Speech Separation based on the GMM pdf Estimation

Yu Xiao *Hu Guangrui*

Department of Electronic and Engineering, Shanghai Jiao Tong Univ.
Shanghai 200030, P.R.China
email: yuxiaoc@online.sh.cn

ABSTRACT

In this paper, the speech separation task will be regarded as a convolutive mixture Blind Source Separation (BSS) problem. The Maximum Entropy (ME) algorithm, the Minimum Mutual Information (MMI) algorithm and the Maximum Likelihood (ML) algorithm are main approaches of the algorithms solving the BSS problem. The relationship of these three algorithms has been analyzed in this paper. Based on the feedback network architecture, a new speech separation algorithm is proposed by using the Gaussian Mixture Model (GMM) pdf estimation in this paper. From the computer simulation results, it can be concluded that the proposed algorithm can get faster convergence rate and lower output Mean Square Error than the conventional ME algorithm.

1. INTRODUCTION

With the development of speech recognition technology, robust speech recognition is becoming the research stress in the speech signal processing. The "cocktail party effect" is the ability to focus one's listening attention on a single talker among a cacophony of conversations and background noise. On the other hand, it also can be regarded as the speech separation ability of human. In this paper, the speech separation task will be regarded as a convolutive mixture Blind Source Separation (BSS) problem. In BSS, the problem is how to recover the original sources without knowing anything except for the independence among the original sources. Utilizing the independence among the speech sources and appropriate BSS algorithm, the speech separation task can be accomplished by separating the clean speech signals from the observed mixture inputs.

Blind Source Separation has been used in many areas such as speech recognition, biomedical signal processing, image processing and data communication etc, which has been received a lot of attention since 90's. Several different BSS algorithms have been proposed in recent years: Widrow's well-known LMS algorithm was proved that the signal-to-noise ratio (SNR) at system outputs is at best the noise-to-signal ratio at the reference input if there are cross-talk inputs in the reference channel [1]. A.J.Bell proposed the Maximum Entropy (ME) algorithm based on the information maximization theory by using nonlinear sigmoid function on the system outputs [2]. P.Comon established the Independent Component Analysis (ICA) theory [3] to solve the BSS problem. S.amari used the Gram-Charlier expansion to estimate the marginal probability density function (pdf) of outputs in the Minimizing Mutual Information (MMI) algorithm [4] in stead of P.Comon's Edgeworth expansion in ICA. A.Belouchrani introduced the Maximum Likelihood (ML) algorithm [5] to solve the BSS problem.

In this paper, a new speech separation algorithm is proposed based on the feedback architecture and the Gaussian Mixture Model (GMM) pdf estimation. From the simulation results, the proposed algorithm demonstrates its better performance in the speech separation task than the conventional ME algorithm.

2. SPEECH SEPARATION BASED ON THE FEEDBACK ARCHITECHTURE

Without loss of generality, a two-channel-input Speech separation system can be illustrated as Figure 1. The observed mixture speech signals can be presented as formula (1) in Z-transform domain, where the cross-talk filters $A_{12}(z)$, $A_{21}(z)$ and the weight filters $W_1(z)$, $W_2(z)$ are assumed to be FIR filters.

$$\begin{cases} X_1(Z) = S_1(Z) + S_2(Z)A_{21}(Z) \\ X_2(Z) = S_2(Z) + S_1(Z)A_{12}(Z) \end{cases} \quad (1)$$

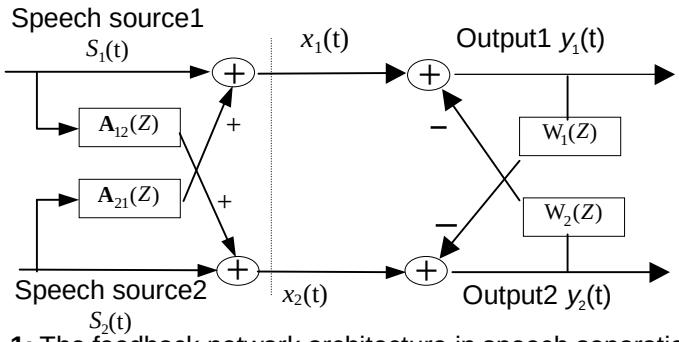

Figure 1: The feedback network architecture in speech separation task.

