International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.5 (2015), pp.303-314
http://dx.doi.org/10.14257/ijmue.2015.10.5.28

Binaural Sound Source Localization based on Sub-band SNR Estimation

Zhou Lin, Zhao Xiao-Yan, Cheng Xu and Wu Zhen-Yang

School of Information Science and Engineering, Southeast University, Nanjing, China
Linzhou@seu.edu.cn

Abstract

Sound Source Localization (SSL) has a wide application in speech separation, recognition and enhancement. Binaural sound source localization based on human spatial hearing mechanism is an important research field of SSL. The recent binaural SSL research is focused on the system robust against noise and reverberation. In order to improve the localization performance in degraded environment, this paper proposes an algorithm to adaptively select the 'good' sub-bands to compute the binaural localization cues. Firstly, sub-band Signal-Noise Ratio (SNR) is estimated based on the auto-correlation matrix of binaural sound signals. Then, Inter-aural Time Difference (ITD) is computed by adaptively selecting the sub-bands which have the high SNR. Since the ITD is calculated through the sub-bands which are less affected by the noise, the sound source azimuth is estimated more accurate. The simulation results show that compared to the conventional binaural SSL algorithm, the localization accuracy of the proposed algorithm has been improved significantly.

Keywords: *Binaural sound source localization; Sub-band signal-noise ratio estimation; Binaural localization cues*

1. Introduction

Sound Source Localization (SSL) has a wide application in Robot navigation, speech recognition and blind source separation. One method to localize the sound source is based on microphone array, in which multi-channel signals need to be processed. That means a large computational complexity. The other method is based on the human hearing mechanism, which uses binaural sound source to estimate the sound source direction. The binaural localization method has the advantage of simple structure, small complexity and accurate localization, which makes it an important research field of SSL.

Lord Rayleigh's [1] 'duplex theory' firstly analyzed the physical properties of binaural perception. Rayleigh pointed out that two physical cues are used to perceive the sound locations, which are Inter-aural Time Difference (ITD) and Inter-aural Intensity Difference (IID). ITD is due to the distance difference from sound source to two ears, and IID is due to the intensity difference from shading effect of head. Li [2] explored the ITD, IID and frequency spectrum of binaural signals to localization sound source based on Bayes criterion. Raspaud and Evangelista [3] used the Fourier Transformation (FT) to estimate the sound direction for clean sound. Kim [4] proposed the algorithm based on Zero-Crossing Time Difference (ZSTD) to estimate the sound source direction, and extended to multi-band ZSTD [5]. Stern [6] utilized the Short-Time Fourier Transformation (STFT) and Gammatone weight to calculate the ITD. In order to reduce the effect of noise and reverberation on binaural SSL, Tobias [7] trained the ITD and IID of each band in reverberation environment using Gaussian Mixed Model (GMM). This method can realize the binaural SSL for reverberation signal. Rodemann [8] also trained the signal in reverberation environment, and utilized the signal onset to imitate the

ISSN: 1975-0080 IJMUE

precedence effect human hearing. Wang [9, 10] realized the binaural SSL based on ITD, IID and pitch information. Also, Wang [11] traced the moving speaker using Hidden Markov Model (HMM) and binaural cues. Stern [12-14] improved the performance of binaural SSL through re-reverberation and echo-suppression. For noisy environment, Uncini [15] employed the pre-filter in spectrum domain to reduce the noise disturb. Tobias [16] utilized the GMM model to ITD and IID, which realized the localization and speaker recognition in reverberation and noisy environment. Also, Karim [17] provided the binaural cues as inputs for a neural network. This learning approach obtained good azimuth and elevation angles estimates performance.

Since the performance of binaural SSL degraded rapidly in noisy and reverberation environment, the research of anti-noise and anti-reverberation is the key point of SSL. Based on the SNR estimation method for microphone array [18], this paper presents an algorithm to estimate sub-band SNR of each frame, which utilizes the auto-correlation matrix of binaural signals in sub-band. Then, the proposed algorithm adaptively selects the sub-bands with high sub-band SNR to estimate the ITD and localize sound source. Since ITD are estimated by the sub-bands which are less affected by noise, the proposed algorithm significantly improves the performance of binaural SSL.

This rest of the paper is organized as follows: Section 2 introduces the principle of binaural SSL. Section 3 describes the method of sub-band SNR estimation and binaural SSL in detail. Section 4 gives the simulation results and analysis.

2. The Structure of Binaural Sound Source Localization

In the process of sound propagation from source to two ears of human, the sound wave is reflected and scattered. All those influence can be integrated by the transfer function, which is called Head Related Transfer Function (HRTF), the corresponding time form is Head Related Impulse Response (HRIR). HRTF is defined as the spectral acoustic transform function from the sound source to eardrum in free sound field, which reflects the acoustic filter effect of human physiological structure on sound wav. The HRTF is described in Eq.(1):

$$\begin{aligned} H_L &= H_L(r,\theta,\phi,f) = \frac{P_L(r,\theta,\phi,f)}{P_0(r,f)} \\ H_R &= H_R(r,\theta,\phi,f) = \frac{P_R(r,\theta,\phi,f)}{P_0(r,f)} \end{aligned} \tag{1}$$

where H_L and H_R denote HRTF of left ear and right ear respectively; P_L and P_R are the spectral sound pressure of left ear and right ear from the point sound source; P_0 is the spectral sound pressure in the original position of head when the human moves.

