\tau) \approx -A_1\sin(\omega_1 t_1+\phi_1)\cdot\omega_1(\tau-d_1) - A_2 \times \sin(\omega_2 t_1+\phi_2)\cdot\omega_2(\tau-d_2). \tag{2}$$

Since $x_2(t_1+\tau)=0$,

$$\tau(A_1\omega_1\sin(\omega_1 t_1+\phi_1)+A_2\omega_2\sin(\omega_2 t_1+\phi_2)) \approx A_1\omega_1\sin(\omega_1 t_1+\phi_1)\cdot d_1 + A_2\omega_2\sin(\omega_2 t_1+\phi_2)\cdot d_2. \tag{3}$$

Since $x_1(t_1) = A_1\cos(\omega_1 t_1+\phi_1)+A_2\cos(\omega_2 t_1+\phi_2)=0$ and τ is obtained from the nearest zero-crossing point,

$$\tau \approx \begin{cases} \dfrac{\omega_1\sqrt{A_1^2-A_2^2\cos^2(\omega_2 t_1+\phi_2)}\cdot d_1 + A_2\omega_2|\sin(\omega_2 t_1+\phi_2)|\cdot d_2}{\omega_1\sqrt{A_1^2-A_2^2\cos^2(\omega_2 t_1+\phi_2)}+A_2\omega_2|\sin(\omega_2 t_1+\phi_2)|} & \text{if } A_1 \geqslant A_2, \\ \dfrac{A_1\omega_1|\sin(\omega_1 t_1+\phi_1)|\cdot d_1 + \omega_2\sqrt{A_2^2-A_1^2\cos^2(\omega_1 t_1+\phi_1)}\cdot d_2}{A_1\omega_1|\sin(\omega_1 t_1+\phi_1)|+\omega_2\sqrt{A_2^2-A_1^2\cos^2(\omega_1 t_1+\phi_1)}} & \text{otherwise.} \end{cases} \tag{4}$$

We assume that the frequencies ω_i are distributed over a narrow band, so one of them is approximately same as the other. In addition, by assuming that the phases ϕ_i are uniformly distributed over the interval $(-\pi,\pi)$, we may consequently assume that $\psi_i = \omega_i t_1 + \phi_i$ are also uniformly distributed over the same interval. Since Eq. (4) is periodic with period π, one may obtain the mean of the estimated ITD, $\bar{\tau}$, which can be approximated by

$$\bar{\tau} \approx g_1'(A_1,A_2)\cdot d_1 + g_2'(A_1,A_2)\cdot d_2, \tag{5}$$

where

$$g_1'(A_1,A_2) = \begin{cases} \int_0^\pi \dfrac{\sqrt{A_1^2-A_2^2\cos^2(\psi_2)}}{\sqrt{A_1^2-A_2^2\cos^2(\psi_2)}+A_2\sin(\psi_2)}\dfrac{\mathrm{d}\psi_2}{\pi} & \text{if } A_1 > A_2, \\ \int_0^\pi \dfrac{\sqrt{1-\cos^2(\psi_2)}}{\sqrt{1-\cos^2(\psi_2)}+\sin(\psi_2)}\dfrac{d\psi_2}{\pi} & \text{if } A_1 = A_2, \\ \int_0^\pi \dfrac{A_1\sin(\psi_1)}{A_1\sin(\psi_1)+\sqrt{A_2^2-A_1^2\cos^2(\psi_1)}}\dfrac{\mathrm{d}\psi_1}{\pi} & \text{otherwise,} \end{cases} \tag{6}$$

and

$$g_2'(A_1,A_2) = \begin{cases} \int_0^\pi \dfrac{A_2\sin(\psi_2)}{\sqrt{A_1^2-A_2^2\cos^2(\psi_2)}+A_2\sin(\psi_2)}\dfrac{d\psi_2}{\pi} & \text{if } A_1 > A_2, \\ \int_0^\pi \dfrac{\sin(\psi_2)}{\sqrt{1-\cos^2(\psi_2)}+\sin(\psi_2)}\dfrac{d\psi_2}{\pi} & \text{if } A_1 = A_2, \\ \int_0^\pi \dfrac{\sqrt{A_2^2-A_1^2\cos^2(\psi_1)}}{A_1\sin(\psi_1)+\sqrt{A_2^2-A_1^2\cos^2(\psi_1)}}\dfrac{\mathrm{d}\psi_1}{\pi} & \text{otherwise.} \end{cases} \tag{7}$$

Using the MATLAB `symbolic integration` function we obtain

$$\bar{\tau} \approx g(A_1,A_2)\cdot d_1 + (1-g(A_1,A_2))\cdot d_2, \tag{8}$$

where

$$g(A_1,A_2) = \begin{cases} \dfrac{\pi(A_1^2-\frac{A_2^2}{2})-A_1^2\arctan\left(\frac{A_2}{\sqrt{A_1^2-A_2^2}}\right)-A_2\sqrt{A_1^2-A_2^2}}{\pi(A_1^2-A_2^2)} & \text{if } A_1 > A_2, \\ \frac{1}{2} & \text{if } A_1 = A_2, \\ \dfrac{-\frac{\pi}{2}A_1^2+A_2^2\arctan\left(\frac{A_1}{\sqrt{A_2^2-A_1^2}}\right)+A_1\sqrt{A_2^2-A_1^2}}{\pi(A_2^2-A_1^2)} & \text{otherwise.} \end{cases} \tag{9}$$

