Blind Separation of Single Channel Mixed Speech Signals

Dongxu Han
School of Computer Science and Engineering
Nanjing University of Science and Technology,
Nanjing, China (People's Republic of)
e-mail: tunghsu@outlook.com

Zheyuan Fu
School of Computer Science and Engineering
Nanjing University of Science and Technology,
Nanjing, China (People's Republic of)
e-mail: zheyuanfu@gmail.com

Abstract—Current signal separation of single channel mixed speech principally depends on the training samples of the source, for this problem, a new method of blind separation is proposed. According to distribution of different source vary in band, we use SVD method recombine the single channel of dual-source mixed signals as double channels, in that case, an complex underdetermined problem is translated into a simple even-determined problem, and ultimately, the source signals can be separated by the informax algorithm. The simulation results show that the method do not require more prior knowledge of the source, and which have a better separation performance for the mixed source signal partial overlapped in band. The method can separate the single channel of dual-source mixed speech signals blindly. The study provides important theoretical significance and application value.

Keywords—blind separation; single channel; informax; SVD

I. INTRODUCTION

Separation and identification of mixed speech signals is a major difficulty for the machine hearing, and it is also a hot topic in the field of artificial intelligence. In reality, people can easily extract their concerned and interested sound in noisy environment, that is, to separate the desired source signals from the mixed speech signals. Above all, how to realize the human capability by computer, which is called as cocktail party problem in the field of signal processing[1].

According to the relationship between the number of mixed signal (M) and source signals (N), the cocktail party problem can be divided into three parts, that are the over-determined(M> N), the even-determined (M = N) and the underdetermined(M <N). Blind Source Separation (BSS) aimed at the over-determined and the even-determined, which generally need less assumption about the source signal and can be separated. While the underdetermined has became a major challenge for current speech separation, especially for single channel speech separation, which has already attracted wide attention of researchers in recent years. For such underdetermined problems, it need obtain the full of the source signal and cognitive assumptions to achieve the source separation. The usual procedure is to use the sparse characteristics of the speech signal to separate the signal sparse, which is called sparse component analysis (SCA). In references [2-5], it is based on the theory of sparse decomposition to separate single channel mixed speech, which firstly construct a joint dictionary composed by a over complete dictionary related to the speakers, then use optimization algorithms reconstruct the separate the speech signals. In references [2], it constructs the dictionary based on the space non-negative matrix factorization (SNMF). In references [3], it constructs the dictionary spectrum by using the Norm optimization algorithm in the domain of mel spectrum. In references [4], it chooses a normalized amplitude spectrum of typical speaker's speech as codebook. In references [5], it implements sparse decomposition for the mixed-signal in the domain of KLT, and then use autocorrelation matrix of certain training speech signals as the template. The autocorrelation matrix of each source signal is matched by the method of OMP. Finally, the single-channel speech signals are separated by the l0-norm optimization algorithm. The method provided above, to some extent, can resolve the problem of single channel speech separation, however, it extremely relies on the training samples of the source signals and the complete dictionary constructing and template matching is considerable complex. After all, it unsuited for real-time processing and also don't equip the characteristic of blind processing.

In reality, different speech signals has obvious different energy distribution within a certain band, for example, the energy of the male's speech displays more component at low-frequency, though the female' high-frequency component is relatively strong. Instrument such as drums, electric bass, etc. sound mainly in the band of low frequency, but cello, piano can performance a relatively higher pitch. Based on the differences distributed in the characteristics of different bands of source, it regarded single-channel mixed speech signal source as the object of the study, and a new method signal separation is proposed. It resolves the single channel of dual-source mixed signal by the method of dichotomous SVD, thereby a complex underdetermined problem is translated into a simplified even-determined problem. And then the mixed signal is separated by the informax algorithm finally. Throughout the process of separation, it does not require any training samples of the sources. Simulation results show the effectiveness of the method for blind separation.