According to Eq.(1), HRTF is the function of distance r from sound source to head, sound source azimuth θ, elevation ϕ and frequency f. HRTF integrates the localization information of human ear, and each spatial position corresponds to a pair of HRTF.

Based on HRTF, we introduce the binaural SSL process. Here we assume the sound source is $s(t)$. The left-ear signal $x_L(t)$ and right-ear signal $x_R(t)$ are defined as binaural signals, which are given by:

$$\begin{aligned} x_L(t) &= h_L * s(t) + n_L(t) \\ x_R(t) &= h_R * s(t) + n_R(t) \end{aligned} \tag{2}$$

where h_L and h_R represent the HRIR of left-ear and right ear; " $*$ " denotes linear convolution; $n_L(t)$ and $n_R(t)$ represent noise signal of left-ear and right-ear, which include reverberation and additive noise. Here, reverberation is supposed to be related with source, while additive noise is unrelated with the source.

After binaural signals are pre-processed (framed and windowed) and Voice Activity Detected(VAD), $x_L(t)$ and $x_R(t)$ become discrete signals $x_L(i,n)$ and $x_R(i,n)$,where i indicates the number of frame. ITD are extracted for each frame. The ITD is the delay

corresponding to the maximum cross-correlation function of binaural signals. ITD_i of the ith frame is derived in time domain through Eq.(3):

$$ITD_i = \arg\max_{\tau} R_{x_L x_R}(i,\tau) = \arg\max_{\tau}\left(\sum_{n=0}^{N-1} x_L(i,n)x_R(i,n+\tau)\right) \tag{3}$$

where N represents the frame length; τ is time delay; $R_{xLxR}(i,\tau)$ is the cross correlation function of binaural signals in the ith frame.

Since Eq.(3) involves a larger number of multiplications, ITD_i is generally computed in frequency domain. First, the spectral magnitude $X_L(i,k)$ and $X_R(i,k)$ of binaural signals in the ith frame are calculated through FT with Eq.(4):

$$\begin{aligned} X_L(i,k) &= \sum_{n=0}^{N-1} x_L(i,n)\exp(-j2\pi kn/N) \\ X_R(i,k) &= \sum_{n=0}^{N-1} x_R(i,n)\exp(-j2\pi kn/N) \end{aligned} \tag{4}$$

Cross-Power Spectral Density (PSD) of binaural signals is given by:

$$P_{LR}(i,k) = X_L(i,k) * X_R^*(i,k) \tag{5}$$

$R_{xLxR}(i,\tau)$ is derived by inverse transform of the cross PSD $P_{LR}(i,k)$ of Eq.(5):

$$R_{x_L x_R}(i,\tau) = \sum_{k=0}^{N-1} P_{LR}(i,k)\exp(j2\pi nk/N) \tag{6}$$

Once the ITD_i is estimated, it is compared with ITD models. And then based on the certain criterion, the direction of sound source in each frame is determined. But in real environment, $R_{xLxR}(i,\tau)$ has several pseudo-peaks due to the existence of interference. That will cause performance degradation of localization method based on ITD.

3. Binaural Sound Source Localization based on Sub-band SNR Estimation

Since the sound source and noise has different distribution in spectrum, the noise interference on sound in different frequency within one frame is different. Also, because of the non-stationary characteristics of sound source and noise, the interference in different frames is also different. According to the above considerations, this paper estimates sub-band SNR estimation for each frame. The process of binaural SSL based on sub-band SNR estimation is described in detail as follows.

Based on Eq.(2), the signal model in frequency domain for a frame is expressed in Eq.(7):

$$\begin{aligned} X_L(i,k) &= H_L(k)S(i,k) + N_L(i,k) \\ X_R(i,k) &= H_R(k)S(i,k) + N_R(i,k) \end{aligned} \tag{7}$$

where $H_L(k)$ and $H_R(k)$ are HRTF, which are unrelated to the frame number i; $X_L(i,k)$ and $X_R(i,k)$ represent left-ear and right-ear spectrum of the kth frequency bin in the ith frame respectively; $N_L(i,k)$ and $N_R(i,k)$ are spectrum of noise in both ears of the kth frequency bin in the ith frame.

Binaural spectral signal vector is defined as $X(i,k)=[X_L(i,k), X_R(i,k)]^T$, the corresponding auto-correlation matrix is derived in Eq.(8):

$$R(i,k) = E[X(i,k)X(i,k)^H] = \begin{bmatrix} E[X_L^2(i,k)] & E[X_L(i,k)X_R(i,k)] \\ E[X_L(i,k)X_R(i,k)] & E[X_R^2(i,k)] \end{bmatrix} \tag{8}$$

In Eq.(8), since the auto-correlation matrix of the kth frequency bin is estimated only by $X_L(i,k)$ and $X_R(i,k)$ in current frame, the estimation error is relatively large. Thus, this paper doesn't intend to estimate the auto-correlation matrix of each frequency bin, but estimates the auto-correlation matrix of each sub-band instead. The proposed method divides spectrum of each frame into M sub-bands, so each sub-band has $L=N/M$ frequency

bins. The method estimates the auto-correlation matrix of each sub-band using *L* frequency, which improves the estimation accuracy.

