Zero-Crossing-Based Ratio Masking for Sound Segregation

Sung Jun An\textsuperscript{1}, Young-Ik Kim\textsuperscript{2}, and Rhee Man Kil\textsuperscript{1}

\textsuperscript{1}Department of Mathematical Sciences, \textsuperscript{2}Brain Science Research Center
Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea
curtaincall@kaist.ac.kr, youngik@kaist.ac.kr, rmkil@kaist.ac.kr

1. Introduction

In the human auditory system, a sound source is localized by the differences of signals obtained from both ears. The main cues are inter-aural time differences (ITDs) and inter-aural intensity differences (IIDs). It has been known that the human auditory system is able to select a specific sound source among multiple sound sources even in noisy environments using various cues including these localization cues, for instance, the capability of handling the cocktail party problem. This concept of sound source localization can be applied to sound source segregation using two sensors. In our approach, a target sound source mixed with multiple sound sources is segregated using the masking scheme in which sound segments in the time-frequency domain originated from a target sound source are selected and other sound segments originated from interfering sound sources are blocked. This approach is feasible under the assumption that the signals have non-overlapping time-frequency representation supports. This assumption has been shown to be approximately true for speech signals \cite{1}. This direction of research was considered from Bregman’s theory of auditory scene analysis (ASA) \cite{2} in which the notion of auditory grouping, that is, elements that are likely to have arisen from the same environment event are combined, was used to segregate the target sound source from an acoustic mixture. Among this direction of works, the masking methods \cite{3, 4} in which the masks in time-frequency domain to select the target source are trained using a priori knowledge of the target and interferences, showed outstanding performances of speech segregation. One of the problems in these works is that they require the 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 investigate the zero-crossing-based method for the possible solutions of these problems. In this work, 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 \cite{5} in which the time difference is actually measured using delay components and coincidence detectors. In this approach, one of the notable properties in the statistics of ITD estimates is that their variances are closely related to the signal-to-noise ratios (SNRs) of filtered signals, enabling us to identify reliable samples according to the variances of ITD estimates. As a result, the proposed approach \cite{6} was effective in localizing multiple sound sources in noisy environments while offering significantly less computational complexity compared to cross-correlation-based methods. However, this approach can be used only for low frequency channels due to phase ambiguity in high frequency channels. To cope with this problem, we consider to use IID estimates in addition to ITD estimates to resolve the phase ambiguity in high frequency channels. After localizing the sound sources, the segments in time-frequency domain are further analyzed in such a way that the direction of signal between two adjacent upward-zero-crossings is tested using the ITD-IID estimates. Then, the ratio of the target source power to the total power in the segment is estimated and the decision of masking for each segment is determined by this estimated power ratio. We show that this power ratio is optimal from the view point of reconstructing the target speech signal. As a result, the suggested method is able to provide an accurate estimate of multiple sound source directions and also a good masking scheme for sound segregation while offering significantly less computational complexity compared to cross-correlation-based methods. Furthermore, the suggested method does not require the training of masks according to the spatial configuration of sound sources.

2. Zero-Crossing-Based Sound Source Localization

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 zero-crossing peak-amplitude (ZCPA) cod-
ing, a series of bandpass filter is applied to each left and right sensor signals. Here, let us denote \(x_i(t)\) as the output signal of the \(i\)th 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 \(i\)th 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 \(t_i^L\) for \(n = 1, 2, \cdots, N\), and in the right channel as \(t_i^R\) 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 the auditory processing, the ITDs and IIDs convey same information of sound source directions. From this observation, we consider the method of selecting valid ITD-IID sample pairs. First, for \(t_i^L\) we consider the following candidates of ITD samples:

\[
\Delta t_i(n, m) = t_i^L - t_i^R 
\]

where \(t_i^R\) is selected within a window in which the time interval is given by the range of \(t_i^L = t_i^L + W\) for the time span \(T\). The time span \(T\) is determined as 1 ms. Here, the problem is to determine the proper \(t_i^R\) for the given \(t_i^L\). To solve this problem, we consider the following IID samples:

\[
\Delta p_i(n, m) = 10 \log_{10} \frac{p_i^L}{p_i^R} 
\]

where \(p_i^L\) and \(p_i^R\) represent the left power at time \(n\) and the right power at time \(m\) respectively. They are defined by

\[
p_i^L = \frac{1}{2T} \sum_{t = t_i^L - W}^{t_i^L + W} (x_i^L(t))^2 \tag{3}
\]

\[
p_i^R = \frac{1}{2T} \sum_{t = t_i^R - W}^{t_i^R + W} (x_i^R(t))^2 \tag{4}
\]

where \(W\) is set to 5/cycle frequency of the \(i\)th channel according to the Ghitta’s perceptual window [7].

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 angles measured from the frontal axis versus ITD values and also a map of sound source angles versus IID values. These maps can be made by investigating the ITD or IID values for the corresponding sound source angles versus IID values. Here, we can identify the corresponding angle for the ITD value \(\theta_{ITD}\) from the map of angles versus ITD values and also the corresponding angle for the IID value \(\theta_{IID}\) from the map of angles versus IID values. Then, we can search the best matching ITD-IID sample pair by searching the minimum angle difference between \(\theta_{ITD}\) and \(\theta_{IID}\), that is, we can find the time index \(k\) for the proper ITD value as

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

where \(\theta_{ITD}(\Delta t_i(n, m))\) and \(\theta_{IID}(\Delta p_i(n, m))\) represent the angle values for the measured values of \(\Delta t_i(n, m)\) and \(\Delta p_i(n, m)\) respectively. As a result, we get the ITD sample \(\Delta t_i(n)\) at the \(i\)th channel as

\[
\Delta t_i(n) = t_i^L - t_i^R 
\]

where \(k\) satisfies (5).

Although we select the best matching ITD-IID sample pair, there are perturbations when we measure the value of (6) due to the environmental noise and/or measurement error. From this point of view, we investigated the relationship between the variance of \(\Delta t_i(n)\) and the SNRs of left and right channels, and concluded the following result [6]:

\[
V a r(\Delta t_i(n)) \approx \frac{\Delta t^2_i}{2w_i^2} \left( \frac{1}{10^{SNR^L_i/10}} + \frac{1}{10^{SNR^R_i/10}} \right) 
\]

where \(w_i\) represents the frequency of the signal from the \(i\)th channel, and \(SNR_L\) and \(SNR_R\) represent the SNRs of the signals from the \(i\)th channels of the left and right sensors respectively. If there is no intensity difference between two sensors, the SNR of the filtered signal can be approximated as

\[
SNR_i \approx 10 \log_{10} \frac{1}{w_i^2 V a r(\Delta t_i(n))}. 
\]

Using the formula of (8), 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. Using this approach [6], the reliable estimation of ITDs is possible under noisy multi-source environments while offering significantly less computational complexity compared to cross-correlation based methods.

3. Zero-Crossing Based Mask Estimation

We now describe the algorithm for sound segregation based on the sound source localization using zero-crossings. Here, we assume that the sound source is 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 \(L\). Here, we assume that the sound source localization using zero-crossings is done so that we have sound source ITD values \(ITD(j)\), \(j = 1, \cdots, K\) corresponding to \(K\) sound sources. Afterwards, the sound segregation algorithm based on the sound source localization using the bинаural zero-crossing time differences (ZCTDs) is described as follows:

**Step 1.** (Selection of consistent ITD-IID sample pairs) 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 powers of sound sources) For each time frame \(\tau\) and frequency \(i\), the following procedure is applied:

- For the ITD samples \(\Delta t_i(n)\) of (6), \(n = 1, \cdots, N\) associated with \(N\) zero-crossing points in a time frame, each zero-crossing point is assigned to the selected sound source corresponding to the nearest sound source ITD value, that is, one of \(ITD(j), j = 1, \cdots, K\).
- For each zero-crossing point in a time frame, the signal powers \(p_i(n)\), \(n = 1, \cdots, M\) between the current and previous zero-crossings are calculated.
- For each sound source, the signal powers associated with the assigned zero-crossings are accumulated: for \(j = 1, \cdots, K\),
\( P_j \leftarrow 0. \)
\[ \text{for } n = 1, \ldots , M, \]
\( P_j \leftarrow P_j + p_n \text{ if } |\Delta t_i(n) - ITD(j)| = \min_l |\Delta t_i(n) - ITD(l)|. \)

