An adaptive speech separation method based on spherical-regular-tetrahedral-microphone array

Yu Pang[1, 2], Li Pei [2] and Jia-Yu Lin[1]
[1]*College of Electronic Science and Engineering, National University of Defense Technology, Changsha, China*
[2]*The Chinese Armed Police Force 8610, Panjin, China*
E-mail: PangYu_NUDT@163.com
www.nudt.edu.cn

For speech separation method, getting the accurate weight vector is very important for the performance of method. Aiming at how to reduce the signal cancellation, this paper focuses on obtaining the weight vector in the speech separation process accurately, and it designs an adaptive speech separation method based on spherical-regular-tetrahedral-microphone array. The simulation results show that this paper's method can reduce the signal cancellation effectively in speech separation process, so as to have a better separation performance than the similar method base on two-microphone.

Keywords: Speech Separation; Microphone Array; Signal Cancellation.

1. Introduction

Speech separation technology uses a certain method to calculate and obtain the individual sound source signal from the mixing speech signals received. Compared to single microphone, linear microphone array and planar microphone array, speech separation method based on spherical-microphone array method can better obtain the spatial information of the receiving signals and calculate the desired weight vector in response to the power requirements of the methods in different directions, so that the signal in the target direction can be gained, and the signal in the non-target direction can be suppressed[1]. But if the signal cancellation occurs, which means while the speech activities of the sound source is active, the adaptability of the method is also working, the sound source signal will be regarded as a noise and be cancelled in the output[2]. Therefore, in order to get the accurate weight vector, it is need to design a kind of speech separation method which can reduce the cancellation phenomenon and separate the sound source signal accurately.

2. Method Design and Implementation

Consider both execution performance and the equipment cost, we design an adaptive speech separation method based on spherical-regular-tetrahedral-microphone array, of which four array elements are respectively arranged on the four vertices of tetrahedron inscribed in spherical model. The method consists of four main parts, which are the Mode Transition (MT) module, the Speech Activity Classification (SAC) module, the Adaptive Dual-beam Speech Separation (ADSS) module and the Post Check (PC) module. Fig.1 shows the method architecture.

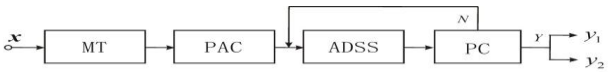

Fig. 1 Speech separation method architecture based on spherical-regular-tetrahedral-microphone array

2.1. *MT module*

The MT module is used to transform the receiving signal from element space to the harmonic space, so that the weight vector and the cross-power spectrum matrix are calculated more concise and efficient.

Let $x(k,r,\Omega)$ be the sound pressure of incident signal, k is the wave number, and $(r,\Omega)=(r,\theta,\varphi)$ is the spatial position of spherical coordinates. The Spherical Fourier transform [1] is

$$x_{vw}(k,r)=\int_{\Omega\in S^2} x(k,r,\Omega)[Y_v^w(\Omega)]^* d\Omega = b_v(kr)[Y_v^w(\Omega_0)]^* \tag{1}$$

Where $x_{vw}(k,r)$ is the spherical harmonic coefficient, $b_v(kr)$ is the modal energy, and $Y_v^w(\Omega)$ is the spherical harmonics of v-order and w-degree. The inverse Spherical Fourier transform is

$$x(k,r,\Omega)=\sum_{v=0}^{\infty}\sum_{w=-v}^{v} x_{vw}(k,r)Y_v^w(\Omega) \tag{2}$$

2.2. *SAC module*

SAC module is used to control the method's adaptability by classifying the speech activity of sound sources, which is the key link to reduce the signal cancellation in the method. Four kinds of speech activity for two sound sources are: only sound source 1 is active (S1); only sound source 2 is active (S2); both two sound sources are active (MIX); both two sound sources are not active (NONE).