By introducing the signal-to-interference ratio (SIR) in decibels (dBs) according to the relation

$$\mathrm{SIR} = 20\log_{10}\frac{A_1}{A_2}\ (\mathrm{dB}), \tag{10}$$

where A_1 and A_2 denote the amplitudes for the target and interference signals, respectively,

$$\bar{\tau} \approx g(\mathrm{SIR})\cdot d_1 + (1-g(\mathrm{SIR}))\cdot d_2, \tag{11}$$

where

$$g(\mathrm{SIR}) = \begin{cases} \dfrac{\pi(10^{\mathrm{SIR}/10}-\frac{1}{2}) - 10^{\mathrm{SIR}/10}\arctan\left(\frac{1}{\sqrt{10^{\mathrm{SIR}/10}-1}}\right) - \sqrt{10^{\mathrm{SIR}/10}-1}}{\pi(10^{\mathrm{SIR}/10}-1)} & \text{if SIR} > 0\ \text{dB}, \\ \frac{1}{2} & \text{if SIR} = 0\ \text{dB}, \\ \dfrac{-\frac{\pi}{2} + 10^{-\mathrm{SIR}/10}\arctan\left(\frac{1}{\sqrt{10^{-\mathrm{SIR}10}-1}}\right) + \sqrt{10^{-\mathrm{SIR}/10}-1}}{\pi(10^{-\mathrm{SIR}/10}-1)} & \text{otherwise}. \end{cases} \quad (12)$$

Fig. 2 displays the mixing function $g(\mathrm{SIR})$. Using Eqs. (11) and (12), one can easily relate the estimated ITDs to the SIRs in a mixture. Note that Eq. (11) describes a monotonic function that is suitable for one-to-one mapping and that does not depend on any frequency-specific parameters, so the same function can be used in all frequency bands.

Since most systems that perform binaural analysis estimate ITDs using cross-correlation rather than zero-crossing methods, we also consider the corresponding relationship between the measured ITD in a frequency band and the relative signal strength of the desired target signal based on cross-correlation analysis. We show in the Appendix that the relationship between the ITD estimated using cross-correlation methods and signal amplitudes A_i of the target and interfering source can be expressed as the relationship:

$$\begin{aligned} \tau_{cc}\big[&A_1^2(\sin(\omega_1 T)\omega_1\cos(2\phi_1) + \omega_1^2 T) \\ &+ \frac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T/2)(\omega_1^2+\omega_2^2)\cos(\phi_1+\phi_2) \\ &+ \frac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T/2)(\omega_1^2+\omega_2^2)\cos(\phi_1-\phi_2) \\ &+A_2^2(\sin(\omega_2 T)\omega_2\cos(2\phi_2)+\omega_2^2 T)\big] \\ \approx\ & d_1\Big[A_1^2(\sin(\omega_1 T)\omega_1\cos(2\phi_1)+\omega_1^2 T) \\ &+ \frac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T/2)\omega_1^2\cos(\phi_1+\phi_2) \\ &+ \frac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T/2)\omega_1^2\cos(\phi_1-\phi_2)\Big] \\ &+ d_2\Big[\frac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T/2)\omega_2^2\cos(\phi_1+\phi_2) \\ &+ \frac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T/2)\omega_2^2\cos(\phi_1-\phi_2) \\ &+A_2^2(\sin(\omega_2 T)\omega_2\cos(2\phi_2)+\omega_2^2 T)\big] \\ &-\big[A_1^2\sin(\omega_1 T)\sin(2\phi_1) + 2A_1A_2\sin((\omega_1+\omega_2)T/2) \\ &\times\sin(\phi_1+\phi_2) + 2A_1A_2\sin((\omega_1-\omega_2)T/2)\sin(\phi_1-\phi_2) \\ &+A_2^2\sin(\omega_2 T)\sin(2\phi_2)\big]. \end{aligned} \quad (13)$$

The value of τ_{cc} is, of course, easily obtained by dividing both sides of Eq. (13) by the expression on the left side of the equation in brackets. In deriving this equation we assume that $\omega_1 \neq \omega_2$ because the two frequencies are regarded as random variables that are uniformly distributed within the bandwidth of an analysis band.

Many of the terms of the complicated Eq. (13) above serve primarily to reflect the effects of the interaction between the period of the sinusoidal waveforms and the finite-duration of the observation window T. As the observation duration T increases, these "fringe effects" will become decreasingly important, and for very large T Eq. (13) converges to

$$\tau_{cc} = \frac{10^{\mathrm{SIR}/10}\omega_1^2 d_1 + \omega_2^2 d_2}{10^{\mathrm{SIR}/10}\omega_1^2 + \omega_2^2} \quad (14)$$

with SIR defined as in Eq. (10).

2.3. Estimation of a desired signal using continuously-variable weighting factors

Given the relationship between the ITD and the relative contribution of a source as described by Eq. (11), the ITDs estimated at each zero-crossing point are used to estimate weighting factors according to the extent to which the desired target signal is assumed to be present in each time-frequency region of the mixture. This processing results in a series of estimated weights for the desired signal corresponding to each zero-crossing in each frequency band. Weighting factors for all samples in time are obtained by developing a piecewise-linear amplitude modulation function that passes through the values of the weighting coefficients for the desired signal that are calculated at each zero-crossing point. An estimate of the desired signal in a frequency band is then obtained by multiplying the input signal after band-pass filtering by the cor-

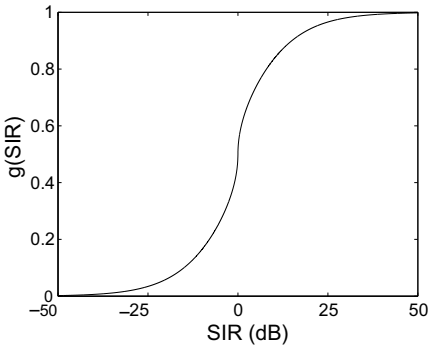


Fig. 2. The function $g(\mathrm{SIR})$ described by Eq. (12).