**Step 3.** (Selection of segments in time-frequency domain) For each time frame \( \tau \) and frequency \( i \), the ratio \( M_r(\tau, i) \) of the accumulated power for the selected sound source (or target sound source) to the total powers in the segment is calculated:

\[ M_r(\tau, i) = \frac{P_i}{\sum_{j=1}^{R} P_j} \]  
(9)

where \( P_i \) represents the estimated target power, that is, \( M_r(\tau, i) \) represents the ratio of the target source power to the total power in the segment at the time frame \( \tau \) and frequency \( i \).

**Step 4.** (Speech segregation) For each segment, the signal is multiplied by the ratio mask \( M_r(\tau, i) \) and the masked signals are reconstructed.

In this algorithm, the masking strategy for each segment is determined by the estimated ratio of the target source power to the total power from a view point of mean square error between the target and mixed signals. First, let us consider the following signal \( s_i \) for the segment at the time frame \( \tau \) and frequency \( i \):

\[ s_i(n) = s_i^T(n) + s_i^f(n) \quad \text{for } n \in \Lambda(\tau, i) \]  
(10)

where \( \Lambda \) represents the index set of signals for the segment, and \( s_i^T \) and \( s_i^f \) respectively represent the targets for the signal and the mixture of interferences. Let us consider the ratio of masking as \( r \). Then, the error \( e_i \) between the target signal and the signal with masking is given by

\[ e_i(n) = s_i^T(n) - rs_i(n) = s_i^T(n) - r(s_i^T(n) + s_i^f(n)) \]  
(11)

and the mean square error \( E \) for the segment is given by

\[ E(\tau, i) = \sum_{n \in \Lambda(\tau, i)} e_i^2(n) \]  
(12)

Since the above equation has the quadratic form, we can get the optimal ratio \( r^* \) by finding the extremum point of \( r \), that is, from

\[ \frac{dE(\tau, i)}{dr} \bigg|_{r=r^*} = 0, \]  
(13)

we get

\[ r^* = \frac{\sum_{n \in \Lambda(\tau, i)} s_i^T(n)(s_i^T(n) + s_i^f(n))}{\sum_{n \in \Lambda(\tau, i)} (s_i^T(n) + s_i^f(n))^2}. \]  
(14)

Here, we assume that target and interference signals are uncorrelated so that

\[ \sum_{n \in \Lambda(\tau, i)} (s_i^T(n))^2 \gg \sum_{n \in \Lambda(\tau, i)} s_i^T(n)s_i^f(n). \]  
(15)

Then, the optimal ratio \( r^* \) can be approximated as

\[ r^* \approx \frac{\sum_{n \in \Lambda(\tau, i)} (s_i^T(n))^2}{\sum_{n \in \Lambda(\tau, i)} (s_i^T(n) + s_i^f(n))^2}. \]  
(16)

This target-to-total power ratio, that is, the ratio of the target source power to the total power in the segment, is used in our masking scheme in step 3.

In principle, the suggested algorithm determines which sound source is more suitable for each segment in the time-frequency domain. Then, the masking is done by comparing the powers associated with the selected target. The advantages of the suggested algorithm are 1) the robustness to noise due to the dominant frequency principle of zero-crossings, 2) the less computational complexity involved in estimating ITDs than the cross-correlation-based methods, and 3) no need to train the masks for sound segregation since the decision of masking is made by the target-to-total power ratio in each segment.