For the state of NONE, it can be detected by using some speech endpoint detection methods [3].For the state of S1, S2 and MIX, we use beamformers

$\{w_1, w_2\}$, where w_i is the beamformer targeting sound source $i(i=1,2)$. The output processed by $\{w_1, w_2\}$ is [1]

$$\hat{y}_n = \sum_{v=0}^{1} \sum_{w=-v}^{v} w_{vw,n}(k) x_{vw}(k) = \boldsymbol{w}_{vw,n}^H \boldsymbol{x}_{vw}, \ (n=1,2) \tag{3}$$

Notice that $\{\hat{y}_1, \hat{y}_2\}$ is equivalent to be restored from the two sound source signals $\{s_1, s_2\}$. Therefore, the power of $\{\hat{y}_1, \hat{y}_2\}$, which is calculated more easier, reflects the power of $\{s_1, s_2\}$ in a certain extent. Logarithmic of the power ratio of $\{\hat{y}_1, \hat{y}_2\}$ is named logarithmic-of-beamformer-power-ratio (LogBPR) and is defined as follows

$$r(k,L) \square \log\left(\frac{\sum_{l\in L}\left|\hat{y}_1(k,l)\right|^2}{\sum_{l\in L}\left|\hat{y}_2(k,l)\right|^2}\right) \tag{4}$$

Where k is the window length, L is the frames of the fragment. It is observed that the LogBPR is positive in the state of S1, the LogBPR is negative in the state of S2, and the LogBPR fluctuates around 0 in the state of MIX. And the states of S1, S2 and MIX can be classified more easier by being simplified to S1-MIX and S2-MIX two sub-classification cases for the supports of S1 LogBPR and S2 LogBPR are well separated[4][5]

$$\begin{cases} \mathrm{S1} & r(k_1,L) \ge \theta_1 \\ \mathrm{S2} & r(k_2,L) \le \theta_2 \\ \mathrm{MIX} & otherwise \end{cases} \tag{5}$$

Tran has provided the method of finding the optimal window value k and the threshold value θ. And the LogBPR in harmonic space also has a similar approximate Gauss distribution, as it is shown in Fig.2.

We can find the optimal window value k_1, k_2, and the threshold θ_1, θ_2 by a similar procedure of Tran's method. Fig.3 shows the distribution of LogBPR in harmonic space.

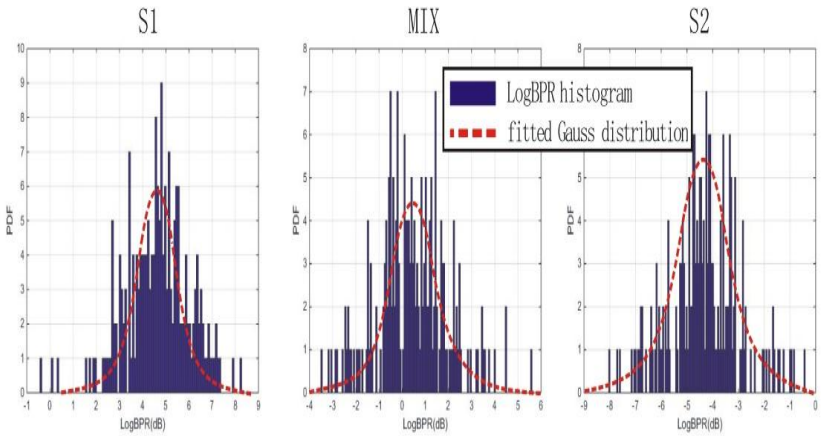

Fig. 2 LogBPR histogram and fitted Gauss distribution of each state in harmonic space

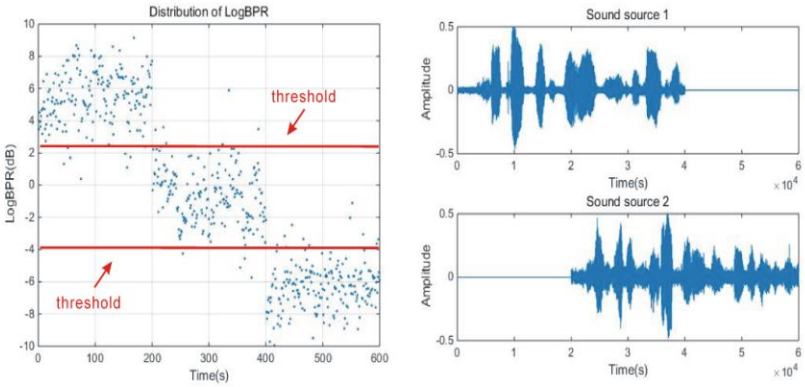

Fig. 3 Distribution of LogBPR and reference signal in harmonic space

2.3. *ADSS module*

ADSS module is used to calculate the time-delay compensation and the amplitude weights ,which make the beam in each dimension has the same main lobe width in the harmonic space and generate the corresponding FIR filter bank according to the state of adaptability of method controlled by the classification results of SAC module. So that, the mixed speech signals can be separated accurately in continuous phase condition [6].