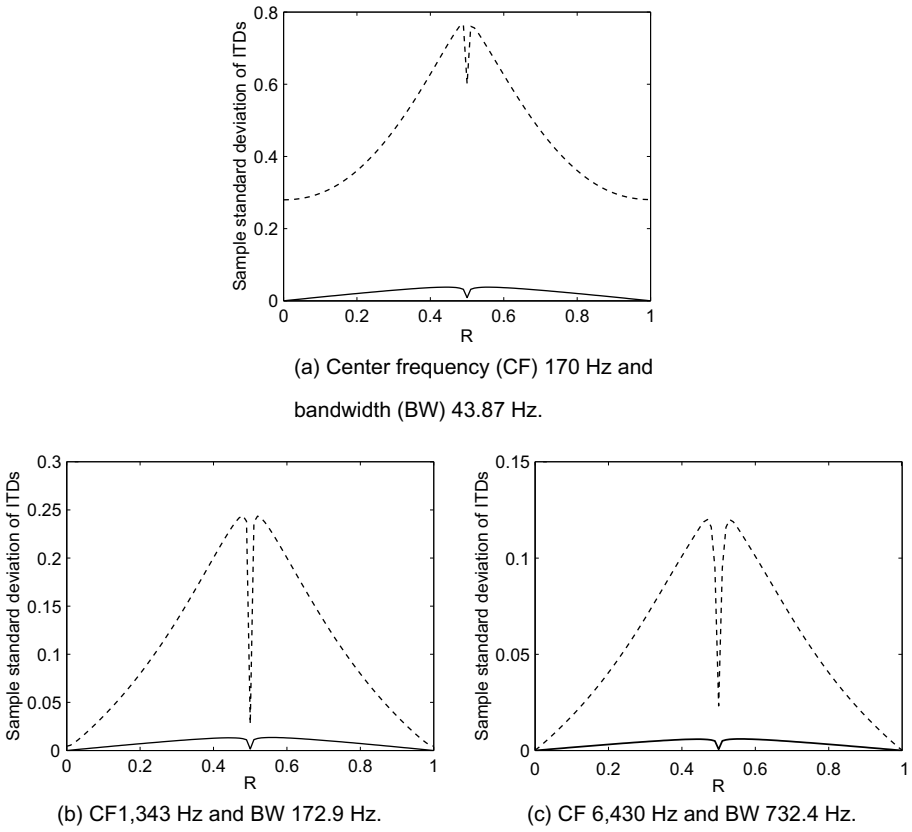


Fig. 3. Sample standard deviation of estimated ITDs as a function of the relative signal strength R at three frequencies, 170, 1343, and 6340 Hz. The solid and dashed lines correspond to results obtained using the zero-crossing-based and cross-correlation-based methods, respectively. The vertical axis is normalized by dividing by d_2.

responding amplitude modulation function. Finally, the enhanced full-band signal is recovered by summing the subband signals across all frequencies after compensating for the phase distortion introduced by the Gammatone filters as described above.

3. Comparison of standard deviations of ITDs derived from zero-crossings and cross-correlation

In this section we compare the reliability of estimates of ITD obtained using the zero-crossing and cross-correlation methods in terms of the sample standard deviation of the ITDs obtained by each of two analysis methods. Defining the relative signal strength R to be

$$R = \frac{A_1}{A_1 + A_2} = \frac{1}{1 + 10^{-\mathrm{SIR}/20}}, \tag{15}$$

we obtained 100,000 samples of ITDs using Eqs. (4) and (13) using randomly-generated values of the phases ϕ_i and frequencies ω_i and various values of R. For each sample, ϕ_i and ω_i were uniformly distributed over the intervals $(-\pi, \pi)$ and $(\omega_{\mathrm{cf}} - \mathrm{BW}/2, \omega_{\mathrm{cf}} + \mathrm{BW}/2)$, respectively, where ω_{cf} and BW are the center frequencies and bandwidths of the frequency channels. We obtained the sample values with parameter values $d_1 = 0$, $d_2 = 1/16,000$ s.

Fig. 3 shows the sample standard deviation of estimates of the ITD as a function of the relative signal strength R for three different frequency bands. In this simulation, the frame length was $T = 25.6$ ms, which corresponds to the frame length used for analysis in the speech recognition experiments that are described in the following section. Since the frequency of zero-crossings is approximately twice the center frequency, the sample standard deviation for the zero-crossing method was computed from averaged ITDs whose number approximately corresponded to the number of zero-crossings in a frame. In each of the three frequency bands, the variability of ITD estimates obtained using the zero-crossing method is much less than that of estimates obtained using cross-correlation, which suggests that the zero-crossing method is likely to provide more reliable estimates of ITD than the cross-correlation method.

It should be noted that the large differences between the sample standard deviations observed using the zero-crossing and cross-correlation methods are primarily a consequence of the short-duration of the analysis frames and the consequent significance of the interaction between the cross-correlation computation and the frame boundaries, as specified in Eq. (13). If the calculations were carried out using a longer frame duration such that Eq. (14) were valid, the difference between standard deviations observed

based on zero-crossing-based versus cross-correlation-based estimates of ITD would be small.

