Gaussian Mixture Model for Singing Voice Separation from Stereophonic Music

Minje Kim[1], Seungkwon Beack[1], Keunwoo Choi[1], and Kyeongok Kang[1]

[1]*Realistic Acoustics Research Team, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea*

Correspondence should be addressed to Minje Kim (mkim@etri.re.kr)

ABSTRACT

This paper presents an adaptive prediction method about source-specific ranges of binaural cues, such as inter-channel level difference (ILD) and inter-channel phase difference (IPD), for centrally positioned singing voice separation. To this end, we employ Gaussian mixture model (GMM) to cluster underlying distributions in the feature domain of mixture signal. By regarding responsibilities to those distinct Gaussians as unmixing coefficients of each mixture spectrogram sample, the proposed method can reduce artificial deformations that previous center channel extraction methods usually suffer, caused by their imprecise or rough decision about ranges of central subspaces. Experiments on commercial music show superiority of the proposed method.

1. INTRODUCTION

Singing voice separation (SVS) or vocal source separation, which aims to separate lead singer's playing from music, draws much attention in various research fields and applications. First of all, in music information retrieval (MIR) area, well-separated vocal sources can be utilized in some important tasks, such as automatic singer identification [1] and main melody extraction [2].

Another important application of SVS can be found in the Karaoke market. We expect that a decent SVS method will let users cheaply enjoy their Karaoke services with better sound quality than the traditional MIDI-based ones. Furthermore, object-based audio services and their standard [3] further allow users not only to take away singing voice, but to control the other instruments. To this end, they also require music to be separated well in advance.

There have been two different approaches to separating singing voices: monophonic ans stereophonic methods. In the monophonic methods, tracking a dominant melody from multiple pitches plays great role in effective separation of vocal sources. For instance, a method of masking salient pitches showed promising results combined with reconstruction of the other instruments using binary weighted nonnegative matrix factorization (NMF) [5]. A more sophisticated estimation of the main melody was made with source-filter model along with matrix decomposition concepts as well [6].

One the other hand, stereophonic methods mainly rely on the assumption that main singers' voices are usually positioned at the central subspace; both of their channels are more similar than the other surround instruments are. The distinction between center and surround channels can be made by binaural cues, such as inter-channel intensity difference (IID), inter-channel phase difference (IPD), and inter-channel coherence (ICC). Azimuth discrimination and resynthesis (ADress) is one important technique that finds out a sound source which has a particular IID value [7]. While ADress provides acceptable separation performance in various recordings, it still suffers musical noise which is cause by its *hard decision* manner. A post-processing method, based on independent component analysis (ICA), was introduced to enhance the ADress results [8].

In this paper, we propose an alternative clustering scheme based on Gaussian mixture model (GMM) [9]. The GMM on binaural cues, inter-channel level difference (ILD) and IPD in this case, produces responsibilities of each sample to the center subspace, and ends up allowing the concept of *soft decision* to the mixture samples that do not totally belong to one specific source (we use the term ILD for log-energy difference as defined in (5) to distinguish it from IID as an amplitude discrimination in [7]).

This paper consists of following sections. Section 2 describes problems that can be caused by improper decision mechanism. Section 3 provides details about the pro-

posed separation method using GMM on binaural cues. Section 4 shows empirical assessment of the proposed soft decision method on real-world commercial music. Finally, Section 5 concludes the work.

2. SOFT VS HARD DECISION

Separating the c_ith channel of jth target source $S_j^{(c_i)}$ from c_ith channel of a short-time Fourier transformed (STFT) stereophonic mixture $X^{(c_i)}$ can be represented as an element-wise weighting process like,

$$S_j^{(c_i)}(t,f) = W_j(t,f)X^{(c_i)}(t,f), \text{ for } 0 \le W_j(t,f) \le 1, \quad (1)$$

where c_i indicates each channel, $c = [1,2]^\top$ in stereophonic case, t and f respectively designate a specific frame and frequency bin. Equation (1) covers instantaneous mixing environments where all unmixing coefficients $W_j(t,f)$ are the same with different t and f indices. Furthermore, (1) can also model more complicated mixing environments with the nonlinear filtering, by considering each $W_j(t,f)$ has a distinct value.