II. DICHOTOMOUS SVD METHOD

In 2010, Zhao Xuezhi proposed the theory of multi-resolution SVD packets [6], which decomposes signals

into two parts of low and high frequency by singular value decomposition theory, similar to the way of wavelet packet decomposition. Dichotomous SVD is a special form of the singular value decomposition. According to the SVD decomposition theory, for any matrix $A \in R^{m \times n}$, there always exists two orthogonal matrix $U = (u_1, u_2, \mathrm{L}\ , u_m) \in R^{m \times m}$ and $V = (v_1, v_2, \mathrm{L}\ , v_n) \in R^{n \times n}$ which produce a equation as follows,

$$A = USV^T \quad (2.1)$$

In which, the matrix $S = (diag(\sigma_1, \sigma_2, \mathrm{L}\ , \sigma_q), O)$, $S \in R^{m \times n}$, $q = \min(m, n)$, $\sigma_1 \geq \sigma_2 \geq \mathrm{L}\ \geq \sigma_q \geq 0$, σ_i is the singular value. Dichotomous SVD method makes blocks of signa $X = (x_1, x_2, \mathrm{L}\ x_N)$ and rearrange as a matrix by two rows,

$$A = \begin{bmatrix} x_1 & x_2 & \mathrm{L} & x_{N-1} \\ x_2 & x_3 & \mathrm{L} & x_N \end{bmatrix} \quad (2.2)$$

After the process of SVD, it just obtain two singular values. We can obtain a approximation of the original signal refer to the larger singular value, and obtain the detail signal by the smaller singular value, in other word, original signal is decomposed into a approximately component dominant by low frequency and a detail component dominant by high frequency. Go on the procedures of decomposition layer and layer, it can reach the purpose of multi-resolution. Suppose the original signal just processed of one layer using dichotomous SVD method, then we can get two matrix space of sub-signals,

$$\begin{cases} A_h = \sigma_1 U_h \times V_h^T \\ A_d = \sigma_2 U_d \times V_d^T \end{cases} \quad (2.3)$$

where U_h and V_h^T are respectively the low-frequency approximated component of the left and right singular vectors, U_d and V_d^T are respectively the details of the high frequency component of the left and right singular vectors. Singular values σ_1 and σ_2 indicate the energy of each component. Literature [6] shows that, for the singular values normal of two pure signal , there always exists $\sigma_1 ?\ \sigma_2 \approx \varepsilon$ (two values of the white noise singular exists σ_1 ; σ_2), this is because of the upper and lower rows of the signal matrix is highly correlated while the noise matrix is not. Furthermore, the low frequency component of A_h indicates the approximately profile of the original signal and the high frequency component of A_d indicates the detail. Fig .1 shows the results of a certain period of laughter processed by dichotomous SVD. Exam the spectrum of each decomposed signal, we find that the low frequency component of the approximately is dominant, the high frequency component of the detail is dominant identically. Energy of the detail is much less than the approximately. The decomposition signals remain the phase unchanged compared to the source signal, and using that signals can quickly reconstruct the original source with minimal energy loss.

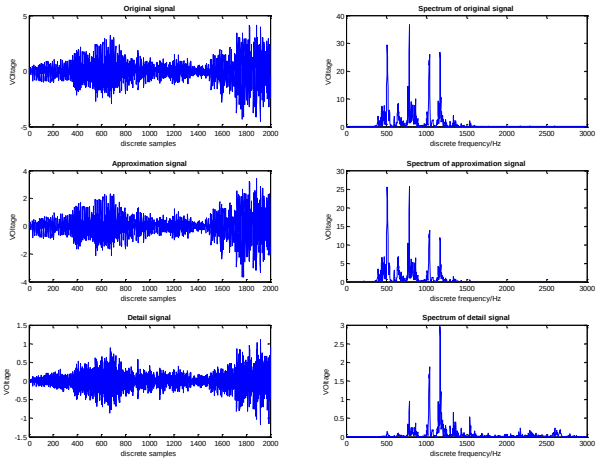

Figure1. Results of dichotomous SVD for laughter voice

From above analysis, we can draw a conclusion. After dichotomous SVD process of the mixed speech, two sub-signals is obtained, and the waveform of these signal is very similar with the source, though there are significant differences in the distribution of energy within same band, i.e., a single channel of dual-source mixed speech processed by the dichotomous SVD, it can be recombine as a two channels mixed signals with different method of combination, thereby, for a single channel mixed-signal vary in band, an underdetermined problem can be translated to the underdetermined after above treatment, thus the difficulty of single channel mixed signal separation is greatly decreased.

