ZERO-CROSSING BASED BINAURAL MASK ESTIMATION FOR MISSING DATA SPEECH RECOGNITION

Young-Ik Kim, Sung Jun An, and Rhee Man Kil

Division of Applied Mathematics
Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea
rmkil@kaist.ac.kr

ABSTRACT

This paper presents a new method of zero-crossing based binaural mask estimation for missing data speech recognition under the condition that multiple sound sources are present simultaneously. The masking is determined by the estimated directions of sound sources using the spatial cues such as inter-aural time differences (ITDs) and inter-aural intensity differences (IIDs). In the suggested method, the estimation of ITDs is utilizing the statistical properties of zero-crossings generated from binaural filter-bank outputs. We also consider the estimation of ITDs with the aid of IID samples to cope with the phase ambiguities of ITD samples in high frequencies. As a result, the proposed method is able to provide an accurate estimate of sound source directions and a good masking scheme for speech recognition while offering significantly less computational complexity compared to cross-correlation based methods.

1. INTRODUCTION

In the human auditory system, a sound source is localized by the differences of signals obtained from both ears. The main cues four sound source localization are inter-aural time differences (ITDs) and inter-aural intensity differences (IIDs). Sound source localization in the human auditory system is used to select a specific sound source among multiple sound sources even in noisy environments. The cocktail party problem is one example of this type of situation. This concept of sound source localization can be applied to automatic speech recognition systems using two sensors. In this application, we consider the case that multiple sound sources are present simultaneously. Here, the directions of sound sources can be identified using the estimation of ITDs and IIDs. Then, the speech segments of the selected sound source in time-frequency domain can be extracted using the information of sound source directions for each segment. After that, the extracted speech segments can be reconstructed to recover the selected sound source or recognized using the missing data technique[1]. For these problems, Roman et al.[2] suggested a masking method in which the masks on the ITD-IID space are trained using a priori knowledge of the target and interferences. However, this method requires a training procedure for every spatial configuration, that is, if the number of sound sources and/or the directions of sound sources are changed, these masks should be retrained. In this sense, it is not favorable for implementing this method in real applications. From this context, we consider a method of identifying the directions of sound sources and segregating the selected sound without the need of training the masks.

Zero-crossings were used to find noise-robust speech features. One example of such a feature is the zero-crossing peak-amplitudes (ZCPAs) proposed by Kim et al[3]. They demonstrated that the auditory model based on zero-crossing features is more robust in noisy environments than other popularly used feature extraction methods, such as LPCC or MFCC. This is mainly due to the dominant frequency principle[4] which states that the number of zero-crossings per unit time is close to two times the frequency of the dominant signal when one exists. From this observation, we propose a method of estimating ITDs using the zero-crossing time differences (ZCTDs) detected from the filter-bank outputs of the left and right sensors. This approach is in accordance with Jeffress's hypothesis[5] in the sense that the time difference is actually measured using delay components and coincidence detectors. In this approach, one of the notable properties in the statistics of the ITD samples is that their variances are closely related to the signal-to-noise ratios (SNRs) of the measured samples, enabling us to identify reliable samples according to the variances of ITD samples. By combining the ZCPA coding and the identification of reliable samples, the suggested method is able to provide an accurate estimate of ITDs. However, it is difficult to estimate ITDs in high frequencies due to the phase ambiguities of zero-crossings. From this point of view, we consider a method of selecting ITD samples according to the estimation of sound source directions using IID samples. As a result, the proposed method is able to provide an accurate estimate of sound source directions and a good masking scheme for speech recognition while offering significantly less computational complexity compared to cross-correlation based methods.

2. ESTIMATION OF ITDS USING ZERO-CROSSINGS

In the estimation of ITDs, it is important to obtain reliable samples, for instance, samples with high SNR, as much as possible. First, as done in the ZCPA coding, a series of band-

This research was supported by *Brain Science and Engineering Research Program* sponsored by Korean Ministry of Commerce, Industry, and Energy.

pass filter is applied to each left and right sensor signals. Here, let us denote $x_i(t)$ as the output signal of the ith channel of the filter-bank. The estimation of ITDs is performed separately for each channel. Suppose there are N (upward) zero-crossings, and zero-crossing times are represented by t_n, $n = 1, 2, \cdots, N$ satisfying $x_i(t_n) = 0$. To distinguish the signals generated from the left and right sensors, we use $x_i^L(t)$ and $x_i^R(t)$ as the signal at the ith channel of the left and right sensors, respectively. We now describe the principle of determining the ITD using zero-crossings. Let us define zero-crossing time in the left channel as t_n^L for $n = 1, 2, \cdots, N$, and in the right channel as t_m^R for $m = 1, 2, \cdots, M$, where N and M represent the number of zero-crossings detected from the left and right channel signals, respectively.

