Dual-channel speech separation using interaural time difference with Generalized Gaussian Mixture Model

Zhaogui Ding[1, a] , Liming Zhang[2,b], Longbiao Wang[3,c] and Weifeng Li[4,d*]

[1] Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China

[2]Faculty of Science and Technology University of Macau, Avenida da Universidade, Macau, China.

[3]Top Runner Incubation Cente, Nagaoka University of Technology, Japan.

[4] Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China.

[a]dingzhaogui@126.com, [b]lmzhang@umac.mo, [c]wang@vos.nagaokaut.ac.jp, [d]li.weifeng@sz.tsinghua.edu.cn

Keywords: interaural time difference (ITD) statistics, Generalized Gaussian Mixture Model, correlation coefficient, time-frequency mask

Abstract. In this letter we present a novel speech separation scheme using two microphones. The proposed method utilizes the estimation of interaural time difference (ITD) statistics for the separation of mixed speech sources. The novelties of this paper consist in the use of Generalized Gaussian Mixture Model (GGMM) for speech separation frame by frame and cross-correlation coefficient for distributed parameter selection. The proposed model can be extended to audio enhancement. Our objective quality evaluation experiments demonstrate the effectiveness of the proposed methods and show significant quality improvements over the conventional dual ITD based methods.

Introduction

In order to mimic the sound separation abilities of human listeners, computational auditory scene analysis (CASA) has been developed [1]. Binaural CASA systems localize sound sources by measuring the interaural time differences (ITD) and the interaural intensity differences (IID). The computational goal of the binaural CASA systems is to obtain an ideal binary mask. Interaural phase differences (IPD) have been used in [2]. In [2] the author proposed a speech enhancement algorithm which utilizes phase-error based filters that depend only on the phase of the signals. Instead of a fixed threshold, [3] employed a statistical modeling of angle distributions together with a channel weight to determine which signal component belongs to the target signal and which is part of the background noise. While the common Gaussian Mixture Model (GMM) could be used in ITD, [4] employed the Laplace Mixture Model (LMM) to fit the ITD statistics and theoretically proved the rationality of his model. However, [4] utilized some approximation to illustrate the rationality of LMM, which means that LMM was not the best model for ITD statistics.

In this paper, we present a new ITD statistics based technique capable of separating speech signals through two microphones. Considering that both GMM and LMM are special cases of Generalized Gaussian Mixture, we novelly employ Generalized Gaussian Mixture to estimate the ITD statistics in each frame. Moreover, we utilize correlation coefficient to select the best distribution for every sentence instead of each frame. The framework of our approach is illustrated in Fig.1.

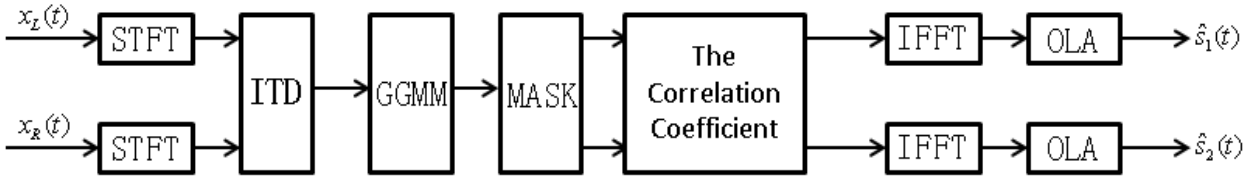

Fig. 1. Block diagram of the proposed approach. STFT: Short Time Fourier Transform, ITD: Interaural Time Difference, GGMM: Generalized Gaussian Mixture Model, IFFT: Inverse Fast Fourier Transform, OLA: Over-Lapping and Adding.