III. INFORMAX(INFORMATION MAXIMIZATION)

Under the conditions of the source signal and input channel parameters unknown, according to the statistical characteristics of the input signal separate each individual ingredient of the source signal by the observed signal detection, which called blind source separation[7], or independent Component Analysis (ICA). The basic idea of the method assumes that all of the source signals are independent, which can compensate the lack of prior information of channel. In many cases, the assumption of the statistical independence is possible. According the mechanism of mixture, blind source separation model in mathematical can be divided into two mixture model, that is the instantaneous and the convolved. Linear instantaneous model is simple and principle in blind source separation model, it can be expressed as,

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{n}(t) \quad (3.1)$$

Where , $\boldsymbol{s}(t) = [s_1(t), s_2(t), \mathrm{L}\ , s_n(t)]^T$ is a n-dimensional source signal vector, $\boldsymbol{x}(t) = [x_1(t), x_2(t), \mathrm{L}\ , x_m(t)]^T$ is a m-dimensional vector of observed signals, m × n-dimensional matrix called the mixed matrix whose elements represent

the signal mixed case, $\boldsymbol{n}(t)=\left[n_1(t), n_2(t), \mathrm{L}\ , n_m(t)\right]^T$ is a m-dimensional noise.

The purpose of the blind source separation is assumption that the source signals and the mixing matrix $\boldsymbol{A}$ is unknown. Without regard of noise $\boldsymbol{n}(t)$, only estimate separation matrix $\boldsymbol{W}$ from the observed data vector $\boldsymbol{x}(t)$, we can get

$$\boldsymbol{y}(t)=\boldsymbol{W}\boldsymbol{x}(t)=\boldsymbol{W}\boldsymbol{A}\boldsymbol{s}(t) \quad (3.2)$$

Where $\boldsymbol{y}(t)=\left[y_1(t), y_2(t), \mathrm{L}\ , y_n(t)\right]^T$ is a n-dimensional isolated signal vector, order $\boldsymbol{WA}=\boldsymbol{P}$, If each row and each column of the matrix $\boldsymbol{P}$ have single element of greater than the others, it is able to achieve the purpose of recovered source signal.

In 1995, Bell and Sejnowski[8] proposed the maximum entropy method based on the Infomax, whose main idea is to put separation signal through a nonlinear function (instead of an estimate of the higher-order statistics) and consider that we obtain approximate values of the separating matrix when the output signal information is maximized. Assuming that $\boldsymbol{y}$ is the output signal of a non-linear network, i.e.

$$\boldsymbol{y}=f(\boldsymbol{u})=f(\boldsymbol{W}\boldsymbol{x}) \quad (3.3)$$

Multivariate probability density function of network output is $p_y(\boldsymbol{y})=p_x(\boldsymbol{x})\big/\left|J_f\right|$, where $\left|J_f\right|$ is absolute of the transformation function of the Jacobian determinant

$$\left|J_f\right|=(\det(\boldsymbol{W}))\prod_{i=1}^{N}\frac{dy_i}{du_i} \quad (3.4)$$

In order to make the output information maximize, considering the mutual information between the input and the output, the relationship can be described by equation (3.5),

$$I(\boldsymbol{y};\boldsymbol{x})=H(\boldsymbol{y})-H(\boldsymbol{y}\mid\boldsymbol{x}) \quad (3.5)$$

the gradient of the mutual information about the networks weight is

$$\frac{d}{d\boldsymbol{W}}I(\boldsymbol{y};\boldsymbol{x})=\frac{d}{d\boldsymbol{W}}(H(\boldsymbol{y})-H(\boldsymbol{y}\mid\boldsymbol{x}))=\frac{d}{d\boldsymbol{W}}H(\boldsymbol{y}) \quad (3.6)$$