Here defines the auto-correlation matrix $R(i,m)$ of the mth sub-band in the ith frame:

$$R(i,m)=\begin{bmatrix} E[X_L^2(i,m)] & E[X_L(i,m)X_R(i,m)] \\ E[X_L(i,m)X_R(i,m)] & E[X_R^2(i,m)] \end{bmatrix} \tag{9}$$

The diagonal elements $E[X_L^2(i,m)]$ and $E[X_R^2(i,m)]$ respectively indicate the mean square value of left-ear spectrum, right-ear spectrum of the mth sub-band in the ith frame. $E[X_L(i,\mathrm{m})X_R(i,m)]$ is the correlation function of left-ear spectrum and right-ear spectrum of the mth sub-band in the ith frame.

Sound source $S(i,k)$ is assumed to unrelated to $N_L(i,k)$ and $N_R(i,k)$ Also, $N_L(i,k)$ and $N_R(i,k)$ are unrelated. Based on the above assumptions and Eq.(7), each variable of $R(i,m)$ is computed as follows:

$$\begin{aligned} E[X_L^2(i,m)] &= E\left[\left(H_L(m)S(i,m)+N_L(i,m)\right)\left(H_L(m)S(i,m)+N_L(i,m)\right)\right] \\ &= E\left[H_L^2(m)S^2(i,m)+N_L^2(i,m)\right] \\ &= P_L(i,m)+\sigma_{N_L}^2(i,m) \end{aligned} \tag{10}$$

$$E[X_R^2(i,m)] = P_R(i,m)+\sigma_{N_R}^2(i,m) \tag{11}$$

where $P_L(i,m)$ and $P_R(i,m)$ represents the spectral power of received left-ear clean sound and received right-ear clean sound of the mth sub-band in the ith frame; $\sigma_{NL}^2(i,m)$ and $\sigma_{NR}^2(i,m)$ are variance of $N_L(i,k)$ and $N_R(i,k)$ respectively.

Also:

$$\begin{aligned} E[X_L(i,m)X_R(i,m)] &= E\left[\left(H_L(k)S(i,k)+N_L(i,k)\right)\left(H_R(k)S(i,k)+N_R(i,k)\right)\right] \\ &= E\left[H_L(k)S(i,k)H_R(k)S(i,k)\right] \end{aligned} \tag{12}$$

Since the spectrum is divided into several sub-bands, the ratio $H_L(k)/H_R(k)$ is supposed to be a constant within a sub-band, that is IID_m. Thus, the Eq.(12) is rewritten as:

$$E[X_L(i,m)X_R(i,m)] = E[IID_m H_R^2(m)S^2(i,m)] = IID_m P_R(i,m) \tag{13}$$

Based on Eq.(10), (11)and(13), $R(i,m)$ is derived as follows:

$$R(i,m)=\begin{bmatrix} P_L(i,m)+\sigma_{N_L}^2(i,m) & IID_m P_R(i,m) \\ IID_m P_R(i,m) & P_R(i,m)+\sigma_{N_R}^2(i,m) \end{bmatrix} \tag{14}$$

Here we suppose that $\sigma_{NL}^2(i,m)$ and $\sigma_{NR}^2(i,m)$ are the same, that is $\sigma^2(i,m)$. It should be noted that, the noise variance varies with the sub-band and frame. And also based on the formulation $P_L(i,m)/P_R(i,m)=IID_m^2$, $R(i,m)$ is derived in the following:

$$R(i,m)=\begin{bmatrix} IID_m^2 P_R(i,m)+\sigma^2(i,m) & IID_m P_R(i,m) \\ IID_m P_R(i,m) & P_R(i,m)+\sigma^2(i,m) \end{bmatrix} \tag{15}$$

According to the Eq.(15), the diagonal elements of $R(i,m)$ is the sum of sound source power and noise variance in the mth sub-band, the non-diagonal elements is the sound source power.

Also, there are *L* frequency bins in the mth sub-band to estimate each component in Eq.(9):

$$\begin{aligned} E[X_L^2(i,m)] &= \frac{1}{L}\sum_{k=(m-1)*L+1}^{m*L} X_L^2(i,k) \\ E[X_R^2(i,m)] &= \frac{1}{L}\sum_{k=(m-1)*L+1}^{m*L} X_R^2(i,k) \\ E[X_L(i,m)X_R(i,m)] &= \frac{1}{L}\sum_{k=(m-1)*L+1}^{m*L} X_L(i,k)X_R(i,k) \end{aligned} \tag{16}$$