Time Difference Model

Suppose that there are I ($I = 2$) sources (we use s_1 to represent the target source and s_2 to represent the interfering source) in a sonic environment. The signals from two different microphones are defined respectively as:

$$\begin{aligned} x_L(t) &= \sum_{i=1}^{I} a_i^L s_i(t) \\ x_R(t) &= \sum_{i=1}^{I} a_i^R s_i(t - t_i) \end{aligned} \tag{1}$$

where a_i^L and a_i^R denote the weighted coefficients of the recordings of the left and right microphone from i-th source separately. t_i is the time delay of arrival (TDOA) of i-th source between two microphones. With the short-time Fourier transform (STFT), the signals can be expressed as:

$$\begin{aligned} X_L[m,k] &= \sum_{i=1}^{I} a_i^L S_i[m,k] \\ X_R[m,k] &= \sum_{i=1}^{I} a_i^R S_i[m,k] \times e^{-jw_k t_i[m,k]} \end{aligned} \tag{2}$$

where m is the frame index and $w_k = 2p\,k/K$. Here k and K are the frequency index and total frequency bins respectively. Time delay measured in time-frequency [m, k] can be expressed as:

$$t[m,k] = \frac{\angle X_L[m,k] - \angle X_R[m,k] + 2p\,r}{w_k} \tag{3}$$

where $\angle\cdot$ indicates the phase of signal. r is an integer which makes the value of $\angle X_L[m,k] - \angle X_R[m,k]$ limited between [$-\pi$, π].

Proposed Approach

We use a probabilistic approach to the problem of ITD estimation. Several mixture models are used to obtain the distribution characteristics, e.g., Gaussian Mixture Model (GMM), Laplace Mixture Model (LMM) [4]. As both the Gaussian distribution and Laplace distribution are special cases of Generalized Gaussian distribution, we can utilize the Generalized Gaussian Mixture Model (GGMM) to obtain a precise fitting.

In the GGMM, $\tau[m, k]$ is modeled as:

$$p_{mix}(t \mid \Theta^G[m]) = \sum_{i=1}^{I} w_i\, p_i(t \mid J_i^G[m]) \tag{4}$$

where $\Theta^G[m]$ is the set of Generalized Gaussian Mixture Model parameters and $\sum_{i=1}^{I} w_i = 1$. More specially,

$$\begin{aligned} J_i^G[m] &= \{m_i[m], s_i[m], l\} \\ \Theta^G[m] &= \{w_1, w_2, J_1^G[m], J_2^G[m]\} \end{aligned} \tag{5}$$

and

$$p_i(t_i \mid J_i^G[m]) = A(l)\, e^{-B(l)\left|\frac{t - m_i[m]}{s_i[m]}\right|^l} \tag{6}$$

where $A(l)=\frac{l\sqrt{\Gamma(3/l)}}{2s_i[m]\Gamma(1/l)}$, $B(l)=\left[\frac{\Gamma(3/l)}{\Gamma(1/l)}\right]^{l/2}$. $\Gamma(\cdot)$ denotes the gamma function. μ_i and σ_i are the pdf mean and standard deviation respectively. The parameter λ controls the details of the pdf. Utilizing the EM algorithm [6], we have

$$w_i[m] \leftarrow \frac{\sum_{k=1}^{K} b_i[m,k]}{K} \tag{7}$$

$$m_i[m] \leftarrow \frac{\sum_{k=1}^{K} p_i(t \mid J_i^G[m]) \mid t[m,k]-m_i[m]\mid^{l-1} t[m,k]}{\sum_{k=1}^{K} p_i(t \mid J_i^G[m]) \mid t[m,k]-m_i[m]\mid^{l-2}} \tag{8}$$

$$s_i[m] \leftarrow \left[\frac{\sum_{k=1}^{K} l\,B(l)\,p_i(t \mid J_i^G[m]) \mid t[m,k]-m_i[m]\mid^{l}}{\sum_{k=1}^{K} p_i(t \mid J_i^G[m])}\right]^{1/l} \tag{9}$$

where $\beta_i[m, k]$ can be obtained by Bayesian principle:

$$b_i[m,k]=\frac{w_i p_i(t \mid J_i^G[m])}{\sum_{l=1}^{I} w_l p_l(t \mid J_l^G[m])} \tag{10}$$

Normally, λ can be updated refer to [7] for each frame. Differing from the traditional method, we assume that every mixed sentence exists a special λ which is optimal to fit the ITD statistics. While λ = 1 (LMM) and $\lambda = 2$ (GMM) are two special cases of our hypothesis. The detailed information about the selection of λ will be described in the following section. The results fitted by both GMM, LMM and GGMM are shown in Fig. 2.