In the absence of noise, the output conditional entropy diverges to infinity. Therefore, in order to acquire maximum output of information via network, only make the output of the entropy maximize

$$\begin{aligned}\frac{d}{d\boldsymbol{W}}H(\boldsymbol{y})&=\frac{d}{d\boldsymbol{W}}(-E\left\{\log(p_y(\boldsymbol{y}))\right\})\\&=\frac{d}{d\boldsymbol{W}}-E\left\{\log(p_x(\boldsymbol{x}))\right\}+E\left\{\log(\left|\boldsymbol{J}_f\right|)\right\}\\&=(W^T)^{-1}+\frac{d}{d\boldsymbol{W}}\log(\prod_{i=1}^{N}\frac{dy_i}{du_i})\end{aligned} \quad (3.7)$$

Rightmost items should be considered one by one, and it depends on the form of nonlinear network, the nonlinear function is

$$\boldsymbol{y}=\tan h(\boldsymbol{W}\boldsymbol{x}) \quad (3.8)$$

Then

$$\frac{d}{d\boldsymbol{W}}H(\boldsymbol{y})=(\boldsymbol{W}^T)^{-1}-2\tanh(\boldsymbol{y})\boldsymbol{x}^T \quad (3.9)$$

Using statistical gradient, the weight adjusting algorithm can write as discrete form ,

$$\boldsymbol{W}_{n+1}=\boldsymbol{W}_n+\mu_n(\boldsymbol{W}_n^{-T}-2\tan h(\boldsymbol{y}_n)\boldsymbol{x}_n^{\ T}) \quad (3.10)$$

Amari.S[9] and Cardoso[10] ects. who perfected the Infomax algorithm, using the natural gradient or relative gradient to replace conventional gradient descent methods which improved the convergence speed, but these algorithms are only limited to separate super-Gaussian signal. Girolami and Fyfe proposed extended informax[11], which can resolve super-Gaussian signal and sub-Gaussian signal simultaneously, the weight adjusting algorithm of separation matrix $\boldsymbol{W}$ is wrote as follows

$$\Delta\boldsymbol{W}\propto\mu[\boldsymbol{I}-K\tan h(\boldsymbol{y})\boldsymbol{y}^T-\boldsymbol{y}\boldsymbol{y}^T]\boldsymbol{W} \quad (3.11)$$

Where, $K=diag(sign(\kappa(y_i)))$ is called switch Function, $\kappa(y_i)$ is the kurtosis of y_i, $sign()$ is a symbolic function, $diag()$ is a function of diagonal matrix. Kurtosis is used to measure non-Gaussian signal in ICA, for a zero-mean random variable x, kurtosis is defined as

$$\kappa(x)=E\left\{x^4\right\}-3(E\left\{x^2\right\})^2 \quad (3.12)$$

The size of kurtosis can be either positive or negative, random variables with negative kurtosis is called sub-Gaussian, which display the characteristic of flat probability density function, while having positive kurtosis of the random variable is called super-Gaussian distribution, which shows the characteristic as sharp. For a super-Gaussian random variable, assuming the nonlinear function is $g(y)=\tan h(y)$, then the probability density function match it [12],

$$p_s(y)=g'(y)=1-\tan h^2(y) \quad (3.13)$$

Therefore, if estimate the probability density function of the source signal, it can gradually update the weight vector with (3.11) and (3.12), when the output maximize, we can obtain the final approximation of the separation matrix $\boldsymbol{W}$, and the blind signal separation achieved.