Even in the instantaneous mixture case, hard decision can cause problems with inappropriate prediction. For instance, after the decision is made based on a certain criteria α_j in the feature domain like,

$$\hat{S}_j^{(c_i)}(t,f) = \begin{cases} X^{(c_i)}(t,f), & \text{if } \left|\Phi\Big(X^{(c)}(t,f)\Big)\right| \le \alpha_j \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

a sample point of c_ith channel of reconstructed source $\hat{S}_j^{(c_i)}(t,f)$ is copied from the mixture sample as is or has zero value. Note that feature transform function $\Phi(\cdot)$ takes all two channels of the mixture signal, $X^{(c)}(t,f) = \left[X^{(1)}(t,f), X^{(2)}(t,f)\right]^\top$, in the stereophonic case. Suppose that the true unmixing coefficient for jth source, $W_j(t,f)$ is less than 1. At the same time, if the sample point is decided as the target source based on the hard decision manner, unnecessary part of interfering sources $(1-W_j(t,f))X^{(c_i)}(t,f)$ will be also extracted. Otherwise, some part of the target source $W_j(t,f)X^{(c_i)}(t,f)$ will be omitted in the reconstructed source $\hat{S}_j^{(c_i)}(t,f)$.

Our goal is to provide a soft decision mechanism, where each unmixing coefficient $W_j(t,f)$ is estimated to have a soft real number from 0 to 1, instead the two integers, 0 or 1. Similarly to (2), the reconstruction can be made from the weighting process using the delicately estimated unmixing coefficient $\hat{W}_j(t,f)$, like

$$\hat{S}_j^{(c)}(t,f) = \hat{W}_j(t,f)X^{(c)}(t,f). \quad (3)$$

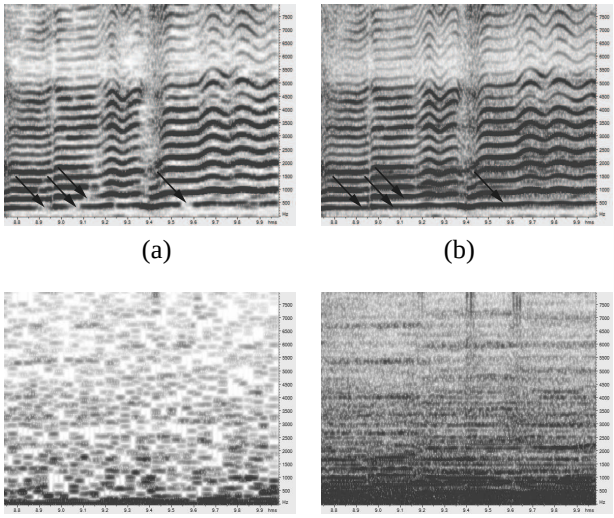
We propose a GMM-based clustering technique in Section 3, where probabilities that each sample belongs to the Gaussian distributions are regarded as unmixing coefficients for the sources. Consequently, the goal of our soft decision mechanism is to get less separation error than hard decision, like

$$\begin{aligned} & \sum_{t,f,i} \left| \Big(\hat{W}_j(t,f) - W_j(t,f)\Big) X^{(c_i)}(t,f) \right|^2 \quad (4) \\ < & \sum_{(t,f)\in\mathscr{C}_j,i} \left| \Big(1 - W_j(t,f)\Big) X^{(c_i)}(t,f) \right|^2 \\ + & \sum_{(t,f)\notin\mathscr{C}_j,i} \left| W_j(t,f)X^{(c_i)}(t,f) \right|^2, \end{aligned}$$

where $\mathscr{C}_j$ means the cluster consists of samples that are classified into the jth target source. The two terms of the right hand side represent errors caused by interfering sources and loss of the target source during the reconstruction process, respectively.