For sound source 1, we use J -order FIR filter to divide the harmonic receiving signal into J slices on the time t_0, and the $j_{st}(j=0,1,...,J-1)$ slice is[1][6]

$$\begin{aligned}\boldsymbol{x}_j &= [x_{00}(t_0 - jT_s), x_{1-1}(t_0 - jT_s), x_{10}(t_0 - jT_s), x_{11}(t_0 - jT_s)]^T \\ &= [x_{00,j}, x_{1-1,j}, x_{10,j}, x_{11,j}]^T\end{aligned} \tag{6}$$

Where the T_s is the time interval of two adjacent slices. The element of $\boldsymbol{x}_j$ is

$$x_{vw}(t_0 - jT_s) = \sum_{m=1}^{4} \alpha_m x_m(t_0 - jT_s)[Y_v^w(\Omega_m)]^* \tag{7}$$

The manifold vector of the j_{st} slice is

$$\boldsymbol{a}_j = [a_{00,j}, a_{1-1,j}, a_{10,j}, a_{11,j}]^T \tag{8}$$

The element of $\boldsymbol{a}_j$ is

$$\boldsymbol{a}_{vw,j} = b_v^s(kr, kr_s)[Y_v^w(\Omega_0)]^* \tag{9}$$

The weighting of the j_{st} slice is

$$\boldsymbol{w}_j = [w_{00,j}, w_{1-1,j}, w_{10,j}, w_{11,j}]^T \tag{10}$$

Combining with MVDR algorithm, the element of $\boldsymbol{w}_j$ is [7]

$$\begin{aligned}\boldsymbol{w}_0(k) &= \boldsymbol{w}_0(k-1) - \Delta_0(k) y_P(k) \boldsymbol{x}_0(k) \\ &\quad - \frac{1}{N}\mathbf{1}\mathbf{1}^T \times [\boldsymbol{w}_0(k-1) - \Delta_0(k) y_P(k) \boldsymbol{x}_0(k)] + \frac{1}{N}\mathbf{1}\end{aligned} \tag{11}$$

$$\begin{aligned}\boldsymbol{w}_{vw}(k) &= \boldsymbol{w}_{vw}(k-1) - \Delta_{vw}(k) y_P(k) \boldsymbol{x}_{vw}(k) \\ &\quad - \frac{1}{N}\mathbf{1}\mathbf{1}^T \times [\boldsymbol{w}_{vw}(k-1) - \Delta_{vw}(k) y_P(k) \boldsymbol{x}_{vw}(k)]\end{aligned} \tag{12}$$

The output signal of sound source 1 separated by the ADSS module on the time t_0 is

$$y = \sum_{v=0}^{V} \sum_{w=-v}^{v} w_{vw}(k) x_{vw}(k) = \boldsymbol{w}_{vw}^H \boldsymbol{x}_{vw} \tag{13}$$

The output signal of sound source 2 can be obtained by a similar procedure. The same processing can be carried out for the receiving signal of each time, so that the mixing speech signal can be separated effectively by the ADSS module.

2.4. *PC module*

In the sub-classification of S1-MIX, if the decision error occurred of SAC module, there will be two results: (1) the decision is S1 while the adaptability of w_1 is open, which is caused by "miss" and can lead a severe influence on the result. (2) The decision is not S1 while the adaptability of w_1 is off, which is caused by "false alarm" but lead a teeny influence [9]. The PC module is just used to keep the method with low miss rate, while not to bring a substantial increase in false alarm rate by calculating the output power of the ADSS module, so that the result of the SAC module is modified and improved. In harmonic

space, we have proved a same effect of the method as in element space. The desired output signal power is [1][8].

$$P = E\{|y|^2\} = E\{|\boldsymbol{w}^H \boldsymbol{x}|^2\} = \boldsymbol{w}^H E\{|\boldsymbol{x}\boldsymbol{x}^H|^2\}\boldsymbol{w} = \boldsymbol{w}^H R\boldsymbol{w} = \frac{1}{\boldsymbol{v}^H R^{-1}\boldsymbol{v}} \tag{14}$$

Where $\boldsymbol{v}$ is the unit vector of the target direction. R is the covariance matrix of receiving signal. When w_1 is open accurately, the output signal power is.

$$P_{true} = \frac{1}{v_{vw}^H R_{I,N}^{-1} v_{vw}} \tag{15}$$

Where $R_{I,N}$ is the covariance matrix of the sum signal of sound source 2, and the interference plus noise. When w_1 is open inaccurately, the output signal power is.