4. Experimental evaluation on speech recognition

As noted in the Introduction, speech-on-speech interference has long been considered to be one of the most challenging problems in ASR, both because the interfering speech tends to be confused with the target speech and because the non-stationary nature of the masking signal renders most conventional noise-compensation methods ineffective (e.g. Singh et al., 2002a,b). In this section we compare the recognition accuracy obtained using the ZCAE method with correlation-based approaches and baseline processing.

4.1. Experimental design and signal generation

Recognition experiments were conducted using the DARPA Resource Management (RM1) database (Price et al., 1988) and the CMU SPHINX-III speech recognition system. The recognition system is based on fully-continuous hidden Markov models, which are trained on 2880 RM1 sentences recorded in a quiet environment. The test set consists of 600 RM1 sentences. Speech recognition is based on the observed values of 13th-order mel-frequency cepstral coefficients with a frame size of 25.6 ms and a frame rate of 10 ms developed in the conventional fashion.

Each test utterance was corrupted by a second (interfering) speech signal which had the same energy as the test utterance, producing a nominal SIR of 0 dB. To simulate the measurement of signals that would be obtained by microphones in close proximity, the target and interfering speech were combined with different simulated delays from sensor to sensor, corresponding to different putative arrival angles for the target and interfering source. Because the original sampling period is too coarse for this purpose, the original target and interfering speech signals were upsampled by a factor of 4 and added together at the 64 kHz sampling rate, after appropriate ITDs were inserted to simulate the different azimuths of the target and interfering signal. These delays were selected independently and randomly in the range of −3 to 3 samples at 64 kHz, except that delays for the target and interference were forced to be different from each other under the assumption that the sources are spatially-separated. As a result, the net difference in ITDs between the target and the interference ranged from 1 to 6 samples or 15.6 to 93.8 μs. (For microphones spaced by 21 mm, this would correspond to differences in azimuth of between about 14.9° and 100.7°.) The target signal was always assumed to be the component with the more positive delay. After combining the target and interfering speech, the resulting signals were downsampled back to 16 kHz. Because we used a nominal spacing of about 21 mm between the sensors to avoid spatial aliasing at 8 kHz, the largest delay between observations from sensor to sensor would be 60.9 μs, which is slightly smaller than the sampling period at 16 kHz.

4.2. Signal separation using continuously-variable masks

To obtain subband signals from each sensor signal according to the ZCAE algorithm, we used a 40-channel bank of Gammatone filters with center frequencies spaced linearly in equivalent rectangular bandwidth (ERB) from 170 Hz to 6430 Hz (O'Mard, 2000). The ITDs of the subband signal at each frequency band are converted into estimates of the relative signal strength of the target according to Eqs. (11) and (12) under the assumption that the actual delays for the desired target and interfering speech are known *a priori*.

For comparison, we also present recognition accuracies based on ITD estimation using cross-correlation. In order to obtain continuously-variable masks from the cross-correlation-based ITDs, we need to derive the relationship between the ITDs and the relative contribution of a source in a mixture. To simplify the derivation, we used the simpler expression in Eq. (14) for the ITD at the maximum of cross-correlation, τ_{cc}.

Since Eq. (14) is the exactly same as the derivation by Roman et al. (2003), we adopt their approximation for the mean of τ_{cc} given by

$$\bar{\tau}_{cc} = \frac{d_1 + d_2}{2} + \frac{1}{w_{cf}} \left\{ \arctan \left[-\frac{(10^{SIR/10} - 1)}{(10^{SIR/10} + 1)} \tan\beta \right] + k\pi \right\}, \quad k \in \{0, \pm 1\}, \tag{16}$$

where $\beta = w_{cf} \cdot (d_2 - d_1)/2 \in [0, \pi]$. If $\beta \leqslant \pi/2$, $k = 0$. Otherwise, $k = 1$ when SIR < 0 dB and $k = -1$ when SIR > 0 dB (Roman et al., 2003). The time lag corresponding to the maximum of the cross-correlation function was estimated by differentiating a 20th-order polynomial approximation to the cross-correlation function in a region around the observed discrete-time maximum and finding the root of the derivative closest to the discrete-time maximum using Newtonian iteration.

4.3. Signal separation using binary masks

In addition to the recognition results that were obtained using the continuously-variable masks as described above, we also evaluated speech recognition accuracy using binary masks. These binary masks were determined by establishing a hard threshold between the ITD representing the azimuth of the target and the ITD representing the azimuth of the interfering source in place of the functions that related the estimated ITDs and the relative contribution of a source according to Eqs. (11) or (16). Specifically, the value of the binary mask for the ith channel and the jth zero-crossing or the jth frame (depending on whether we are using zero-crossing-based or cross-correlation-based ITD extraction) is

$$m(i,j) = \begin{cases} 1, & \text{if } \tau_{\text{est}} \geqslant \frac{d_1+d_2}{2}, \\ 0, & \text{otherwise,} \end{cases} \quad (17)$$

where τ_{est} denotes the estimated ITD regardless of which method was used to obtain it. Recall that $d_1 > d_2$ since the target signal is assumed to have the more positive delay than the interfering signal.