Most of the works on BSS addressed the case of instantaneous and linear mixture [2-5], where the mixture model is assumed as $\mathbf{X}(t) = \mathbf{A}\mathbf{S}(t)$. So the forward network is an optional architecture in such BSS problem. But in real world applications of speech separation task, the observed inputs shall be modeled as a convolutive mixture model because of the propagation delays in medium and the filter responses of the observed sensors. If the forward network is used in such speech separation task, the best system outputs will not be the needed clean speech signal. And it will be distorted by a filter whose Z-transformation is $1 - A_{12}(z)A_{21}(z)$ [6]. A postprocessing filter is needed to cancel such distortion, whose Z-transformation is $1/[1 - A_{12}(z)A_{21}(z)]$. But there is not any prior knowledge about these filter coefficients. On the other hand, the feedback architecture is an effective and concise network structure to solve the convolutive mixture BSS problem with the assumption that the cross-talk filters $A_{12}(z)$ $A_{21}(z)$ shall be strictly causal to ensure the stability of the corresponding BSS algorithm [6]. In this paper, only the feedback architecture is considered like in Figure 1. So the system outputs can be presented as formula (2) in Z-transformation.

$$\begin{cases} Y_1(z) = X_1(z) - W_2(z)Y_2(z) \\ Y_2(z) = X_2(z) - W_1(z)Y_1(z) \end{cases} \tag{2}$$

The best solutions of the speech separation system in Figure 1 can be presented as formula (3). If it is assumed that the channel detectors are close to the corresponding sources and the cross-talk filters $A_{12}(z)$, $A_{21}(z)$ is strictly causal, the first solution in the formula (3) will be easier to be achieved.

$$\begin{cases} y_1(z) = S_1(z) \\ y_2(z) = S_2(z) \end{cases} \quad with \begin{cases} W_1(z) = A_{12}(z) \\ W_2(z) = A_{21}(z) \end{cases} \quad \text{Or} \quad \begin{cases} y_1(z) = A_{21}(z)S_2(z) \\ y_2(z) = A_{12}(z)S_1(z) \end{cases} \quad with \begin{cases} W_1(z) = 1/A_{21}(z) \\ W_2(z) = 1/A_{12}(z) \end{cases} \tag{3}$$

3. THE ME, MMI and ML BLIND SOURCE SEPARARTION ALGORITHM

The ME algorithm and the MMI algorithm are two main approaches of the BSS algorithms. The main idea of the Maximum Entropy algorithm is: transforming the system outputs **Y** to $\mathbf{Z} = \mathbf{G}(\mathbf{Y})$, which can be regarded as the outputs of some nonlinear neurons whose nonlinear function is $G(y)$. It is expected that the mutual entropy of the output vector **Z** will be maximized when the components of **Z** are mutually independent. Then the ME algorithm is deduced by maximizing the joint entropy $\mathrm{H}(\mathbf{Z};\mathbf{W})$ with respect to the weight vector **W** by using the traditional stochastic gradient descent method. The joint entropy of **Z** can be presented as formula (4).

$$\mathrm{H}(\mathbf{Z};\mathbf{W}) = -\int \mathrm{p}(\mathbf{z};\mathbf{w}) \log \mathrm{p}(\mathbf{z};\mathbf{w}) d\mathbf{z} = \mathrm{E}[\ln|\mathbf{J}|] - \mathrm{E}[\ln \mathrm{p}(\mathbf{x})] \tag{4}$$

Where $p(\mathbf{z};\mathbf{w})$ is the joint probability density function (pdf) of the output vector **Z**. And $\mathrm{p}(\mathbf{x})$ is the joint

probability density function (pdf) of the observed input vector **X**. $|\mathbf{J}| = \left| \frac{\partial z_1}{\partial x_1}\frac{\partial z_2}{\partial x_2} - \frac{\partial z_1}{\partial x_2}\frac{\partial z_2}{\partial x_1} \right|$ is the absolute value of the Jacobean polynomial. If the feedback architecture is used and the weight filters $W_1(z)$, $W_2(z)$ are strictly causal or no more than one of the weight filters has the zero delay items, then $|\mathbf{J}| = \left| \frac{\partial z_1}{\partial y_1}\frac{\partial z_2}{\partial y_2} \right| = |g(y_1)g(y_2)|$, where $g(y) = \frac{\partial G(y)}{\partial y}$ is the differential coefficient of the nonlinear function $G(y)$ of the neurons.