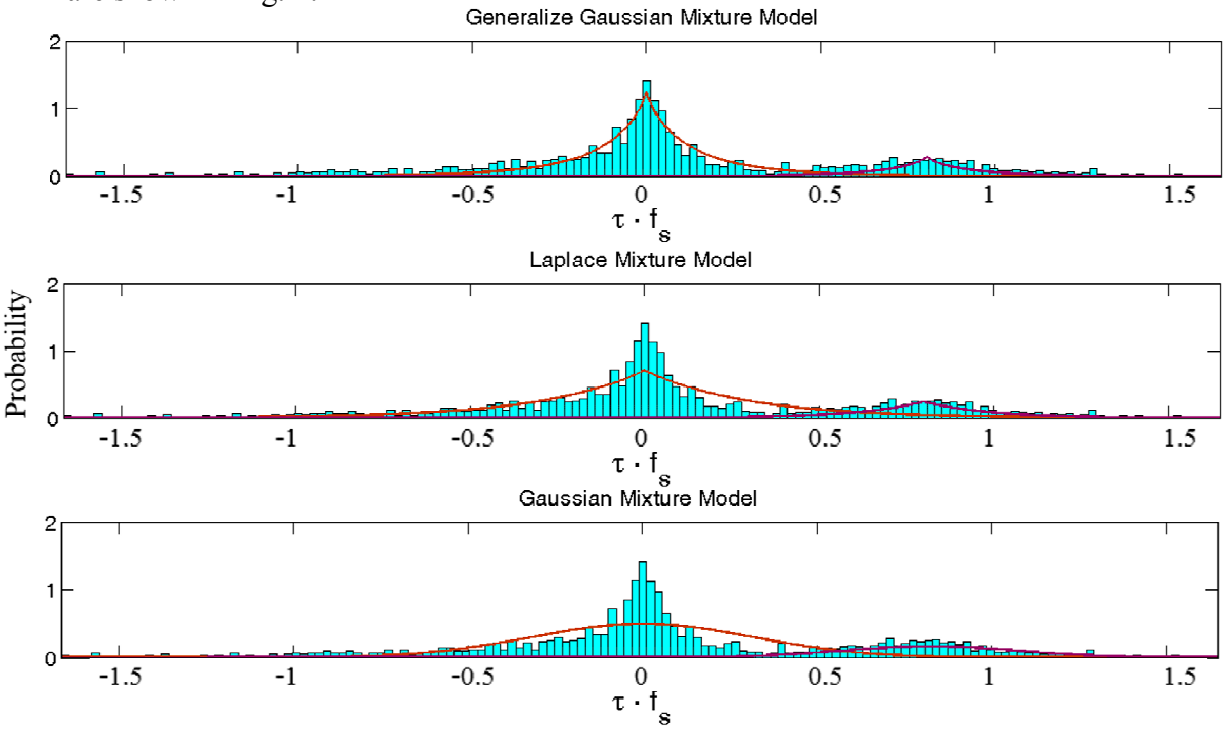

Fig. 2. The results fitted by GGMM, LMM and GMM separately. The horizontal axis is $\tau \cdot f_s$, where τ is the time delay of arrival and f_s is the sampling rate.

After obtaining the rough probabilistic fittings of the ITD, we adopt the masking method to separate the target and interfering sources. In our studies, we employed the Likelihood Ratio Criterion (LRC),

which provides a binary masking. Two hypothesis H_0 and H_1 which respectively indicate the target source plays a dominant role in the mixtures or not can be described as:

H_0: target is dominant

H_1: interference is dominant

The LRC criterion suggests the following decision rule in GGMM:

$$\frac{P^G(H_0 \mid t)}{P^G(H_1 \mid t)} = \frac{w_1[m]p_1(t \mid J_1^G[m])}{w_2[m]p_2(t \mid J_2^G[m])} \underset{H_1}{\overset{H_0}{\gtrless}} 1 \tag{11}$$

where superscript G indicates the likelihood term associated with GGMM. Let

$$M_1[m,k] = \begin{cases} 1 & if\ P^G(H_0 \mid t) \geq P^G(H_1 \mid t) \\ 0 & otherwise \end{cases} \tag{12}$$

be the mask indicator function of target source for time-frequency point $[m, k]$. Then both the target and interfering speeches can be separated as:

$$\begin{aligned} \hat{S}_1[m,k] &= M_1[m,k]X[m,k] \\ \hat{S}_2[m,k] &= (1\text{-}M_1[m,k])X[m,k] \end{aligned} \tag{13}$$

where $X[m, k]$ is defined as:

$$X[m,k] = \frac{1}{2}(X_L[m,k] + X_R[m,k]) \tag{14}$$

Then we introduce the method which utilizes the cross-correlation coefficient to select the optimal λ.