In most of previous methods, for example ADress [7], a range parameter α_j is certainly exploited to tackle *frequency azimuth smearing*, which occurs when there are harmonic overlaps in a given frequency. Although the azimuth subspace width, which ADress provides as a range parameter α_j, helps robust estimation of the azimuth values of sources, it is true that wider range of α_j does not guarantee to avoid problems of hard decision. Instead, wider α_j can increase error from interfering sources, $\sum_{(t,f)\in\mathscr{C}_j,i} |(1-W_j(t,f))X^{(c_i)}(t,f)|^2$ in (4).

Fig. 1 depicts the problems that hard decision can cause. We set a specific criteria α_j on ILD and IPD values, and then collect spectrogram samples that lie in the criteria like in (2). To see the effect of hard decision more clearly, the decision was made not on the mixture spectrogram, but on each of the two sources, singing voice and summed harmonic instruments. If α_j is wide enough to cover all spectrogram samples of vocal source, the reconstructed spectrogram Fig. 1(a) should be the same with the original one in Fig. 1(b). However, there are serious discontinuous regions marked with arrows, where



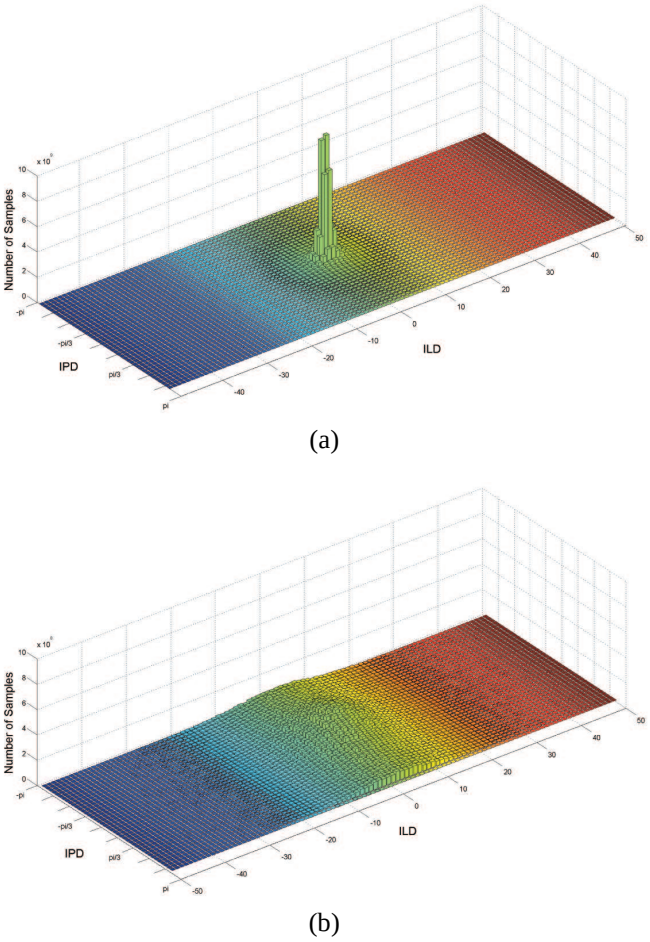
(a) (b) (c) (d)

Fig. 1: Spectrograms of hard decision results on vocal and harmony sources. Loss of vocal harmonics are marked with arrows. (a) Hard decision results on vocal source. (b) Original vocal source. (c) Hard decision results on mixture of harmony sources. (d) Mixture of original harmony sources.

some spectrogram samples of vocal source are misclassified into surround channel group. Another kind of distortion is that minute stereophonic effects of singing voice which might be artificially added in studio cannot be captured well, because they are more likely to spread widely in stereophonic sound field. We can see that the original noise floor between the harmonic crests in Fig. 1(b) is not fully reconstructed in Fig. 1(a). Furthermore, the same value of α_j also produces interfering musical noise in Fig. 1(c), which are incorrectly involved spectrogram samples from the summed surround sources in Fig. 1(d). In practice, the hard decision-based separation on mixture spectrograms in real world separation tasks, spectrograms in Fig. 1(a) and Fig. 1(c) are summed up to reconstruct centered singing voice. Therefore, the reconstructed signals usually suffer irregular loss of vocal sources and irritating peaks from surround harmony sources.