Unlike the ME algorithm, in the MMI algorithm, the weight iterative formulas are deduced by directly minimizing the mutual information (MI) between the system outputs. The MI is measured as formula (5) by the Kullback-Leibler divergence $\mathbf{I(W)}$ between the joint probability density function $\mathrm{P}(\mathbf{Y},\mathbf{W})$ of the output vector **Y** and its factorized version $\tilde{\mathrm{P}}(\mathbf{Y},\mathbf{W}) = \mathrm{p}(y_1,\mathbf{W})\mathrm{p}(y_2,\mathbf{W})$ which is the product of the marginal probability density functions of the vector **Y**:

$$\mathrm{I}(\mathbf{W}) = D[p(\mathbf{Y},\mathbf{W}) \| \tilde{\mathrm{p}}(\mathbf{Y},\mathbf{w})] = \int p(Y,\mathbf{w}) \log \frac{\mathrm{p}(\mathbf{Y},\mathbf{w})}{\tilde{\mathrm{p}}(\mathbf{Y},\mathbf{w})} d\mathbf{y} \tag{5}$$

It is easy to prove $\mathbf{I(W)} \geq 0$ and the mutual information $\mathbf{I(W)}$ will be equal to zero if and only if the output signals y_1, y_2 are mutual independent. In order to deduce the iterative formulas, A.amari [5] used the Gram-Charlier expansion instead of P.Comon's Edgeworth expansion to estimate the marginal pdf of the system output.

Based on the forward architecture, A.Belouchrani introduced the Maximum Likelihood (ML) Algorithm by directly maximizing the Likelihood with respect to the weight filter vectors by using the stochastic density descend method like formula (6), where $f_j(y_j,\mathbf{W}_j)$ is the pdf of the output.

$$\frac{d\hat{G}}{d\mathbf{W}_j} = -\frac{df_j(y_j;\mathbf{W}_j)/d\mathbf{W}_j}{f_j(y_j;\mathbf{W}_j)} \tag{6}$$

4. The PROPOSED SPEECH SEPARATION ALGORITHM BASED ON THE GMM pdf ESTIMATION

The ME algorithm has not yet been rigorously justified except for the case when the sigmoid nonlinear function in ME algorithm happens to be the cumulative density function of the unknown sources. Based on the Independent Component Analysis (ICA) theory, the Mutual Information is one of the best contrast functions because it is invariant under the transforms such as scaling, permutation, and componentwise non-linear transforms. In the ML algorithm, it is needed to know the probability density function form of the needed system output in order to deduce the corresponding iterative formulas. In some way, the ME algorithm can be regarded as a special case of the ML algorithm, where one of special probability density functions is considered in the ME algorithm: the logistic pdf $1/(1+\exp(-x))$.

The ME algorithm and the ML algorithm directly use the instantaneous value in stead of the expectation value of the gradient of the mutual entropy or the Likelihood to deduce the corresponding iterative formulas. On the other hand, in the MMI algorithm, the truncated pdf polynomial expression is used to estimate the formula expression of the expectation value of the mutual information, which is used to deduce the iterative formulas by using statistical gradient descend method. In real application, finite-sample estimation of higher-order cumulants (such as the 3rd cumulant: skewness and the 4th cumulant: kurtosis) are sensitive to outliers, and there may be approximation errors when it is used in some unstable signal such as the speech signal. So in some experiments using speech signals [2], it was concluded that the ME algorithm has better performance than the MMI algorithm. The joint Entropy in the ME algorithm has the relationship with the mutual information in the MMI algorithm as formula (7).