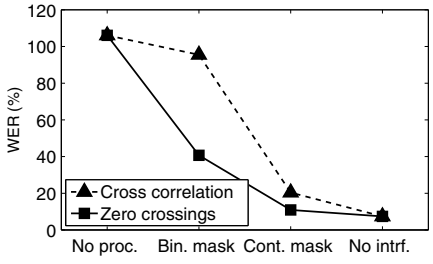
4.4. Incorporation of missing-feature reconstruction for signal separation using binary masks

Many conventional methods have employed missing-feature techniques to achieve noise robustness of ASR system through the use of the binary masks. Typical missing-feature techniques, as reviewed by Raj and Stern (2005), either attempt to obtain optimal decisions while ignoring time-frequency regions that are considered to be unreliable, or they attempt to "fill in" the values of those unreliable features. We employed the cluster-based method of restoring missing-features. Briefly, in cluster-based missing-feature restoration it is assumed that the various spectral profiles that represent speech sounds can be clustered into a set of prototypical spectra. For each input frame we first estimate the cluster to which the incoming spectral features are most likely to belong from the observed spectral components that are believed to be "present" (or reliable) and that are considered to belong to the target signal based on zero-crossing or cross-correlation information, as described above. The remaining "missing" spectral components are obtained using bounded estimation based on the observed values of the components that are considered to be reliable, and based on the knowledge of the spectral cluster to which the incoming speech is assumed to belong. A detailed description of this approach is available in (Raj et al., 2004). When missing-feature reconstruction is used, it is no longer feasible to use the procedures described by Weintraub (1986) and Brown and Cooke (1994) to reconstruct an estimate of the separated desired signal. Instead of reconstructing the desired signal we obtained spectral features by directly computing frame energies after passing the input signals through the same bank of Gammatone filters used in the earlier experiments. These spectral features were transformed into cepstral features in the usual fashion to train and test a second ASR system. The training and test utterances were the same as described in Section 4.1.

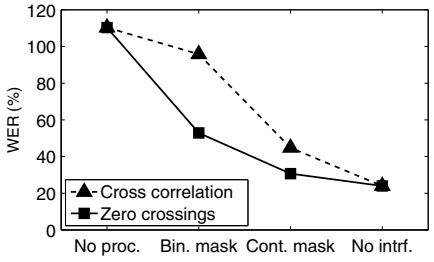
4.5. Experimental results

Fig. 4a presents word error rates (WERs) calculated according to the standard NIST metric for various mixtures of the target and interfering speech combined as described in the previous section. WERs for "clean" target speech presented in isolation are also provided as an upper bound on recognition accuracy, and results for the combined signals without any processing for enhancement are also included as baselines to assess the effectiveness of the processing. As described previously, the signal-to-interference ratio (SIR) of the desired speech source compared to the interfering speech is nominally 0 dB.

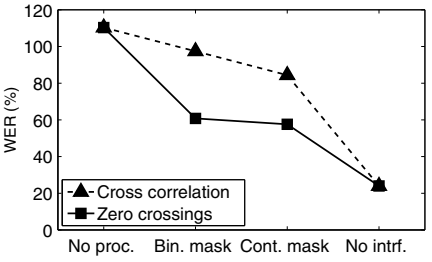
To obtain a crude characterization of the robustness of the methods considered, we also obtained WERs for the same signals in the presence of white Gaussian noise at



(a) No noise added.



(b) Identical white Gaussian noise.



(c) Independent white Gaussian noise.

Fig. 4. Comparison of the WERs obtained using continuously-variable masks versus binary masks and ITD estimation based on zero-crossings versus cross-correlation. In each frame results are shown from left to right with no-processing for enhancement, with binary masks, with continuously-variable masks, and in the absence of an interfering source. Processing using ITDs estimated using zero-crossings is represented by the solid curves, and processing using ITDs based on cross-correlation analysis is represented by the dashed curves. The three frames compare results obtained using (a) no additional background noise, (b) identical white Gaussian background noise, and (c) statistically-independent white Gaussian background noise.

Please cite this article in press as: Park, H.-M., Stern, R.M., Spatial separation of speech signals using amplitude estimation ..., Speech Comm. (2008), doi:10.1016/j.specom.2008.05.012

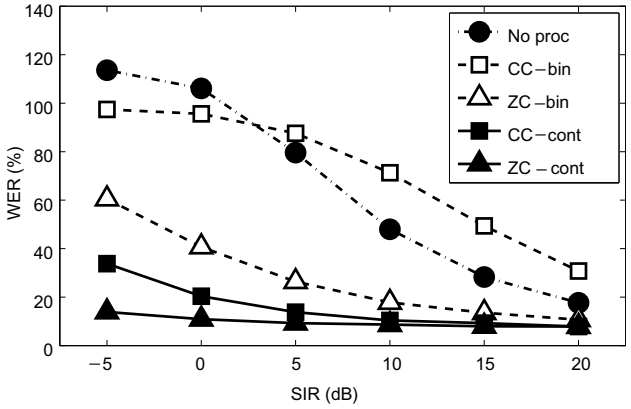


Fig. 5. Dependence of WER on the signal-to-interference ratio (SIR). Observed WER using no-processing (circles), cross-correlation-based ITD extraction (squares), and zero-crossing-based ITD extraction (triangles). In the latter two cases, filled symbols depict results obtained using continually-variable weighting and open symbols depict results obtained using binary weighting.

an SNR of 20 dB. Because there are two sensor signals, this noise was added to the signals in two different ways. Specifically, the two speech signals were corrupted by identical noise (which could be caused by an additional non-speech source of interference with zero ITD), and they were also corrupted by statistically-independent noise (which could be caused by sensor or measurement noise). Results with identical noise, and with statistically-independent noise, are shown in Fig. 4b and c, respectively.