Motivated by [5], here we perform an exhaustive search to find the optimal λ using the cross-correlation coefficient (we restrict the range of λ between 0.5 and 2.5),

$$r(m \mid l) = \frac{\frac{1}{M}\sum_{m=1}^{M} P_1(m \mid l)P_2(m \mid l) - m_{P_1}m_{P_2}}{s_{P_1}s_{P2}} \tag{15}$$

where $P_1(m|\lambda)$ are $P_2(m|\lambda)$ defined as:

$$\begin{aligned} P_1(m \mid l) &= \left(\sum_{k=1}^{K} |\hat{S}_1[m,k]|^2\right)^{a_0} \\ P_2(m \mid l) &= \left(\sum_{k=1}^{K} |\hat{S}_2[m,k]|^2\right)^{a_0} \end{aligned} \tag{16}$$

where $a_0 = 0.2$ as in [5]. The optimal $\hat{l}$ is then obtained by minimizing the $|\rho(m|\lambda)|$,

$$\hat{l} = \arg\min_{l} |r(m \mid l)| \tag{17}$$

Once we obtain the optimal estimation of $\hat{l}$, the mixed signals can be separated. Finally, we can obtain the separated speech waveforms using the Inverse Fast Fourier Transform (IFFT) and Over-Lapping and Adding (OLA).

Experimental Results

Dual-channel distorted speech signals were used to evaluate our proposed algorithm. The source signals (100 sentences) were recordings of 2s length obtained from concatenating sentences randomly drawn from the TIMIT database at 16KHz sampling rate. The set of experiments was conducted using simulated reverberant environments in which the target speaker is masked by an interfering speaker. The distance between two microphones is 2 cm. Reverberation simulations were accomplished using the Room Impulse Response (RIR) [8] open source software package based on the image method. In the experiments in this section, we assumed room dimensions of 6 × 4 × 2.5, with microphones that are

located at the center of the room. The reverberation time is about 0.1s. For all speakers, the distances between the speaker locations and the center of the microphones are 1.5m. We generate 100 mixtures respectively for two environments (S_1 is the target source. $S_2^{(1)}$ and $S_2^{(2)}$ are the interferer source respectively). The illustration of microphone and source placement is shown in Fig. 3.

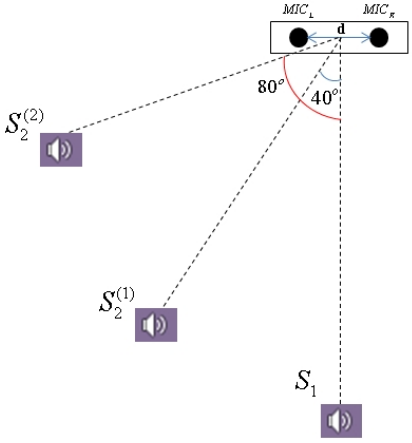

Fig. 3. Placement of the microphones and sound sources. S_1 is the target source. For one environment, $S_2^{(1)}$ is the interfering source, while $S_2^{(2)}$ is the interfering source for another environment.

We evaluate the quality of the recovered speech using SIR [9], PESQ [10], C_{sig}, C_{bak} and C_{ovl} [11]. The ratings are based on the $1-5$ MOS scale, range from 1 (bad) to 5 (excellent). With an overlap of 75%, we assign the window length as 1024 samples. In order to obtain enough information of τ, we add up every $2N+1$ frames. Namely, when we obtain the information of $\tau[m]$ for the frame m, we will add up $\sum_{q=m-N}^{m+N} t[q]$. Here we set N as 2. Furthermore, we make the constraint that the angle between target source and interfering source is larger than 10^o. When we initialize the update equations, $K-$means can be used to obtain the rough initial values of μ. In our experiment, the number of the sources is 2.