Thus, Eq.(15) and Eq.(16) are equal:

$$R(i,m)=\begin{bmatrix} IID_m^2 P_R(i,m)+\sigma^2(i,m) & IID_m P_R(i,m) \\ IID_m P_R(i,m) & P_R(i,m)+\sigma^2(i,m) \end{bmatrix} = \begin{bmatrix} \frac{1}{L}\sum_{k=(m-1)*L+1}^{m*L} X_L^2(i,k) & \frac{1}{L}\sum_{k=(m-1)*L+1}^{m*L} X_L(i,k)X_R(i,k) \\ \frac{1}{L}\sum_{k=(m-1)*L+1}^{m*L} X_L(i,k)X_R(i,k) & \frac{1}{L}\sum_{k=(m-1)*L+1}^{m*L} X_R^2(i,k) \end{bmatrix} \quad (17)$$

According to Eq.(17), since the value of sound spectral power and noise variance are positive, we can obtain a unique solution of $P_L(i,m)$, $P_R(i,m)$,$\sigma^2(i,m)$ and IID_m from Eq.(17). Then, $SNR(i,m)$ which means the SNR of the mth sub-band in the ith frame is defined:

$$SNR(i,m)=10\log\left(\frac{P_L+P_R}{2\sigma^2}\right) \quad (18)$$

The $SNR(i,m)$ is compared with the threshold β to obtain the $SNRIndex(i,k)$,which means the SNR index of each frequency bin within the mth sub-band in the ith frame.

$$SNRIndex(i,k)=\begin{cases} 1, & SNR(i,m)\ge\beta \\ 0, & SNR(i,m)<\beta \end{cases},\quad (m-1)L+1\le k\le mL \quad (19)$$

According to $SNRIndex(i,k)$, cross PSD of Eq.(5) is modified as follows:

$$P_{LR}(i,k)=X_L(i,k)*X_R^*(i,k)*SNRIndex(i,k) \quad (20)$$

Based on Eq.(20), the frequency whose sub-band SNR is larger than threshold are utilized to compute the cross PSD. Otherwise, the frequency is neglected. All that means, the frequency which is less affected by the noise is used to estimate the cross PSD.

$R_{xLxR}(i,\tau)$ is calculated by reverse FT of $P_{LR}(i,k)$ through Generalized Cross Correlation (GCC) algorithm:

$$R_{x_Lx_R}(i,\tau)=\sum_{k=0}^{N-1}\frac{P_{LR}(i,k)}{|P_{LR}(i,k)|}\exp(j2\pi nk/N) \quad (21)$$

Then ITD_i is the estimated:

$$ITD_i=\arg\max_{\tau}\left(R_{x_Lx_R}(i,\tau)\right) \quad (22)$$

The ITD_i is calculated through above procedures. Then azimuth estimation based on ITD_i is divided into two steps: the off-line training stage and the testing stage.

Training stage: The azimuth model is set up. The input signal is the binaural signals of known azimuth. After pre-processed, ITD of each frame is computed based on GCC. The ITD is modeled using VQ for each azimuth. In this paper, the convolution result of MIT HRIR and white noise is set as training data for a certain azimuth. MIT HRIR used is the HRIR of KEMAR in the front horizontal plane. The azimuth is uniformly sampled with the steps of 5°. The range of azimuth is in the [-90° 90°]. -90° correspond to a point directly to the left, and 90° corresponds to a point directly to the right.

Testing stage: ITD_i of testing data is estimated based on sub-band SNR estimation algorithm. That is, Eq.(18)-(22) are utilized to compute the ITD_i. azimuth $\overline{\theta_i}$ of each frame is estimated according to Eq.(23):

$$\overline{\theta_i}=\arg\min_{\theta}|ITD_i-ITD_\theta| \quad (23)$$

where ITD_θ represents ITD value of VQ model for azimuth θ.

The structure of binaural SSL based on sub-band SNR estimation is depicted in Figure 1.

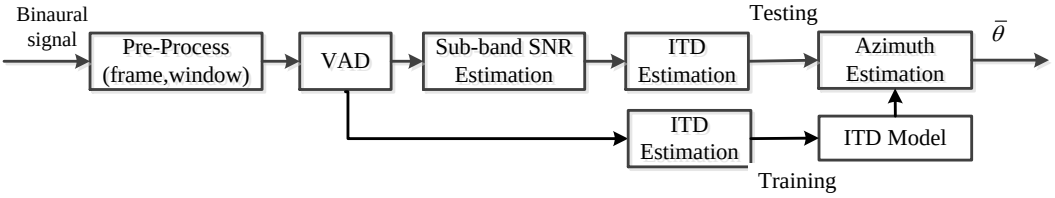


Figure 1.Binaural Sound Source Localization Structure based on Sub-band SNR Estimation

The proposed method adaptively selects the sub-bands with less affected by noise to estimate the sound source direction, which improves the localization accuracy.

In the process of sub-band SNR estimation algorithm, the number of sub-band and value of threshold need to be set. On the one hand, if we increase the number of sub-band, thus the number of frequency bins utilized to compute $R(i,m)$is reduced, which reduces the estimation accuracy of $R(i,m)$. On the other hand, since we assume the IID_m of all frequency within the sub-band is the same in Eq.(13). If we decrease the number of sub-band, then the IID of each frequency bin within one sub-band will have larger difference, which means above assumption does not hold.