$$\mathrm{H}(\mathbf{Z};\mathbf{W}) = \mathrm{H}(\mathbf{Y};\mathbf{W}) + \sum_{i=1}^{2} \int p(y_i,w) \log g_i(y_i) dy_i = -\mathrm{I}(\mathbf{Y};\mathbf{W}) - \sum_{i=1}^{2} \mathrm{D}\left[\mathrm{p}(y_i(w)) \| g_i(y_i(w))\right] \tag{7}$$

From the formula (7), it can be concluded that the ME algorithm will be same as the MMI algorithm if the Kullback-Leibler divergence $\sum_{i=1}^{2} \mathrm{D}\left[\mathrm{p}(y_i(w)) \| g_i(y_i(w))\right]$ is equal to zero, where $\mathrm{p}(y_i, w)$ is the marginal pdf of the needed speech output, and $g_i(y_i, w)$ can be seen as the estimation of $\mathrm{p}(y_i, w)$ in the ME algorithm. So the mutual entropy in the ME algorithm will also become one of the best contrast functions if $g_i(y_i, w)$ is chosen as the good estimation of the needed speech output's pdf.

Based on upper analysis and the feedback architecture in Figure 1, a new speech separation algorithm is proposed here by using Gaussian Mixture Model (GMM) probability density function (pdf) estimation of system outputs. The main advantage of the new algorithm is that the GMM pdf estimation of the speech signals is closer to the original speech pdf than the logistical pdf estimation in the conventional ME algorithm. So it is expected that the proposed algorithm can get better performance than the conventional ME algorithm. The use of GMM for modeling speech signals in the speech separation task is motivated by such two interpretations. First, GMM can be used to approximate arbitrary probability density function of different sources especial in the non-Gaussian signal [7,8]. Second, the GMM has been successfully used in the speaker verification task [9] by D.A.Reynold et.al.

In the proposed speech separation algorithm, the output y_j is regarded as a random variable at time t. the output's GMM pdf will be presented like formula (8), which is the weighted sum of some component probability density functions (pdfs) $b_i(y_j), i = 1,2,\cdots M$. Each Gaussian component pdf is a Gaussian function like formula (9). p_i are the mixture weight coefficients, and M is the order of GMM. Then every system output's pdf can be depicted by a vector $\lambda_j = \{\mathbf{m}^{\mathrm{M}}, \sigma^{\mathrm{M}}, \mathbf{p}^{\mathrm{M}}\} = \{\{m_1 \cdots m_M\}, \{\sigma_1 \cdots \sigma_M\}, \{p_1 \cdots p_M\}\}$, which is composed of the means, variances and mixture coefficients of the corresponding output.

$$\mathrm{P}(y_j | \lambda) = \sum_{i=1}^{M} p_i b_i(y_j) \tag{8}$$

$$b_i(\mathbf{y}_j) = \frac{1}{(2\pi)^{1/2}\sigma_i} \exp\left(-\frac{(y_j - m_i)^2}{2\sigma_i^2}\right) \tag{9}$$

In the proposed speech separation algorithm, the traditional Expectation Maximization (EM) algorithm [7,8] is used to estimate the GMM coefficient vectors λ_i of N point outputs, like the formula (10).

$$\begin{cases} p_i^{k+1} = \dfrac{\sum_{t=1}^{N} h_i^k(t)}{N} \\ m_i^{k+1} = \dfrac{\sum_{t=1}^{N} h_i^k(t) y_j(t)}{\sum_{t=1}^{N} h_i^k(t)} \\ \sigma_i^{k+1} = \dfrac{\sum_{t=1}^{N} h_i^k(t)(y_j(t) - m_i^k)(y_j(t) - m_i^k)}{\sum_{t=1}^{N} h_i^k(t)} \end{cases} \tag{10}$$

Where the posterior probabilities $h_i^k(t)$ are defined as formula (11):

$$h_i^k(t) = \frac{p_i^k \, \mathrm{P}(y_j(t) | m_i^k, \sigma_i^k)}{\sum_{i=1}^{\mathrm{M}} p_i^k \, \mathrm{P}(y_j(t) | m_i^k, \sigma_i^k)} \tag{11}$$