We evaluate our proposed method on the generated speech signals. Meanwhile, we compare our approach with Gaussian Mixture Model, Laplace Mixture Model based methods and other existing dual-channel speech separation approaches. For convenience, these comparing approaches are referred to as GMM, LMM [4], AUTO [5], SMAD [3] and NMF [12] in our experiments.

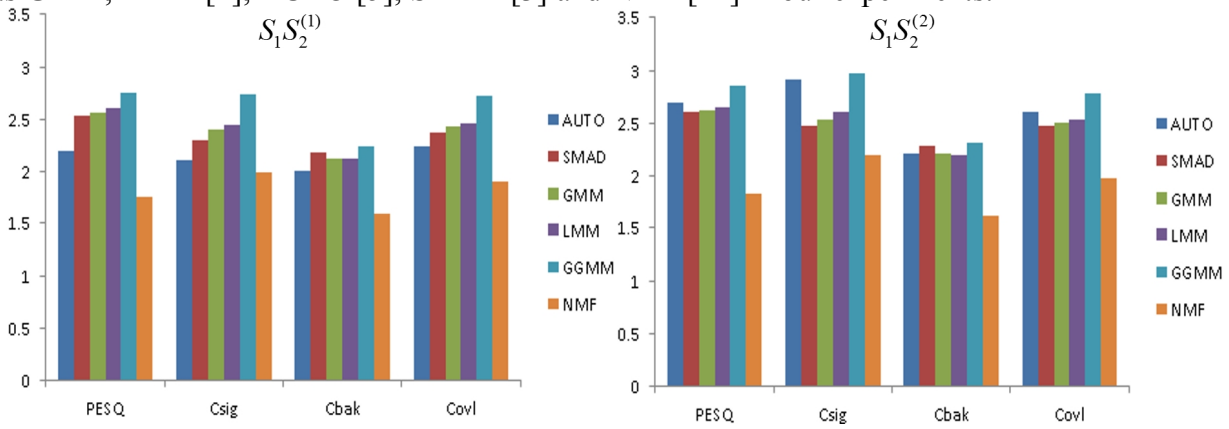

Fig. 4. Separation performance of different methods on $S_1S_2^{(1)}$ and $S_1S_2^{(2)}$ scenarios in terms of Perceptual Evaluation of Speech Quality (PESQ) score, C_{sig}, C_{bak}, and C_{ovl}.

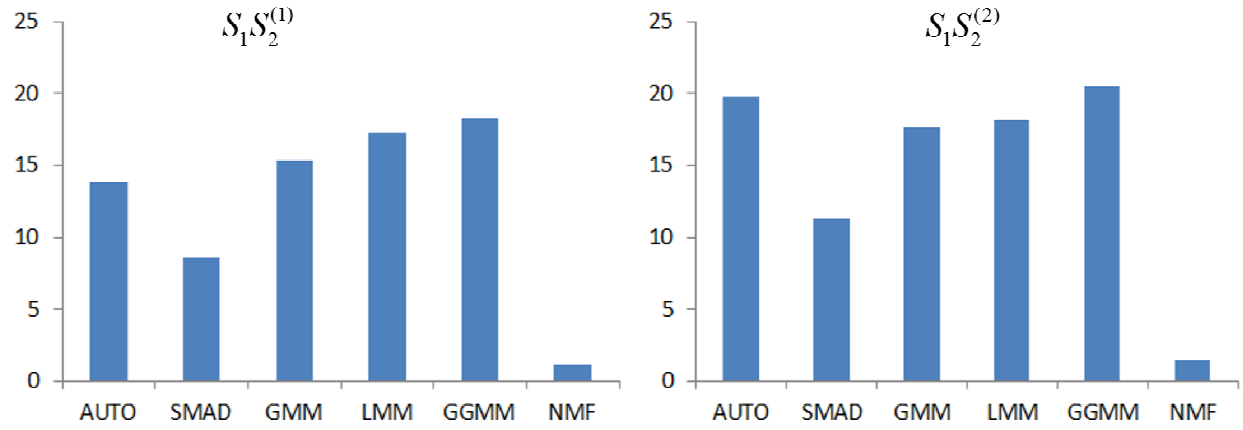

Fig. 5. Separation performance of different methods on $S_1S_2^{(1)}$ and $S_1S_2^{(2)}$ scenarios in terms of Signal-to-Interference Ratio (SIR).