$$P_{false} = \frac{1}{v_{vw}^H (R_1 + R_{I,N})^{-1} v_{vw}} \tag{16}$$

Where R_1 is the covariance matrix of the sound source 1. Here assumes no correlation between R_1 and $R_{I,N}$. Let

$$\Gamma(v_{vw}, R) \square\ v_{vw}^H R^{-1} v_{vw} \tag{17}$$

Combining Eqs (15), (16) and (17), we have

$$\begin{aligned} \Gamma(v_{vw}, R_1 + R_{I,N}) &= v_{vw}^H (R_1 + R_{I,N})^{-1} v_{vw} = v_{vw}^H (R_{I,N}^{-1} - C) v_{vw} \\ &= \Gamma(v_{vw}, R_{I,N}) - v_{vw}^H C v_{vw} \end{aligned} \tag{18}$$

Where $C = R_{I,N}^{-1}(R_{I,N}^{-1} + R_1^{-1})^{-1} R_{I,N}^{-1}$, and $v_{vw}^H C v_{vw} > 0$ to any v_{vw} [9]. Hence

$$\Gamma(v_{vw}, R_1 + R_{I,N}) < \Gamma(v_{vw}, R_{I,N}) \tag{19}$$

This conclusion points that the $\Gamma(v_{vw}, R)$ in the state of the adaptability opened inaccurately is smaller than in the state of the adaptability opened accurately. Therefore, a threshold can be used to improve the result of the SAC module. If the $\Gamma(v_{vw}, R)$ of the fragment is lower than the threshold, the SAC classification result is modified to MIX. There is a similar checking method for S2-MIX. Fig.4 shows the distribution of $\Gamma(v_{vw}, R)$ in harmonic space.

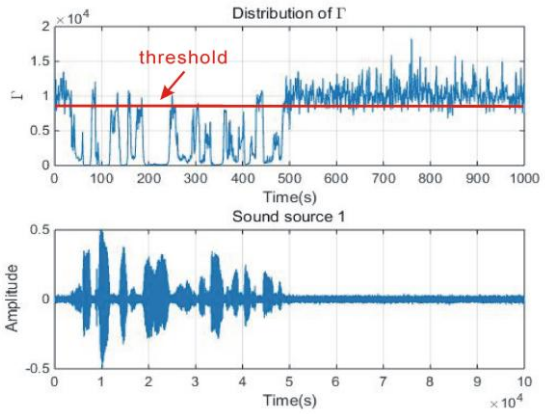

Fig. 4 Distribution of $\Gamma(v_{vw}, R)$ and reference signal in harmonic space

3. Simulation and Experiment

5 groups of sentences of equal length (approximately 3s) randomly from TIMIT[10] speech sample library is selected as the test samples, and sampled by 8kHz and 16bit. In order to avoid signals aliasing, half wavelength (approximately 10cm) of the center speech frequency (1700Hz) is selected as the space distance of the adjacent elements of the spherical-regular-tetrahedral-microphone array. Two sound sources are respectively arranged at points of both distances of 50cm, both azimuth angles of 45°, and pitching angles of ±60° from the array center. Generate the mixing signals received of four array elements according to spatial signal model [1]. And add Gauss white noise of 20dB to the signals above. Being compared with the similar method based on two-microphone[9], this paper's method is evaluated by evaluating the sample speech signals before and after mixing and separation with perceptual evaluation of speech quality (PESQ) standard[11]. Fig.5 shows an example of this paper's method for a group of sample sentences. Tab.1 shows the PESQ values of the two methods.