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

IV. EVALUATION INDEX OF BLIND SEPARATION

Establish an objective evaluation of the signal separation is necessary to the performance of blind separation algorithm, the evaluation criteria commonly used in the simulation environment including global coefficient matrix row elements advantage index, mutual channel interference measurement error index, similarity index of signal waveform and signal to noise ratio of the separated signals[13]. In view of convenience, it is often regard the cross-correlation between separated signals and the source signals as a separation index to evaluate the separation performance

$$\zeta_{ij} = \zeta(y_i, s_j) = \frac{\sum_{k=1}^{LN} y_i(k) s_j(k)}{\sqrt{\sum_{k=1}^{LN} y_i^2(k) \sum_{k=1}^{LN} s_j^2(k)}} \quad (4.1)$$

For the single channel of dual-source signal blind separation problem in this paper, there is a high similarity between the source signal and separated signals processed by dichotomous SVD. In order to better compare the performance of blind separation after informax procession, the cross-correlation between the separated signals are also regarded as a reference index

$$\eta_{ij} = \eta(y_i, y_j) = \frac{\sum_{k=1}^{LN} y_i(k) y_j(k)}{\sqrt{\sum_{k=1}^{LN} y_i^2(k) \sum_{k=1}^{LN} y_j^2(k)}} \quad (4.2)$$

Evaluate the performance of blind separation by joint examination correlation coefficient in equation(4.1) and equation(4.2), the larger the cross-correlation of the separated signal corresponding to respective source signals, meanwhile the smaller of which calculated in source signals, the better the performance of blind separation.

1) them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig .1", even at the beginning of a sentence.

V. SIMULATION AND PERFORMANCE

In order to verify effectiveness of the single channel of dual-source mixed speech signal blind separation proposed in this paper, numerical simulation is conducted by using MATLAB Voice Box, the source signals are selected a period of gong and train which intercepted from the voice are both super-Gaussian signal aliasing band about 550Hz. Data sampling is 2000 point per second, the mixed signal are equably linear superposition of above signals, as shown in Fig .2.

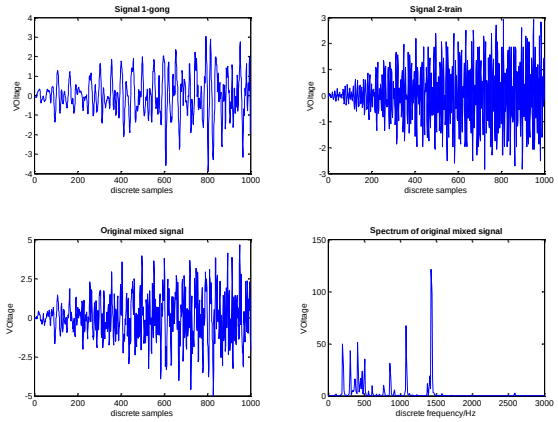

Figure 2. Pure voice signal and linear superposition

First, the mixed signal of dual-source $s(t)$ is processed by dichotomous SVD computation and obtain the detail signal $s_d(t)$ and approximate signal $s_h(t)$, then the $s_d(t)$ and $s_h(t)$ amplitude scale normalized and composed the mixed signal vector $\boldsymbol{x}(t) = [s_d(t) \quad s_h(t)]'$. Finally, the mixed signal is separated by using informax algorithm. The weight adjusting algorithm selected as equation(3.8) , search step is 0.1, the initial separation matrix $\boldsymbol{W}_0$ is two rows two line random in the range of 0～1.

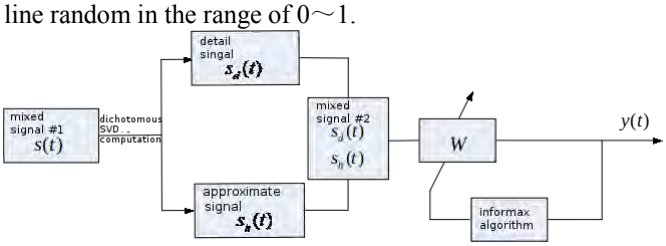

Figure 3. Block diagram of blind signal separation

Fig .4 shows the results of equably linear superposition signal blind separation by dichotomous SVD. The above four figures are comparison chart in time- domain by using the information maximization algorithm before and after the separation, the following four figures are the comparison in frequency -domain. It can be found, in fact, after the SVD, the double channels mixed signals have already initially separated the mixed source signal, continue decompose by using the information maximization algorithm, the final signals are more similar with the corresponding source signals in frequency-domain. Evaluate the similarity index of Fig .5 and Fig .6, the value of cross-correlation coefficients of signals computed before and after the blind signal separation respectively are 0.603 and 0.23. The correlation coefficients between gong and train and their original signal are 0.93 and 0.92 after separation, which shows a well performance of signal separation.