The same problem also exists for the threshold selection. If the threshold value is larger, the less frequency is utilized to compute the $P_{LR}(i,k)$, which reduces the estimation accuracy of $P_{LR}(i,k)$. But if the threshold value is too small, the more frequency interfered by noise is utilized to compute the ITD_i, which reduces the estimation accuracy of ITD_i. Thus we will select the appropriate parameters based on the simulation results, in order to obtain reliable sound source localization.

4. Simulation and Result Analysis

4.1. Simulation Condition

Sound sources for the simulation are randomly taken from the CHAINS Speech Corpus. The speech data includes female speech and male speech with mono channel. The sampling frequency is 44.1 kHz, and the duration of sound source is about 1 minute. The mono channel speech is convolved with MIT HRIR in the steps of 10° to create the directional and clean binaural testing sound. Noise signal is white noise. According to the global SNR level (Here global SNR value is 0, 5, 10, 15 and 20dB), the noise is weighted and added to the clean binaural signal to get the noisy testing sound.

The noisy binaural testing sound is framed and windowed. The frame length is 40ms, with the frame shift of 20ms. Each frame is windowed using Hamming window.

Two indexes are used to evaluate the system performance, the percentage of correct localization and Root Mean Square Error (RMSE). The percentage of correct localization is defined in Eq.(24):

$$P = k_c / K \tag{24}$$

where k_c represents the number of correct localization frame; K is the number of frames after VAD. The correct localization is defined that the estimation $\bar{\theta}_i$ lies within the $\pm 5°$ of the true azimuth θ.

RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{K}\sum_{i \in K} |\bar{\theta}_i - \theta|^2} \tag{25}$$

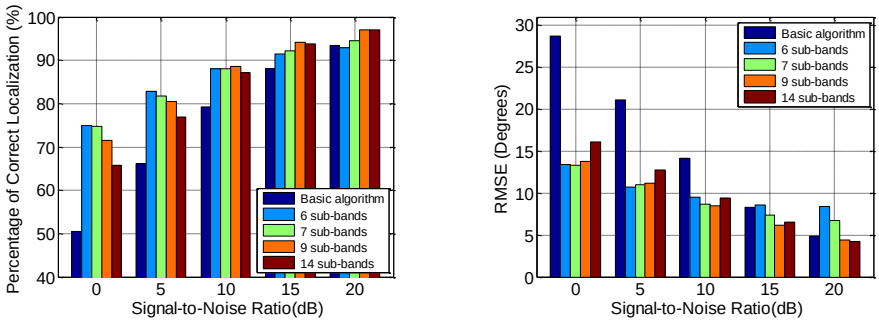
RMSE not only considers the localization error of correct localization frames, but also takes the localization error of all frames after VAD into account. Thereby, RMSE reflects the overall localization accuracy.

The method which utilizes the all frequency to estimate azimuth is called basic algorithm. In this section, we compare the performance of basic algorithm and the proposed algorithm.

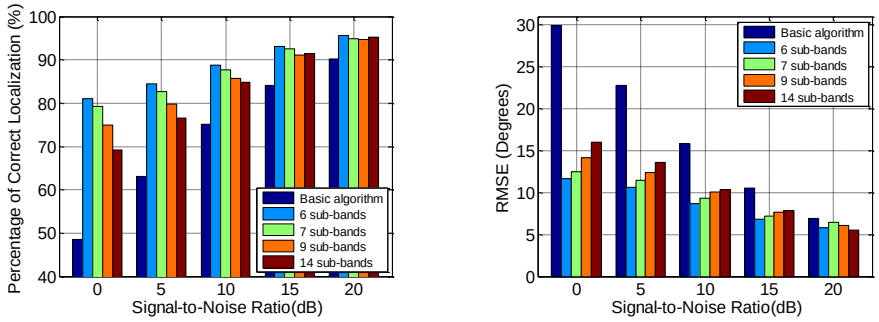
4.2. Simulation 1: Influence of Sub-band Number on Localization Performance

This sub-section focuses on the influence of sub-band number on the localization performance. The number of sub-band is set to 6, 7, 9 and 14. The value of threshold β is set to 0.

The performance of basic algorithm is compared to that of the proposed algorithm with different sub-band number. The results are depicted in Figure 2.



(a) Result for Female Speech



(b) Result for Male Speech

Figure2. Performance Comparisons of Basic Algorithm and the Proposed Algorithm with Different Sub-band Number

In Figure 2, the abscissa represents the global SNR, while ordinate is the performance index. Different bars denote the performance of different algorithms.