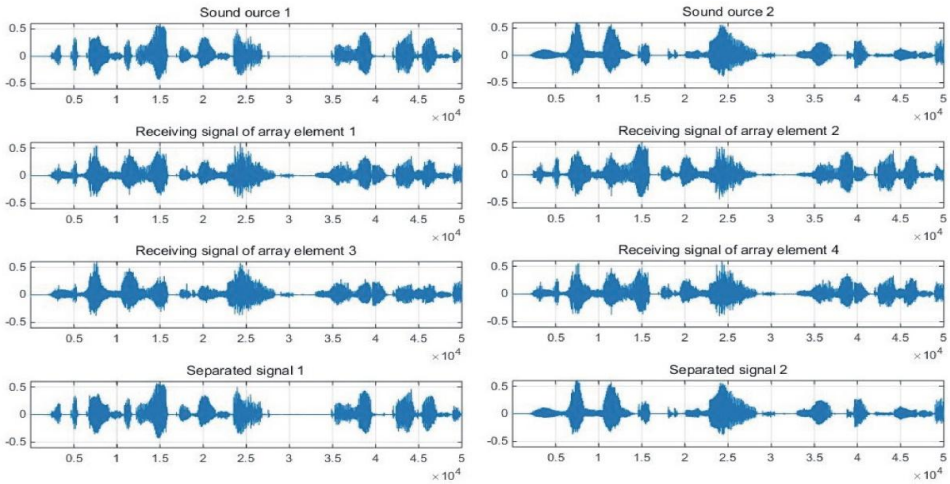

Fig. 5 An example of this paper's method for a group of sample sentences

Tab. 1 The PESQ values of the two methods

Sound source ID	PESQ value before and after mixing	PESQ value before and after separation
Method based on two-microphone		
1	1.875	2.783
2	1.970	2.309
3	1.276	2.270
4	1.817	2.336
5	1.312	2.603
6	2.017	2.739
7	1.911	2.929
8	2.228	3.183
9	1.962	3.037
10	2.035	2.483
Mean Value	1.840	2.667
Standard Deviation	0.095	0.103
Method based on spherical-regular-tetrahedral-microphone array		
1	1.957	3.212
2	1.791	2.976
3	1.312	3.262
4	1.633	3.178
5	2.014	3.036
6	2.008	2.984
7	1.838	3.073
8	2.032	3.179
9	2.058	3.216
10	2.005	2.899
Mean Value	1.865	3.102
Standard Deviation	0.056	0.015

In Tab. 1, the PESQ mean values and standard deviations of the two speech separation methods before and after mixing are both small and similar, which shows that the mixing degrees of the two methods' receiving signal are roughly the same. Comparing the PESQ mean values and standard deviations of the two speech separation methods before and after separation, we can see the mean values of this paper's method is higher and standard deviation is smaller than the method based on two-microphone, which shows that the separation performance of this paper's method is more optimal, and the stability is better.

4. Conclusion

In this paper, we design and implement an adaptive speech separation method based on spherical-regular-tetrahedral-microphone array. The method separates mixing speech effectively by using MT module, PAC module, ADSS module and PC module. Simulation results show that this paper's method can obtain better separation performance than the similar method based on two-microphone array for two sound sources under noisy environment.

References

1. Rafaely, Boaz, et al. Spherical Microphone Array Beamforming. Speech Processing in Modern Communication. Springer Berlin Heidelberg, 2010:281-305.
2. Cox, Henry. "Resolving power and sensitivity to mismatch of optimum array processors." Journal of the Acoustical Society of America 54.3(1973):771-785.
3. Song, Qian Qian, and Y. U. Feng-Qin. "Speech Endpoint Detection Based on EMD and Improved Double Threshold Method." Audio Engineering (2009).
4. Tran, T. N., W. Cowley, and A. Pollok. "Voice activity classification using beamformer-output-ratio." *Communications Theory Workshop (AusCTW), 2012 Australian* IEEE, 2012:96-101.
5. Cho, Namgook, and E. K. Kim. "Enhanced voice activity detection using acoustic event detection and classification." IEEE Transactions on Consumer Electronics 57.1(2011):196-202.
6. Uthansakul, M., and M. E. Bialkowski. "Fully spatial wide-band beamforming using a rectangular array of planar monopoles." IEEE Transactions on Antennas & Propagation 54.2(2006):527-533.

7. Bucris, Y., I. Cohen, and M. A. Doron. "Robust focusing for wideband MVDR beamforming." Sensor Array and Multichannel Signal Processing Workshop (SAM), 2010 IEEE IEEE, 2010:1-4.
8. Li, Jian, P. Stoica, and Z. Wang. "On robust Capon beamforming and diagonal loading." IEEE Transactions on Signal Processing 51.7(2003):1702-1715.
9. Tran, Thuy Ngoc, W. Cowley, and A. Pollok. "Automatic adaptive speech separation using beamformer-output-ratio for voice activity classification." Signal Processing 113.2(2015):259-272.
10. Harte, Naomi, and E. Gillen. "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech." IEEE Transactions on Multimedia 17.5(2015):603-615.
11. Rix, A. W., et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings 2001:749-752.