Figs. 4 and 5 show the separation results of S_1. We do not assume that the probabilistic density functions (PDFs) of the spectrum are Gaussian or Non-Gaussian, which is often embedded in non-negative matrix factorization (NMF) based method. The results of $S_1S_2^{(2)}$ by AUTO are better than GMM, LMM, SMAD and NMF. The Auto utilize a fixed threshold by minimize the correlation coefficient, the scope of the threshold is strict because if one separated speech including two speakers and another including almost nothing, the correlation coefficient will also be small. The nearer the distance between two sources, the more possibly this situation happens. As a result, the performances of $S_1S_2^{(2)}$ by AUTO are well, while performances of $S_1S_2^{(1)}$ by AUTO are poor. The methods based on statistics avoid the drawback of AUTO. Results indicate that both the performances of $S_1S_2^{(1)}$ and $S_1S_2^{(2)}$ by SMAD are poor. Unlike traditional model based on ITD, SMAD is based on statistical angles, which requires the situation that distance between two microphones is close while our database does not strictly meet this condition. As illustrated in [4], the performances of LMM are better than GMM in $S_1S_2^{(1)}$ and $S_1S_2^{(2)}$, which indicates that the value of λ affects the separation performance. Our proposed method performs better both in $S_1S_2^{(1)}$ and $S_1S_2^{(2)}$ than other methods.

Conclusions

In this paper we have proposed a novel source separation approach. Our method, for the first time, employs Generalized Gaussian Mixture Model (GGMM) to estimate the statistical information about interaural time difference (ITD) in each frame. Using Generalized Gaussian Mixture Model, a rough expression of the probabilistic density function (PDF) of the ITD can be obtained and a masking filter can be calculated based on the so-obtained probabilistic distributions. Then the accurate expression of the probabilistic density function (PDF) can be obtained using the correlation coefficient and the separated speech can be obtained. Objective evaluations on speech separations demonstrated the effectiveness of our proposed methods in terms of SIR, PESQ, C_{sig}, C_{bak}, and C_{ovl}.

Acknowledgements

The authors would like to thank the anonymous reviews for their constructive comments and suggestions, which are helpful for the improvement of this paper both in technical and literary quality. This work was supported by Shenzhen Basic Research Project (No.JCYJ20140618164241825).

References

[1]D. A and R. Horaud, A computational model of binaural localization and separation, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP 1983 Vol. 8. IEEE, 1983.

[2] A. Parham and G. Shi. IEEE Transactions on 34 (2004) 1763-1773.

[3] K. Chanwoo, C. Khawand, and R. M. Stern, Two-microphone source separation algorithm based on statistical modeling of angle distributions. Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012. 4629-4632.

[4] C. Maximo, J. J. Lopez, and D. Martinez. Two-microphone multi-speaker localization based on a Laplacian mixture model. Digital Signal Processing 21 (2011) 66-76.

[5] K. Chanwoo, C. Khawand, and R. M. Stern. Automatic selection of thresholds for signal separation algorithms based on interaural delay. INTERSPEECH. 2010. 729-732.

[6] M. Geoffrey and T. Krishnan. The EM algorithm and extensions. John Wiley & Sons, Vol. 382. 2007.

[7] A. M. Saïd. Image Processing, IEEE Transactions on 21 (2012) 1452-1464.

[8] Al. J. B, and D. A. Berkley. Image method for efficiently simulating small- room acoustics. The Journal of the Acoustical Society of America 65 (1979) 943-950.

[9] V. Emmanuel, S. Hiroshi. First stereo audio source separation evaluation campaign: data, algorithms and results. Independent Component Analysis and Signal Separation. Springer Berlin Heidelberg, (2007) 552-559.

[10] R. Antony, and J. Beerends. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Acoustics, Speech, and Signal Processing, 2001. 2001 IEEE International Conference on. Vol. 2. IEEE, 2001.

[11] H. Yi, and P. C. Loizou. Audio, Speech, and Language Processing, IEEE Transactions on 16 (2008) 229-238.

[12] O. Alexey, E. Vincent, and F. Bimbot. Audio, Speech, and Language Processing, IEEE Transactions on 20 (2012) 1118-1133.