### 4. Simulation

To show the effectiveness of the suggested ZCTD masking method, we performed experiments on speech segregation for various configurations of target and interfering sound sources. For both experiments, the sound sources were transformed by the Head-Related-Transfer-Function (HRTF) [8] and decomposed by a gamma-tone filter-bank composed of 128 channels, in which the center frequency of each channel was between 0.08 and 5.0 kHz. Then, the suggested ZCTD masking method was applied. In this simulation, we used the data set collected by Cooke [9] which contains ten voiced speech signals and ten noise intrusions, encompassing a variety of common acoustic interferences such as telephone ring, rock music, and other speech utterances. First, we assume that the target sound source was located in the frontal axis, that is, 0 degree azimuth angle. In the case of two sound sources, the interfering sound source was located at an azimuth angle of 5, 30, or 60 degrees. In the case of three sound sources, the interfering sound sources were located at azimuth angles of -5 and 5 degrees, -30 and 30 degrees, or -60 and 60 degrees. We also consider the situation that the target sound source was not located in the frontal axis, that is, we consider the cases that the target and interfering sound sources were respectively located at azimuth angles of 10 and 20 degrees, 40 and 50 degrees, or 70 and 80 degrees.

The simulation of sound segregation was made using the suggested ZCTD ratio masking method. We also consider the ZCTD binary masking method in which the binary masks, that is, the mask value equals to 1 if \( M_r(\tau, i) > 0.5 \), otherwise, we used instead of ratio masks in the speech segregation process of step 4 of the suggested algorithm. To compare with the suggested method, we consider the method using trained masks [3] in which the ITD value was estimated using the cross-correlation method and masking was performed on the ITD-IID space using the previously trained masks. We call this mask as the cross-correlation-based mask with learning (CCL). As the references of segregation performances, we also made the simulation for sound segmentation using the ideal binary and ratio masks which were obtained from a priori knowledge of the target and interferences. To measure the performance of sound segregation quantitatively, a segregated signal is reconstructed from a binary mask and the following SNR, the ratio in decibel of the target signal power to the power of error between the target and reconstructed signals, is considered:

\[ SNR = 10 \log_{10} \frac{\sum_t s_T^2(t)}{\sum_t (s_T^2(t) - s_E(t))^2} \]  
(17)

where \( s_T \) represents the target signal reconstructed using all-one mask under no interference and \( s_E(t) \) represents the estimated target signal reconstructed from the binary mask generated by the masking method.
For the speech segregation experiments, the simulation results of average SNRs for 100 trials were obtained using the ideal binary, ideal ratio, trained binary, ZCTD binary, and ZCTD ratio masking methods and plotted in Figures 1 through 3. These results showed that 1) as we expected, the ratio masking methods outperformed the binary masking methods, 2) the trained masking and the suggested ZCTD ratio masking methods provided the similar segregation performances when the target was located in the frontal axis, 3) the segregation performances of masking methods became smaller when the number of sound sources increased and also when the target was moved far away from the frontal axis, and 4) the suggested ZCTD ratio masking method provided some merits in the segregation performances compared to the cross-correlation-based methods when the target sound sources were not located in the frontal axis. This is due to the fact that the ZCTD method provides the robust estimation of sound source directions for wide range of angles.

5. Conclusion

We have suggested a method of sound source segregation using the masking method in which ITDs are estimated using zero-crossings detected from binaural filter-bank outputs. We also consider the estimation of ITDs with the aid of IID estimates to cope with the phase ambiguities of ITD samples in high frequencies. For the masking method, we consider to use the target-to-total power ratio in each segment of the time-frequency domain. As a result, the proposed method is able to provide an accurate estimate of sound source directions and also a good masking scheme for sound segregation while offering significantly less computational complexity compared to cross-correlation-based methods. Simulation results for sound segregation in various interfering sound sources show that the suggested ZCTD masking method provides the similar performances with the cross-correlation-based masking method using previously trained masks. This makes the suggested method more favorable in real applications since it does not require any training procedure while the trained masking methods require the training procedure for every spatial configuration. Furthermore, the suggested ZCTD masking method gives some merits in segregation performances compared to cross-correlation-based masking methods especially when the target is not located in front of two sensors.

6. References