By using the information maximum theory [2], the proposed algorithm has the iterative formulas like formula (12). In formula (12), η is the stepsize, $\mathbf{W}_i(t)$ is the weight filter vector and $\mathbf{Y}_i(t)$ is the output vector at time t, $\varphi(y_i)$ is the iterative kernel in the proposed speech separation algorithm and it can be measured by formula (13). $\mathrm{var}(\varphi(y_i))$ is a stepsize adaptive coefficient in order to control the fluctuation of the weight filter vector $\mathbf{W}_i$ in the proposed algorithm, whose value is the variance of the iterative kernel $\varphi(y_i)$ in last pass iteration. From the simulation results, it can be concluded that $\mathrm{var}(\varphi(y_i))$ is

very important in improving the performance of the proposed algorithm.

$$\begin{cases}\mathbf{W}_1(t+1)=\mathbf{W}_1(t)-\eta\varphi(y_2)\mathbf{Y}_1(t)/\operatorname{var}(\varphi(y_2))\\ \mathbf{W}_2(t+1)=\mathbf{W}_2(t)-\eta\varphi(y_1)\mathbf{Y}_2(t)/\operatorname{var}(\varphi(y_1))\end{cases} \tag{12}$$

$$\varphi(y)=\frac{\sum_{i=1}^{\mathrm{M}}-\left[\frac{(y-m_i)}{\sigma_i^2}\frac{p_i}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(y-m_i)^2}{2\sigma_i^2})\right]}{\sum_{i=1}^{\mathrm{M}}\frac{p_i}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(y-m_i)^2}{2\sigma_i^2})} \tag{13}$$

The proposed speech separation batch-iterative algorithm of N point mixture inputs is as follow:

1. Initializing the weight vectors $\mathbf{W}_i$, $\operatorname{var}(\varphi_i(y_i))$ and the stepsize η.
2. Initializing the GMM coefficient vectors λ_i by using L iteration of the EM algorithm on the N-point mixture inputs.
3. Compute the N-point outputs by using formula (12) and (13) on one pass N-point mixture inputs and compute the stepsize adaptive coefficients $\operatorname{var}(\varphi_i(y_i))$.
4. Estimate new GMM coefficient vectors λ_i by using L iteration of the EM algorithm on the new N-point outputs.
5. Judge the convergence of weight vectors $\mathbf{W}_i$, if not, come back to procedure 3 to begin a new batch-iteration, else exit the loop and finish the speech separation task.

Because of the low convergent rate of the EM algorithm in getting the coefficients of GMM, the coefficient L is important in the proposed algorithm. The convergent rate of the new algorithm can be faster and the output Mean Square Error (MSE) will be high when L is chosen lower. If L is chosen large, the output GMM pdfs estimation will be closer to the original pdf of the needed output signals. Then the proposed algorithm will have lower output MSE and slower convergent rate. In our simulation, when L=3, the proposed algorithm already get better performance than the conventional ME BSS algorithm.

5. COMPUTER SIMULATION

In order to check the validity of the proposed algorithm, the proposed algorithm is applied in a speech separation task like Figure 1. All of the computer simulation results demonstrate that the new algorithm can accomplish the speech separation task successfully, and it has faster convergence rate and lower Mean Square Error than the conventional ME algorithm.

In this case, two independent sources are all acoustic signals with 8k/s sample-rate: one is a speech signal in English and another is a clap signal. In the simulation, N=12000 which means that the duration of one batch acoustic input is equal to 1.5s. And $\eta=0.0003$, M=8, the order of the weight FIR filters $\mathbf{W}_i$ is chosen as 36. The system mixture model is like formula (14). Because of the low convergent rate of the EM algorithm, L can be chosen as 3~6 to reduce the computation of the proposed algorithm. In this case, L=3.