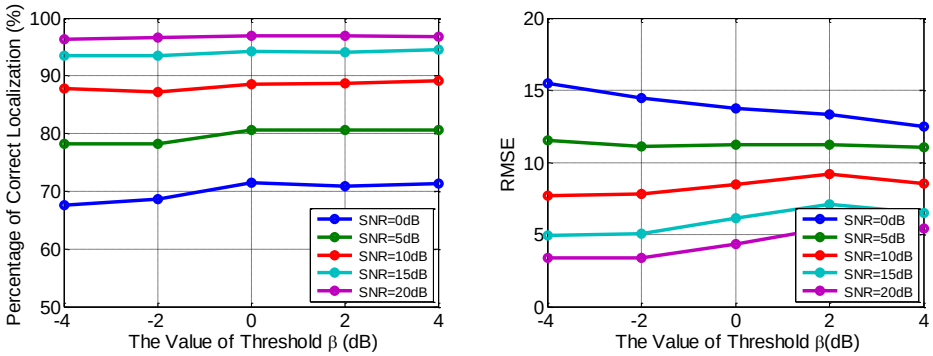
From results, we firstly find that performance of the proposed algorithm is much better than that of basic algorithm. The proposed algorithm has higher percentage of correct localization than basic algorithm, especially in low global SNR. Also, RMSE of the proposed algorithm is much lower than that of basic algorithm. For instance, P of basic algorithm is about 50% in 0 dB, while P of the proposed algorithm exceeds 70%. RMSE of basic algorithm is larger than 30° in 0 dB, while RMSE of the proposed algorithm is less than 16°.

Secondly, the performance of the proposed algorithm varies with the number of sub-band. When the number of sub-band is smaller (such as 6), the proposed algorithm has better performance in the lower global SNR. When the number of sub-band is larger (such as 14), the proposed algorithm has better performance in the higher global SNR. Since the proposed algorithm with 9 sub-bands has a better

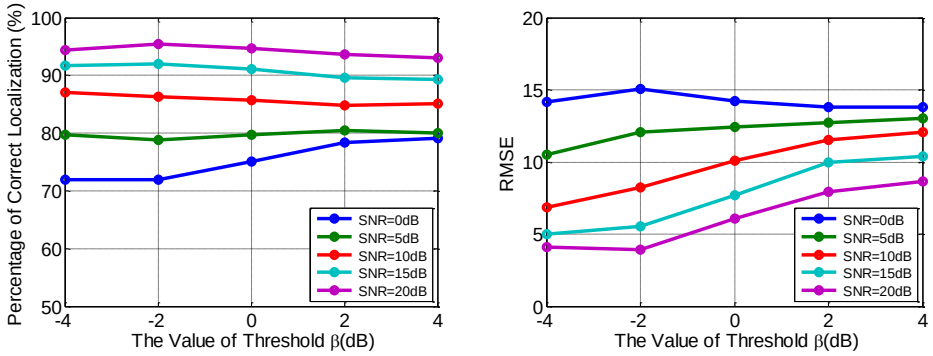
performance both in low global SNR and high global SNR, the number of sub-bands is set to 9.

4.3. Simulation 2: Influence of Threshold Value on Localization Performance

In this sub-section, we observe the performance difference when the threshold is set to -4,-2,-1, 0, 2 and 4. Figure 3 gives the results of the proposed algorithm for female and male speech with different global SNR and different threshold.



(a) Result for Female Speech



(b) Result for Male Speech

Figure3. Performance Comparisons of Basic Algorithm and the Proposed Algorithm with Different Threshold

In Figure 3, the abscissa represents the threshold value. The ordinate is the performance index. Different curves denote the proposed algorithm performance with different global SNR.

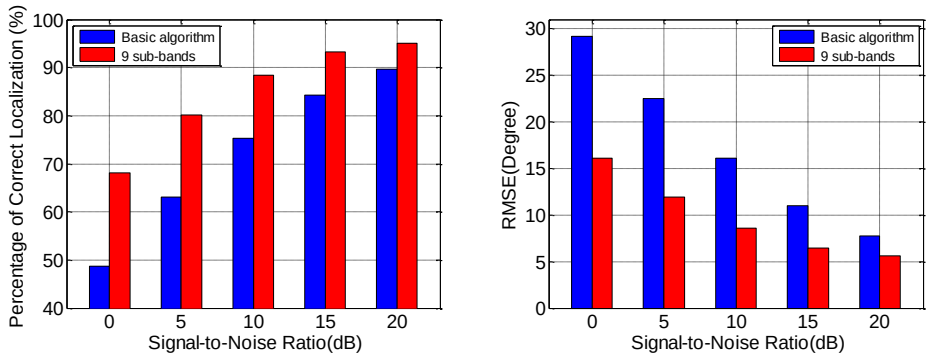
From the results, we find that the performance of the proposed algorithm varies with the value of threshold. In Figure 3(a), for female speech with certain global SNR, the proposed algorithm has the highest percentage of correct localization when the threshold is set to 0, while the proposed algorithm has the lowest RMSE when the threshold is set to -2. Also, for male speech with certain global SNR in Figure 3(b), when threshold is set to -2, the proposed algorithm has better performance. Based on the above simulation results, we set the threshold value of -2.