In auditory processing, the ITDs and IIDs convey same information of source source directions. From this observation, we consider the method of selecting valid ITD-IID sample pairs. First, for t_n^L we consider the following candidates of ITD samples:

$$\Delta t_i(n, m) = t_n^L - t_m^R \tag{1}$$

where t_m^R is selected within a window in which the time interval is given by the range of $t_n^L - T$ and $t_n^L + T$ for the time span T. The time span T is determined as 1 ms to cover the range of azimuth angles between -90 and 90 degrees. Here, the problem is to determine the proper t_m^R for the given t_n^L. To solve this problem, we consider the following IID samples:

$$\Delta p_i(n, m) = 10 \log_{10} \frac{p_n^L}{p_m^R} \tag{2}$$

where p_n^L and p_m^R represent the left power at time n and the right power at time m respectively. They are defined by

$$p_n^L = \frac{1}{2W} \sum_{t=t_n^L-W}^{t_n^L+W} (x_i^L(t))^2 \text{ and} \tag{3}$$

$$p_m^R = \frac{1}{2W} \sum_{t=t_m^R-W}^{t_m^R+W} (x_i^R(t))^2 \tag{4}$$

where W is set to $W = 5/$center frequency of the ith channel according to the Ghitza's perceptual window[6].

Since the ITD and IID samples represent the same information of sound source direction, we consider to make a map of sound source directions represented by the azimuth angles measured from the frontal axis versus ITD values and also a map of azimuth angles versus IID values. These maps can be made by investigating the ITD or IID values for the corresponding sound source angles. Here, we can identify the corresponding azimuth angle for the ITD value θ_{ITD} from the map of azimuth angles versus ITD values and also the corresponding azimuth angle for the IID value θ_{IID} from the map of azimuth angles versus IID values. Then, we can search the best matching ITD-IID sample pair by searching the minimum angle difference between θ_{ITD} and θ_{IID}, that is, we can find the time index k for the proper ITD value as

$$k = \arg\min_m |\theta_{ITD}(\Delta t_i(n, m)) - \theta_{IID}(\Delta p_i(n, m))| \tag{5}$$

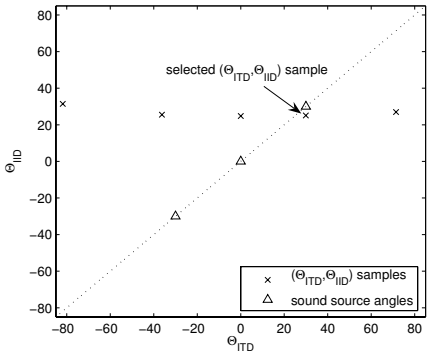


Fig. 1. Selection of a pair of ITD-IID samples: an example of combined azimuth estimate with the channel center frequency of 3.4 KHz where three speech sources are located at azimuth angles of -30, 0, and 30 degrees.

where $\theta_{ITD}(\Delta t_i(n, m))$ and $\theta_{IID}(\Delta p_i(n, m))$ represent the azimuth angle values for the measured values of $\Delta t_i(n, m)$ and $\Delta p_i(n, m)$ respectively. For the illustration of searching the best matching ITD-IID sample pair, refer to Figure 1. As a result, we get the ITD sample $\Delta t_i(n)$ at the ith channel as

$$\Delta t_i(n) = t_n^L - t_k^R \tag{6}$$

where k satisfies (5).

Although we select the best matching ITD-IID sample pair, there is perturbation when we measure the value of (6) due to the environmental noise and/or measurement error. This can be described by the following equations:

$$t_n^L = \bar{t}_n^L + r_n^L \text{ and} \tag{7}$$

$$t_k^R = \bar{t}_k^R + r_k^R \tag{8}$$

where $\bar{t}_n^L$ and $\bar{t}_k^R$ represent the zero-crossing points without noise in the left and right sensor signals respectively, and r_n^L and r_k^R represent the perturbation of zero-crossing points due to noise in the left and right sensor signals respectively. Here, we assume that both r_n^L and r_k^R are identically and independently distributed with mean zero. By investigating these variances, we can show that the variance of $\Delta t_i(n)$ is determined by

$$Var(\Delta t_i(n)) \approx \frac{1}{2w^2} \left(\frac{1}{10^{SNR_L/10}} + \frac{1}{10^{SNR_R/10}}\right). \tag{9}$$

where SNR_L and SNR_R represent the SNRs of the left and right channels respectively. If there is no intensity difference between two sensors, the SNR of the filtered signal can be approximated as

$$SNR \approx 10 \log_{10} \frac{1}{w^2 Var(\Delta t_i(n))}. \tag{10}$$

For more detailed description of the above equation, refer to [7]. The above equation implies that the SNR of the filtered signal can be described by the variance of ITD samples and also the frequency of the filtered signal. Using the formula of

(10), we can estimate the SNR approximately from the variance of ITD samples and the center frequency of bandpass filter used in the filter-bank. Then, the estimated SNR can be effectively used to construct the histogram of ITD samples: the measured ITD samples are weighted by the estimated SNR and accumulated in the histogram of ITD samples. For the sound source localization, the peak values of the histogram are identified and the corresponding ITD values are determined. Then, the sound source direction can be determined from the map of azimuth angles versus ITD values. As a result, we can get the reliable estimation of ITDs in noisy environments. Furthermore, the suggested method offers significantly less computational complexity compared to cross-correlation based methods[7].