It is seen that the use of the continuously-variable masks (unsurprisingly) provides much greater recognition accuracy than the use of binary masks, as had been observed in previous studies (e.g. Srinivasan et al., 2004, 2006). It is also evident, though, that signal separation that is accomplished using ITDs that are estimated using zero-crossing information is more effective in reducing WER than signal separation that is based on cross-correlation information. The WERs are (also unsurprisingly) affected adversely by the addition of noise in all cases.

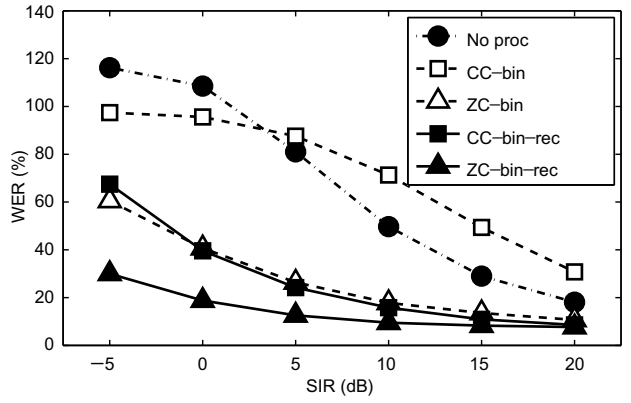


Fig. 6. Comparison of WERs obtained using binary weighting with (filled symbols) and without (open symbols) missing-feature reconstruction. The weights are based on ITD analysis using cross-correlation analysis (squares) and zero-crossing analysis (triangles). See text for additional details.

Fig. 5 depicts speech recognition accuracy as a function of the signal-to-interference ratio (SIR) for the same signal processing procedures that were used for the data in Fig. 4a. It can be seen that the hierarchy of WERs observed in Fig. 4a at 0 dB holds true over a wide range of values of SIR, except that processing using binary masks and ITDs derived from the cross-correlation function produces worse performance than the baseline (no-processing) condition at higher SIRs. Large WERs are observed for the combination of binary masks and cross-correlation-based ITD estimation because signals in frames that are dominated by the interfering signal are removed completely. This typically causes abrupt discontinuities in the spectrogram, which are abnormal for speech.

The best performance is observed when the ZCAE method is used, which employs continuously-variable weighting factors estimated from zero-crossing-based ITDs. The resulting WERs obtained with this approach are generally close to the recognition accuracy observed in the absence of an interfering source.

Fig. 6 describes results obtained using the binary masks described in previous sections combined with the cluster-based missing-feature reconstruction techniques developed by Raj et al. (2004). As noted above, the speech recognition system used in conjunction with missing-feature reconstruction was slightly different from the one that had been used to obtain the previous results. Specifically, cepstral features in the present system were developed directly from the log-spectral values, rather than from a reconstructed speech waveform. The results in Fig. 6 are presented to facilitate comparisons across conditions, and they are the same data that had been presented for the results obtained using binary masks in Fig. 5. By comparing the results presented in Figs. 5 and 6 it can be seen that the use of missing-feature reconstruction can reduce the WERs obtained using binary masks very substantially, but not to the extent enjoyed by the use of the continuous ratio-masks (except for the highest SIRs when the effect of the interfering speech source is minimal).

5. Conclusions

We have described the zero-crossing-based amplitude estimation (ZCAE) method that estimates continuously-variable weighting factors for enhancing the desired signal in the presence of a single interfering source, and we have evaluated it in the context of speech recognition. The use of zero-crossings as the basis for signal separation based on ITD is shown to be superior to the historically more popular use of cross-correlation information to estimate ITDs.

A major contribution of this work is the analytical development of a relationship between estimates of ITD based on zero-crossings in a particular frequency band and the implied ratio of desired signal energy to total

energy in that band. This result complements a similar derivation by Srinivasan et al. (2004, 2006) for ITDs that are estimated from the cross-correlation function of the signals to the two sensors. In contrast to empirically-derived transfer functions that fit empirical data relating ITDs to energy ratios using a specific configuration of sources and sensors, the ZCAE method provides a simple monotonic function which can be used with equal validity in all frequency bands and for any source and signal configuration.

The reliability of continuously-variable weighting factors estimated using the ZCAE method has been assessed both by considering the sample standard deviation of the estimated ITDs and the resulting speech recognition accuracy. In both cases, the use of zero-crossings proved to be superior to the use of cross-correlation for the estimation of ITDs, and the use of continuously-variable weights remains superior to the use of binary masks. A limitation of this approach is that so far it has been applied only to the case of a single interfering source in the absence of reverberation. We are encouraged by these results and are working to extend them.

Acknowledgements

This work was supported by the Information and Telecommunication National Scholarship Program sponsored by the Institute of Information Technology Assessment, Korea, by the Sogang University Foundation Research Grants, and by the National Science Foundation (Grant IIS-0420866).

Appendix A. Derivation of the ITDs based on cross-correlation as a function of source amplitudes

In order to derive the ITDs based on cross-correlation as a function of source amplitudes, we follow almost the same derivation as in the case of the zero-crossing-based method. As before, we start from mixtures given by Eq. (1).