4.4. Simulation 3: the Proposed Algorithm Performance in Reverberation Environment

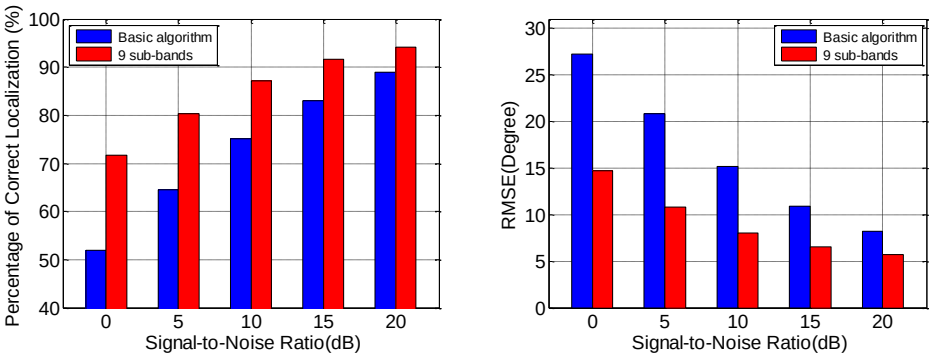
In this sub-section, the reverberation is generated by the Image algorithm with the reverberation time of 0.2s and 0.6s. The number of sub-band is 9, and threshold

value is -2. The clean mono-channel speech is convolved with binaural room impulse response function to derive the reverberation signal. Also, the additive noise is added to the binaural reverberation signal to achieve the required global SNR.

Figure 4 gives the results of basic algorithm and the proposed algorithm when T_{60} is 0.2s. Figure 5 gives the results when T_{60} is 0.6s.

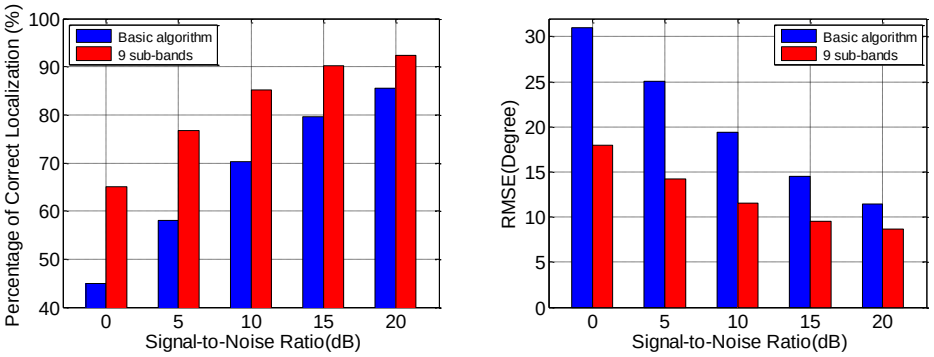


(a) Result for Female Speech



(b) Result for Male Speech

Figure4. Performance Comparisons of Basic Algorithm and the Proposed Algorithm in Reverberation Environment (T_{60} = 0.2s)



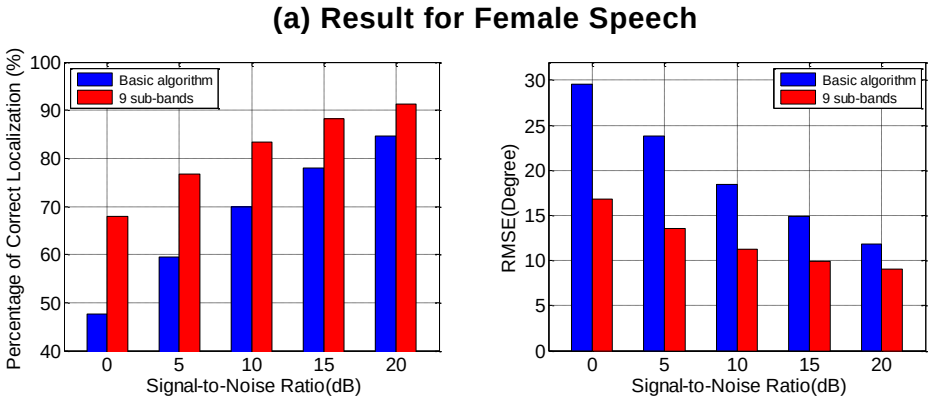


Figure5. Performance Comparison of Basic Algorithm and the Proposed Algorithm in Reverberation Environment (T_{60} = 0.6s)

In Figure 4 and Figure 5, the abscissa represents the global SNR, while ordinate is the performance index.

From the results, we firstly find that the performance of the proposed algorithm is significantly improved, not only in the percentage of correct localization, but also in the RMSE. For example, when the global SNR is 10dB and T_{60} is 0.2s, P of basic algorithm is only 75%, while P of the proposed algorithm is close to 90%. When the global SNR is 15dB and T_{60} is 0.6s, P of basic algorithm is only 80%, while P of proposed algorithm is close to 90%.

Secondly, performance improvement varies with the global SNR. The lower the global SNR, the more obvious performance improvement the proposed algorithm has. For female speech, when the global SNR is 0dB and T_{60} is 0.2s, the improvement of P is about 20%, RMSE reduction is about 13°. But when the global SNR 20 is dB, the improvement of P is only 6%, RMSE reduction is about 2°. When T_{60} is 0.6s and global SNR is 0dB, the improvement of P is about 18%, RMSE reduction is about 10°. But the improvement of P is about 6%, RMSE reduction is about 3° when the global SNR is 20dB.