$$\begin{cases}x_1(t)=S_1(t)-0.6S_2(t-2)+0.4S_2(t-8)\\ x_2(t)=S_2(t)+0.8S_1(t-3)-0.3S_1(t-12)\end{cases} \tag{14}$$

$$WMSE=\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{K}[\mathbf{W}_j^t(i)-\mathbf{W}_{j_{opt}}(i)]^2 \tag{15}$$

The convergent curves of the Weight vector Mean Square Error (WMSE) of weight vector $\mathbf{W}_1$ are showed in Figure 2, where the WMSE is computed by formula (15) in which $\mathbf{W}_{jopt}$ is the jth optimal weight filter vector in the simulation. In Figure 3, the convergent curves of the weight coefficient $W_2(2)$

are also displayed, where the expectation value of $W_2(2)$ is -0.6. It is clear that the proposed algorithm has faster convergent rate and lower WMSE than the conventional ME algorithm. By listening to the output speech signals and the original clean sources, the speech separation results of the proposed algorithm are almost perfect.

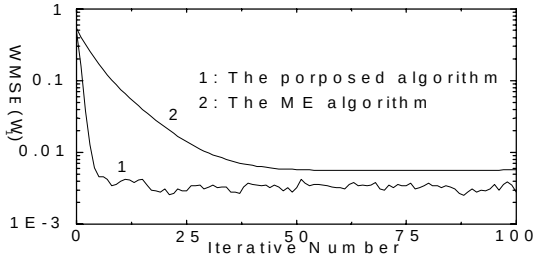

Figure 2: The WMSE(W_1) curves of two algorithms.

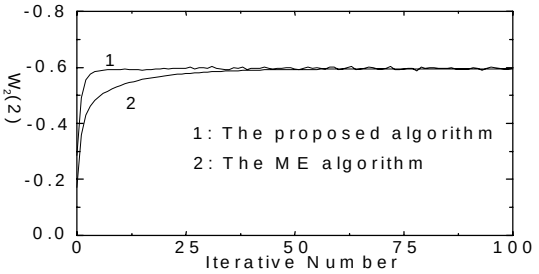

Figure 3: The convergent curves of weight $W_2(2)$.

6. CONCLUSION

In this paper, a new speech separation algorithm by using GMM pdf estimation is proposed based on the feedback network architecture. According to the simulation results, it can be concluded that the proposed algorithm can get faster convergent rate and lower WMSE than conventional ME algorithm. Future work will be mainly concerned on the use of nonlinear filters to improve the performance of speech separation in real world.

REFERENCES

1. B.Widrow and S.Stearns, 'Adaptive Signal Processing', Prentice-Hall, New York 1985.
2. Bell,A.J., and Sejnowski, T.J. 'An information-maximization approach to blind separation and blind deconvolution.' Neural Computation, vol. 7, 1995, pp1129-1159
3. Pierre Comon, 'Independent Component Analysis, A new Concept?' Signal Processing, 1994, vol. 36 pp287-314
4. S.Amari, A.Cichocki, and Yang, H.H. 'A new learning algorithm for blind signal separation', In 'Advances in Neural Information Processing Systems', Editors D.Touretzky, M.Mozer, and M.Hasselmo, MIT Press: Cambridge, MA. 1996,8 pp757-763.
5. A.Belouchrani and J.F.Cardoso, 'Maximum likelihood source separation for discrete sources.' In Proceeding of EUSIPCO, pp768-771, 1994.
6. K.Torkkola. Blind separation of delayed sources based on information maximization. Proceeding of ICASSP, Atlanta, GA, USA, May 7-10,1996
7. Lei Xu, and Michael I,Jordan, 'On Convergence Properties of the EM Algorithm for Gaussian Mixtures', Neural Computation, vol. 8, pp129-151, 1996
8. Yunxin Zhao, Xinhua Zhuang, and Sheu-Jen Ting, 'Gaussian Mixture Density Modeling of Non-Gaussian Source for Autoregressive Process', IEEE Trans on Signal Processing, Vol. 43(4), 1995.4 pp894-903
9. Douglas A.Reynold, Ridchard C.Rose, 'Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models' IEEE Trans on Speech and Audio Processing, Vol. 3(1) 1995 pp72-82