For finite-duration signals in a frame of duration T, the cross-correlation function with time lag τ is expressed by

$$c(\tau) = \frac{1}{T}\int_{t=-T/2}^{T/2} x_1(t)x_2(t+\tau)\,\mathrm{d}t. \tag{A.1}$$

For the values of $x_1(t)$ and $x_2(t)$ considered, the integrand in Eq. (A.1) is

$$\begin{aligned}
&x_1(t)x_2(t+\tau)\\
&= \frac{1}{2}\big[A_1^2(\cos(2\omega_1 t+\omega_1(\tau-d_1)+2\phi_1)+\cos(\omega_1(\tau-d_1)))\\
&\quad + A_1A_2(\cos(\omega_1 t+\omega_2(t+\tau-d_2)+\phi_1+\phi_2)\\
&\quad + \cos(\omega_1 t-\omega_2(t+\tau-d_2)+\phi_1-\phi_2)\\
&\quad + \cos(\omega_1(t+\tau-d_1)+\omega_2 t+\phi_1+\phi_2)\\
&\quad + \cos(\omega_1(t+\tau-d_1)-\omega_2 t+\phi_1-\phi_2))\\
&\quad + A_2^2(\cos(2\omega_2 t+\omega_2(\tau-d_2)+2\phi_2)\\
&\quad + \cos(\omega_2(\tau-d_2)))].
\end{aligned} \tag{A.2}$$

Hence, for $\omega_1 \neq \omega_2$, the derivative of the cross-correlation function $c(\tau)$ can defined as follows:

$$\begin{aligned}
\frac{\mathrm{d}c(\tau)}{\mathrm{d}\tau} = -\frac{1}{2T}\Big[&A_1^2(\sin(\omega_1 T)\sin(\omega_1(\tau-d_1)+2\phi_1)\\
&+\omega_1 T\sin(\omega_1(\tau-d_1)))\\
&+A_1A_2\Big(\frac{2\omega_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T/2)\\
&\times\sin(\omega_2(\tau-d_2)+\phi_1+\phi_2)\\
&+\frac{2\omega_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T/2)\\
&\times\sin(\omega_2(\tau-d_2)-\phi_1+\phi_2)\\
&+\frac{2\omega_1}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T/2)\\
&\times\sin(\omega_1(\tau-d_1)+\phi_1+\phi_2)\\
&+\frac{2\omega_1}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T/2)\\
&\times\sin(\omega_1(\tau-d_1)+\phi_1-\phi_2))\\
&+A_2^2(\sin(\omega_2 T)\sin(\omega_2(\tau-d_2)+2\phi_2)\\
&+\omega_2 T\sin(\omega_2(\tau-d_2)))\Big].
\end{aligned} \tag{A.3}$$

The time lag τ_{cc} at the maximum of the cross-correlation, which corresponds to the estimated ITD, is obtained by setting the derivative of $c(\tau)$ in the equation above to zero. Using the approximation of small $\omega_i(\tau-d_i)$, we obtain the expression below from which τ_{cc} is easily obtained:

$$\begin{aligned}
&\tau_{cc}\Big[A_1^2(\sin(\omega_1 T)\omega_1\cos(2\phi_1)+\omega_1^2 T)\\
&\quad+\frac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T/2)(\omega_1^2+\omega_2^2)\cos(\phi_1+\phi_2)\\
&\quad+\frac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T/2)(\omega_1^2+\omega_2^2)\cos(\phi_1-\phi_2)\\
&\quad+A_2^2(\sin(\omega_2 T)\omega_2\cos(2\phi_2)+\omega_2^2 T)\Big]\\
&\approx d_1\big[A_1^2(\sin(\omega_1 T)\omega_1\cos(2\phi_1)+\omega_1^2 T)\\
&\quad+\frac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T/2)\omega_1^2\cos(\phi_1+\phi_2)\\
&\quad+\frac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T/2)\omega_1^2\cos(\phi_1-\phi_2)\big]\\
&\quad+d_2\Big[\frac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T/2)\omega_2^2\cos(\phi_1+\phi_2)\\
&\quad+\frac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T/2)\omega_2^2\cos(\phi_1-\phi_2)\\
&\quad+A_2^2(\sin(\omega_2 T)\omega_2\cos(2\phi_2)+\omega_2^2 T)\Big]\\
&\quad-\Big[A_1^2\sin(\omega_1 T)\sin(2\phi_1)+2A_1A_2\sin((\omega_1+\omega_2)T/2)\\
&\quad\times\sin(\phi_1+\phi_2)+2A_1A_2\sin((\omega_1-\omega_2)T/2)\sin(\phi_1-\phi_2)\\
&\quad+A_2^2\sin(\omega_2 T)\sin(2\phi_2)\Big].
\end{aligned} \tag{A.4}$$

References

Assmann, P., Summerfield, Q., 2004. The perception of speech under adverse conditions. In: Greenberg, S., Ainsworth, W., Popper, A.N., Fay, R.R. (Eds.), Speech Processing in the Auditory System, . In: Springer Handbook of Auditory Research, Vol. 18. Springer-Verlag, pp. 231–308, Chapter 5.

Barker, J.P., Josifovski, L., Cooke, M.P., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP-2000), Beijing, China, pp. 373–376.

Bodden, M., 1993. Modelling human sound-source localization and the cocktail party effect. Acta Acust. 1, 43–55.

Bodden, M., Anderson, T.R., 1995. A binaural selectivity model for speech recognition. In: Proc. European Conf. on Speech Communication and Technology (EUROSPEECH-1995), Madrid, Spain, pp. 127–130.

Braasch, J., 2005. Modelling of binaural hearing. In: Blauert, J. (Ed.), Communication Acoustics. Springer-Verlag, Berlin, Germany, pp. 75–108, Chapter 4.

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambridge, MA.

Brown, G.J., Cooke, M.P., 1994. Computational auditory scene analysis. Comput. Speech Lang. 8, 297–336.