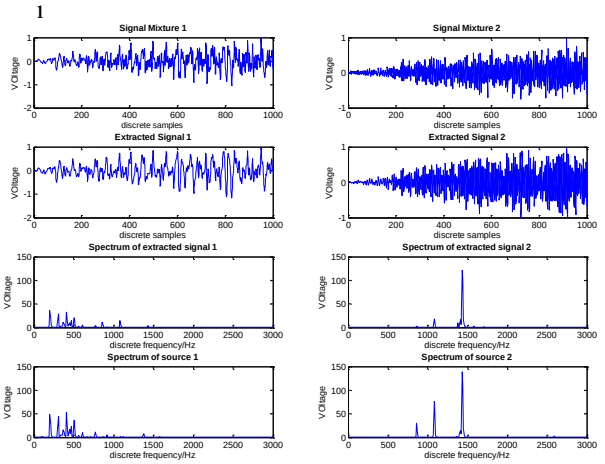

Figure 4. Results of blind separation

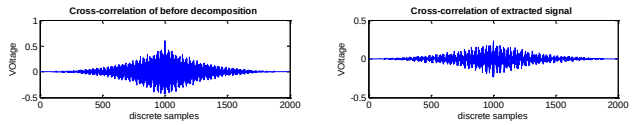

Figure 5. Cross-correlation between signals before and after blind separation

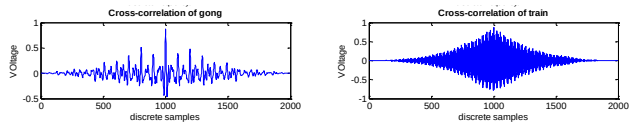

Figure 6. Cross-correlation between the separated signals and original signals before and after blind separation

Given the effective band of original signal gong and train are respectively 200 ~ 1400Hz and 850 ~ 1500Hz, the aliasing bandwidth is approximate 550Hz. In order to study the relationship between aliasing bandwidth and separation performance, before the equable mixture of the two signals, we can take a low-pass filter (or high-pass filter) to change the aliasing bandwidth between the original signals. Here we put gong across a low-pass filter, adjusting the filter parameter and make sure the aliasing bandwidth between outputs and signal train gradually decreased until two signals completely separated in frequency domain. Fig .7 shows the results of separation of different aliasing bandwidth. Three curves from top to bottom in order are respective cross-correlation between separated gong, train and the original signal, and that of after separation.

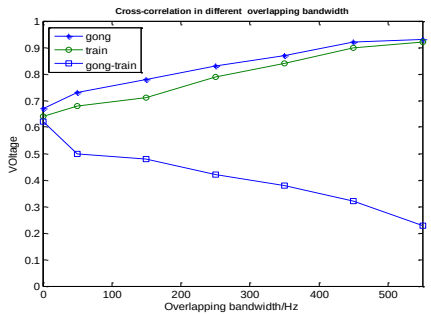

Figure 7. Performance of blind source separation in different aliasing bandwidth

As shown in Fig .7, overlapped in band of the original signal is useful to separation, although the mixed signal detached in frequency domain can be directly separated by high-pass (low-pass) filter in theory, in the condition of "blind" toward the source, it is unable to determine the cut-off frequency of filter apparently. Use blind separation algorithm proposed in this paper, the single channel of dual-source mixed signal can be pre-separated by dichotomous SVD, in that case, the mixed signal of single channel is translate into a double channel. For the mixed signal aliased in frequency domain, it is actually a recombination of the original, while it can be directly separated by dichotomous SVD for signals detached in frequency domain. Obviously, there is no absolutely linear function between the aliasing bandwidth of original signal and the separation performance, if the aliasing band is too wide and the differences of the frequency characteristics between the two signals isn't obvious, which will result in the signal processed by SVD no longer display the characteristic of pre-separation, and therefore decrease the performance of blind separation later. However, when the aliasing band is no more than half of the original signal, it can still achieve a good separation performance, the best aliasing width is about half of the band of the original signal, the blind separation of mixed signals gong and train in this paper is exactly the case.