3. BINAURAL MASK ESTIMATION FOR MISSING DATA SPEECH RECOGNITION

We will consider the binaural mask estimation based on zero-crossings for missing data speech recognition. Here, we assume that the sound sources are captured by two sensors, L and R. Sensor L is used as the reference sensor, that is, $\Delta t_i(n)$ is estimated from sensor L. In this estimation, the identification of the reliable ITD samples with high SNR is necessary to determine the direction of the sound source accurately. For this purpose, we can consider the identification of reliable ITD samples using (10). Here, we assume that the sound source localization using zero-crossings is done so that we get sound source ITDs as $ITD(j)$, $j = 1, \cdots, J$ corresponding to J sound sources. After the sound source localization, the zero-crossing based masking and the speech recognition procedures are performed as explained in the following:

Step 1. (Selection of valid ITD-IID sample pairs) From the sound source localization procedure, we select valid ITD-IID sample pairs $(\Delta t_i(n,k), \Delta p_i(n,k))$ in which k satisfies the condition of (5).

Step 2. (Estimation of the power for each sound source) For each segment at time τ and frequency i, the following procedure is applied:

- For $(\Delta t_i(n,k), \Delta p_i(n,k))$, $n = n_s, \cdots, n_s + S - 1$ value pairs associated with S zero-crossing points in a segment where n_s represents the index of the first ITD-IID sample in a segment, select the sound source whose ITD value is nearest to the measured ITD value $\Delta t_i(n,k)$. Then, the zero-crossing points are assigned to the selected sound sources.
- For each zero-crossing point in a segment, the signal power p_n between the current and previous zero-crossings is calculated.
- For each sound source, the signal powers associated with the assigned zero-crossings are accumulated:

for $j = 1, \cdots, J$
  $P_j \leftarrow 0$
  for $n = n_s, \cdots, n_s + S - 1$,
    $P_j \leftarrow P_j + p_n$ if $j = \arg\min_l |\Delta t_i(n,k) - ITD(l)|$

Step 3. (Mask estimation in time-frequency domain) For each segment at time τ and frequency i, the accumulated power for the selected sound source (or target sound source) is compared with the powers for other interfering sound sources. If the power for the target sound source is larger than the sum of powers for other interfering sound sources, the masking value of the segment is 1 (passing). Otherwise, the masking value of the segment is 0 (blocking). That is, if the target sound source is the qth sound source. Then, the binary mask M_b can be determined by

$$M_b(\tau, i) = \begin{cases} 1 & \text{if } P_q / \sum_{j=1}^{J} P_j > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

We can also use the ratio mask M_r using the power ratio of the target to all sound sources:

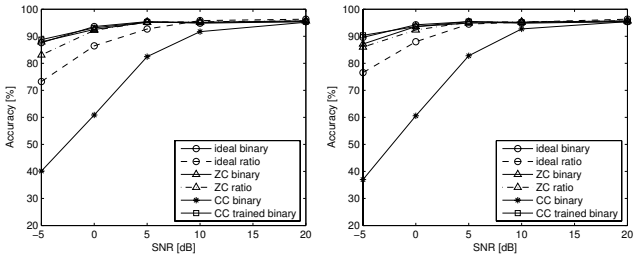
$$M_r(\tau, i) = \begin{cases} P_q / \sum_{j=1}^{J} P_q & \text{if } P_q / \sum_{j=1}^{J} P_j > 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Step 4. (Speech recognition with missing data) From the selection of reliable speech segments of step 3, we can obtain speech features composed of reliable and unreliable parts. These features can be applied to the missing data classifier[1] in which classification results are determined by the features with reliable parts.

In principle, the suggested algorithm determines which source is originated for each zero-crossing using the ITD-IID value pair. The masking is done by comparing the powers associated with zero-crossings.