Colburn, H.S., Kulkarni, A., 2005. Models of sound localization. In: Fay, R., Popper, T. (Eds.), Sound Source Localization, . In: Springer Handbook of Auditory Research, Vol. 6. Springer-Verlag, pp. 272–316, Chapter 8.

Henning, G.B., 1974. Detectability of interaural delay in high-frequency complex waveforms. J. Acoust. Soc. Amer. 77, 1129–1140.

Hermansky, H., 1998. Should recognizers have ears? Speech Comm. 25 (1–3), 3–27.

Jeffress, L.A., 1948. A place theory of sound localization. J. Comput. Physiol. Psychol 41, 35–39.

Juang, B.-H., 1991. Speech recognition in adverse environments. Comput. Speech Lang. 5 (3), 275–294.

Kim, Y.-I., Kil, R.M., 2004. Sound source localization based on zero-crossing peak-amplitude coding. In: Proc. Internat. Conf. on Spoken Language Processing (INTERSPEECH-2004), Jeju, Korea, pp. 477–480.

Kim, Y.-I., Kil, R.M., 2007. Estimation of interaural time differences based on zero-crossings in noisy multisource environments. IEEE Trans. Audio Speech Lang. Process. 15, 734–743.

Kim, Y.-I., An, S.J., Kil, R.M., Park, H.-M., 2005. Sound segregation based on binaural zero-crossings. In: Proc. European Conf. on Speech Communication and Technology (INTERSPEECH-2005), Lisbon, Portugal, pp. 2325–2328.

Lindemann, W., 1986. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation for lateralization for stationary signals. J. Acoust. Soc. Amer. 80, 1608–1622.

Lyon, R., 1983. A computational model of binaural localization and separation. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'83), Boston, MA, pp. 1148–1151.

Meddis, R., Hewitt, M.J., 1991. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification. J. Acoust. Soc. Amer. 89 (6), 2866–2882.

Morris, A., Barker, J., Bourlard, H., 2001. From missing data to maybe useful data: soft data modelling for noise robust ASR. In: Proc. IEEE Internat. Workshop on Intelligent Signal Processing, Budapest, Hungary, pp. 153–164.

Nuetzel, J., Hafter, E.R., 1981. Discrimination of interaural delays in complex waveforms: spectral effects. J. Acoust. Soc. Amer. 69, 1112–1118.

O'Mard, L.P., 2000. Development system for auditory modelling. dsam-2.6.62.tar.gz. <http://www.essex.ac.uk/psychology/hearinglab/dsam/>.

Palomäki, K.J., Brown, G.J., Wang, D., 2004. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Speech Comm. 43, 361–378.

Price, P., Fisher, W.M., Bernstein, J., Pallet, D.S., 1988. The DARPA 1000-word resource management database for continuous speech recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'88), New York, NY, pp. 651–654.

Raj, B., Stern, R.M., 2005. Missing-feature methods for robust automatic speech recognition. IEEE Signal Process. Mag. 22 (5), 101–116.

Raj, B., Parikh, V., Stern, R.M., 1997. The effects of background music on speech recognition accuracy. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'97), Munich, Germany, pp. 851–854.

Raj, B., Seltzer, M.L., Stern, R.M., 2004. Reconstruction of missing features for robust speech recognition. Speech Comm. 43 (4), 275–296.

Roman, N., Wang, D., Brown, G.J., 2003. Speech segregation based on sound localization. J. Acoust. Soc. Amer. 114 (4), 2236–2252.

Singh, R., Raj, B., Stern, R.M., 2002a. Model compensation and matched condition methods for robust speech recognition. In: Davis, G. (Ed.), CRC Handbook on Noise Reduction in Speech Applications. CRC Press, pp. 245–276, Chapter 10.

Singh, R., Stern, R.M., Raj, B., 2002b. Signal and feature compensation methods for robust speech recognition. In: Davis, G. (Ed.), CRC Handbook on Noise Reduction in Speech Applications. CRC Press, pp. 219–244, Chapter 9.

Srinivasan, S., Roman, N., Wang, D., 2004. On binary and ratio time-frequency masks for robust speech recognition. In: Proc. Internat. Conf. on Spoken Language Processing (INTERSPEECH-2004), Jeju, Korea, pp. 2541–2544.

Srinivasan, S., Roman, N., Wang, D., 2006. Binary and ratio time-frequency masks for robust speech recognition. Speech Comm. 48, 1486–1501.

Stern, R.M., Colburn, H., 1978. Theory of binaural interaction based on auditory-nerve data. IV. A model of subjective lateral position. J. Acoust. Soc. Amer. 64, 127–140.

Stern, R.M., Trahiotis, C., 1996. Models of binaural perception. In: Gilkey, R., Anderson, T.R. (Eds.), Binaural and Spatial Hearing in Real and Virtual Environments. Lawrence Erlbaum Associates, pp. 499–531, Chapter 24.

Strutt, J.W., 1907. On our perception of sound direction. Philos. Mag. 13, 214–232.

Tessier, E., Berthommier, F., Glotin, H., Choi, S., 1999. A casa front-end using the localisation cue for segregation and then cocktail-party speech recognition. In: Proc. Internat. Conf. on Speech Processing (ICSP-99), Seoul, Korea, pp. 97–102.

Wang, D., Brown, G.J. (Eds.), 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken, NJ.

Weintraub, M., 1986. A computational model for separating two simultaneous talkers. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'86), Tokyo, Japan, pp. 81–84.

Please cite this article in press as: Park, H.-M., Stern, R.M., Spatial separation of speech signals using amplitude estimation ..., Speech Comm. (2008), doi:10.1016/j.specom.2008.05.012