VI. CONCLUSION

This paper presents a new method of blind separation for single channel of mixed signal, which can be applied to the dual-source mixed signal with apparent difference in frequency distribution. It combines the method of dichotomous SVD and informax and achieves an effective separation for the partial aliasing band of mixed voice. The results of MATLAB simulation shows that the method needn't know too much prior information of source signal, the mixed signal of single channel is recombined as double channel by using the method of dichotomous SVD, thereby a complex underdetermined problem is translated into a simplified even-determined problem. And then the mixed signal is separated by the informax finally. The method is simple and efficient, which provides an important reference value for the blind separation of single channel mixed signals.

Considering of the complexity of mixed mode of voice, the study of this method is aimed at linear instantaneous mixed signal, though it isn't detailed discuss the convoluted mixed signals, the next step will conduct in-depth research on the issue.

REFERENCES

[1] CHEIIRY E C, "Some ezperimenta on the recogution of apeecEt, with one and with two ears," Journal of the Acoustical Society of Arnerica, 1953, 25(5):pp. 975-979.

[2] Schmidt M N, Olsson R K, "Linear regression on sparse features for single-channel speech separation," IEEE Workshop on Application of Signal Processing to Audio Acoustics. NY, USA, 2007 pp. 26-29.

[3] Pearlmutter B, Olsson R, "Linear program differentation for single-channel speech separation," 16th IEEE Signal Processing Society workshop on Machine Learning for Signal Processing. Maynooth, Ireland, 2006 pp. 421-426.

[4] Nakashizuka N, Okumura H, Iiguni Y, "Single-channel speech separation by using a sparse decomposition with periodic structure," 2008 International Symposium on Intelligent Signal Processing and Communications. Bangkok, Thailang, 2008 pp.1-4.

[5] GUO Hai-yan, YANG Zhen, ZHU Wei-ping, "A New Single-Channel Speech Separation Method Based on Sparse Decomposition." ACTA ELECTRONICA SINICA, China, 2012, 40(4) pp. 762-768.

[6] ZHAO Xue-zhi, YE Bang-yan, "Multi-resolution SVD Packet Theory and Its Application to Signal Processing," ACTA ELECTRONICA SINICA, China, 2012, 40(10) pp. 2039-2046.

[7] SUN Shou-zheng, "Blind Signal Processing Foundation and It's Application," National Defence Industry Press, China, 2010 pp. 50-52.

[8] Bell A J, Sejnowski T J, "An information-maxiization approach to blind separation and blind deconvolution," Neural Computation, 1995, 7(6) pp. 1004-1034.

[9] Amari S, Cichocki, "A Adaptive Blind Signal Processing-neural Network Appoaches," Proceedings of IEEE,1998, 86(10) pp. 2026-2046.

[10] Cardoso J F, Laheld B H, "Equivariant Adaptive Source Separation," IEEE Transactions on Signal Processing, 1996, 44 (12) pp. 3017-3030.

[11] Lee T, Girolami M, Sejnowski T, "Independent Component Analysis Using an Extended Information Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources," Neural Computation, 1999, 9(7): pp. 1483-1492.

[12] ZHOU Zhi-yu, FU Hao, "Research and Survey on Algorithms of Blind Signal Separation Technology," Computer Science, China, 2009, 36(10) pp. 16-18.

[13] ZHANG An-qing, "Study on Blind Separation Technology and Its Application of Underwater Acoustic Signals," Dalian University of Technology, China, 2006 pp. 21-24.