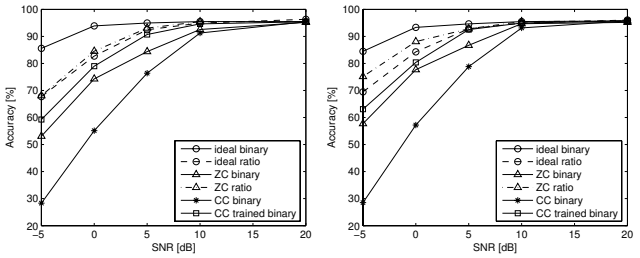
4. SIMULATION

For the simulation of speech recognition, the sound source signals were transformed by the Head Related Transfer Function (HRTF)[8] and decomposed by a gama-tone filter-bank with 32 channels in which the center frequencies are spaced linearly in ERB-scale from 200 to 4000 Hz. In this simulation, we used the TI-DIGIT corpus to train 12 word-level HMMs ('1'-'9', 'oh', 'zero', and 'silence') using auditory rate maps. A subset of 106 male utterances from the test set of the corpus was used as target speech sources. We also selected speech samples randomly from the test corpus for interfering speech sources. The simulation of speech recognition was made by the zero-crossing (ZC) based binary and ratio masking methods as explained in the step 3 of the suggested algorithm. For the comparison, the cross-correlation (CC) based masking methods were tested: one method was to estimate the ITD value from a segment which was most consistent with the IID value and to determine the binary masking according to the decision whether the estimated ITD value was the nearest ITD value of the selected speech source, and another method was to determine the binary masking according to the previously trained masks in the ITD-IID space[2]. We also made the simulation for speech recognition using the ideal binary and ratio masking methods in which masks were obtained from a priori knowledge of the target and interferences.



(a) (b)

Fig. 2. Recognition performance for ZC and CC based masking methods in the case of one interference located (a) at an azimuth angle of 5°, and (b) at an azimuth angle of 30°.



(a) (b)

Fig. 3. Recognition performance for ZC and CC based masking methods in the case of two interferences located (a) at azimuth angles of -5° and 5°, and (b) at azimuth angles of -30° and 30°.

For the classification, we used the missing data speech recognition using the CASA toolkit (CTK)[9]. Here, we assumed that the target speech sources were located at the frontal axis, that is, 0 degree, and the interfering speech sources were located at an azimuth angle of 5 degree or at an azimuth angle of 30 degree in the two source cases while located at azimuth angles of -5 and 5 degrees or at azimuth angles of -30 and 30 degrees in the three source cases. The simulation results of speech recognition according to the SNRs which were measured from the ratio of the target power to the sum of interference powers, were illustrated in Figures 2 and 3. These results showed that 1) the ZC binary masking method obtained the better recognition performance than the CC binary masking method in both cases, 2) the ideal, CC trained, and ZC masking methods obtained the similar recognition performance under the condition of one interference, and 3) the ZC ratio masking method obtained the better recognition performance than the CC trained and ZC binary masking methods under the condition of two interferences. This is due to the fact that the errors in power estimation during the masking process become larger as the number of interferences increases so that the binary masking can make larger variance of errors in the estimation of total powers than the ratio masking. Over all, without the need of training the masks, the suggested ZC based masking methods achieved the comparable performance to the CC trained binary masking method which requires to train masks for every spatial configuration.

5. CONCLUSION

We have suggested an approach of speech recognition with a new method of binaural masking in time-frequency domain using zero-crossings. The distinctive feature of our masking method is that the estimation of ITDs is utilizing the statistical properties of zero-crossings generated from binaural filterbank outputs so that we get more reliable ITD samples in noisy environments. We also consider the estimation of ITDs with the aid of IID samples to cope with the phase ambiguities of ITD samples in high frequencies. As a result, the proposed method provides a good masking scheme for speech recognition while offering significantly less computational complexity compared to cross-correlation based methods. Simulation results for speech recognition under the condition of various interfering sound sources showed that the suggested method achieved the comparable performance to the cross-correlation trained binary masking method which requires to train masks for every spatial configuration.

6. REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communications*, vol. 34, pp. 267–285, 2001.

[2] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *Journal of Acoustic Society of America*, vol. 114, pp. 2236–2252, 2003.

[3] D. Kim, S. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, 1999.

[4] B. Kedem, *Time series analysis by higher order crossings*, IEEE Press, 1994.

[5] L. Jeffress, "A place theory of sound localization," *Journal of Computational Physiology and Psychology*, vol. 41, pp. 35–39, 1948.

[6] O. Ghitza, "Auditory models and human performances in tasks related to speech coding and speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 115–132, 1994.

[7] R. M. Kil and Y. Kim, "Sound source localization based on zero-crossing peak-amplitude coding," in *INTERSPEECH*, Jeju, Korea, October 2004.

[8] W. Gardner and K. Martin, "Hrtf measurement of a kemar," *Journal of Acoustic Society of America*, vol. 97, pp. 3907–3908, 1995.

[9] The RESPITE CASA toolkit project, *A toolkit for computational auditory scene analysis*, http://www.dcs.shef.ac.uk/~jon/ctk.html, 2001.