5. Conclusion

Since binaural SSL aims to imitate the human spatial hearing mechanism to improve the SSL robust with small computation complexity, it becomes an important research topic of SSL. In order to improve the performance of binaural SSL in noisy and reverberation environment, this paper estimates the sub-band SNR based on auto-correlation matrix of binaural signals, and selects the sub-band with the higher SNR to compute the localization cues. Since the proposed algorithm utilizes the reliable frequency to localize the sound source, the performance of the proposed algorithm has been significantly improved, which provides the basis for robust speech segregation and recognition based on localization.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 61201345) and the Beijing Key Laboratory of Advanced Information Science and Network Technology (No. XDXX1308).

References

[1] D. L. Wang and G. J. Brown, "Computational Auditory Scene Analysis: Principles, algorithms and applications", IEEE Press: John Wiley & Sons, Inc., New York **(2005)**.
[2] D. Li and S. Levinson, "A Bayes-rule based hierarchical system for binaural sound source Localization", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, **(2003)** April 6-10, Hong Kong. pp. 521-524.
[3] M. Raspaud, H. Viste and G. Evangelista, "Binaural Source Localization by Joint Estimation of ILD and ITD", IEEE transaction on Audio Speech and Language Processing, vol. 18, no. 1, **(2010)**, pp. 68-77.
[4] Kim Young-IK, "Estimation of Interaural Time Differences Based on Zero-Crossings in Noisy Multisource Environments", IEEE transaction on Audio Speech and language Processing, vol. 15, no. 2, **(2007)**, pp.734-743.
[5] C. Q. Li, S. J. Dai and F. Wu, "Binaural Sound Localization Based on Detection of Multi-band zeros-crossing points", Proceedings of Second International Conference on Intelligent Networks and Intelligent Systems, **(2009)** November 1-3, Tianjin, China, pp. 393-396.
[6] C. Kim, K. Kumar, B. Raj and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain", Proceedings of 10th Annual Conference of the International Speech Communication Association, **(2009)** September 6-10, Brighton, United Kingdom. pp. 2495–2498.
[7] T. May, S. van de Par and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End", IEEE Transactions on Audio Speech and Language Processing, vol. 19, no. 1, **(2011)**, pp. 1-13.
[8] T. Rodemann, G. Ince and F. Joublin, "Using Binaural and Spectral Cues for Azimuth and Elevation Localization", Proceedings of International Conference on Intelligent Robots and Systems, **(2008)** September 22-26, Nice, France, pp. 2185-2190.
[9] J. Woodruff and D. L. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization", IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 7, **(2010)**, pp. 1856-1866.
[10] J. Woodruff and D. L. Wang, "Binaural Detection, Localization and Segregation in Reverberation Environments Based on Joint Pitch and Azimuth Cues", IEEE Transactions on Audio Speech and Language Processing, vol. 21, no. 4, **(2013)**, pp. 806- 815.
[11] N. Roman and D. L. Wang, "Binaural Tracking of Multiple Moving Sources", IEEE Transactions on Audio, Speech and Language Processing, vol. 16, no. 4, **(2008)**, pp. 728-739.
[12] H. M. Park and R. M. Stern, "Spatial Separation of Speech Signals using Amplitude Estimation Based on Interaural Comparisons of Zeros-crossing", Speech Communication, vol. 51, no. 1, **(2009)**, pp. 15-25.
[13] C. Kim, K. Kumar and R. M. Stern, "Binaural sound source separation motivated by auditory processing", In the Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, **(2011)** May 22-27, Prague, Czech Republic, pp. 5072 – 5075.
[14] H. M. Park and R. M. Stern, "Missing Feature Speech Recognition Using Dereverberation and Echo Suppression in Reverberation Environment", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (**2007)** April 15-20, Honolulu, USA, pp. 381-384.
[15] R. Parisi, F. Camoes, M. Scarpiniti and A. Uncini, "A. Cepstrum prefiltering for binaural source localization in reverberant environments", IEEE Signal Processing Letters, vol. 19, no. 2, **(2012)**, pp. 99-102.
[16] T. May, S. van de Par and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation", IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no.7 , **(2012)**, pp. 2016-2030.
[17] Y. Karim, A. Sylvain and Z. Jean-Luc, "A binaural sound source localization method using auditive cues and vision", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, **(2012)** March 25-30, Kyoto, Japan, pp. 217-220.
[18] J. Chen and W. Ser, "Speech detection using microphone array", Electronics Letter, vol. 36, no. 2, **(2008)**, pp. 181-182.

Author

Zhou Lin received the BS and PhD degrees in Signal and Information Processing from School of Information Science and Engineering, Southeast University, Nanjing, China, in 2000 and 2005, respectively. From 2005 to 2009, she was a lecturer with Department of Radio Engineering, Southeast University, China. From 2009 to the present, she is an associate professor with the School of Information Science and Engineering, Southeast University, China. Her research interests include the speech processing and acoustic signal processing, such as spatial hearing, speech recognition and speech separation.