3. CENTERED SOURCE SEPARATION USING GMM ON BINAURAL CUES

The proposed GMM-based clustering is carried out in the feature domain, $\Phi(X^{(C)}(t,f))$. We adopt two widely



(a)

(b)

Fig. 2: Histograms of feature vectors from (a) a centered singing voice source (b) a mixture of surround instruments.

known inter-channel difference measures, ILD and IPD, to compound a feature vector,

$$\Phi\left(X^{(C)}(t,f)\right) = \begin{bmatrix} 10\log_{10}\dfrac{\left|X^{(1)}(t,f)\right|^2}{\left|X^{(2)}(t,f)\right|^2} \\ \angle\left(X^{(1)}(t,f)X^{(2)*}(t,f)\right) \end{bmatrix}, \quad (5)$$

where each element represents ILD and IPD between the two channels of mixture spectrogram.

Fig. 2 provides pictorial examples of two distributions each of which is from a centered vocal source and sum of the other harmonic sources, respectively. Suppose that $S_1^{(C)}(t,f)$ is two-channeled spectrograms of centered

Table 1: GMM-based centered source separation procedure from a stereo mixture.

1. Initialize parameters
   (a) Prepare $S_v^{(C)}(t,f)$ and $S_h^{(C)}(t,f)$, which are spectrograms of stereophonic vocal and harmony source signals for training, respectively
   (b) Calculate binaural cues $\Phi(S_v^{(C)}(t,f))$ and $\Phi(S_h^{(C)}(t,f))$ of training signals
   (c) Calculate means and covariances of training feature vectors, μ_v, μ_h, Σ_v, Σ_h.
      i. If 1. (a) to (b) were done, initialize μ_1, μ_2, Σ_1, Σ_1 with μ_v, μ_h, Σ_v, Σ_h.
      ii. Otherwise, initialize them with random values.
   (d) Initialize mixing parameters $p(j)$ with equal probabilities, 0.5.
2. Prepare input samples for GMM
   (a) Calculate binaural cues $x_{(t-1)F+f} := \Phi(X^{(C)}(t,f))$ of stereophonic mixture signal
3. EM for GMM learning (repeat until convergence)
   (a) E-step: compute responsibilities for all components j and samples x_n
   $r_{jn} = \frac{p(x_n|j)p(j)}{\sum_{j=1}^{M} p(x_n|j)p(j)}$
   (b) M-step: update parameters:
   $\mu_j^{\text{new}} = \frac{\sum_n r_{jn} x_n}{\sum_n r_{jn}}$
   $\Sigma^{\text{new}} = \frac{\sum_n r_{jn}(x_n-\mu_j^{\text{new}})(x_n-\mu_j^{\text{new}})^\top}{\sum_n r_{jn}}$
   $p^{\text{new}}(j) = \frac{1}{N}\sum_n r_{jn}$
4. Reconstruct jth source by substituting $\hat{W}_j(t,f)$ in (3) with $r_{j,(t-1)F+f}$

singing voice and $S_2^{(C)}(t,f)$ is that of summed surround instrumental sources. Fig. 2(a) is a histogram of feature vectors $\Phi\big(S_1^{(C)}(t,f)\big)$ from overall spectrogram samples of the vocal source, $S_1^{(C)}(t,f)$. Compared with the distribution of $\Phi\big(S_2^{(C)}(t,f)\big)$ in Fig. 2(b), ILD and IPD values of singing voice construct way narrower multivariate Gaussian-like sample distribution. Therefore, the variances of the two distributions can be reasonable criteria for separating sources $S_1^{(C)}(t,f)$ and $S_2^{(C)}(t,f)$.

GMM aims at clustering each spectrogram sample based on two learned Gaussian distributions. That means that the binaural cues of the mixture signal consist a mixture distribution of two Gaussians which differ in their means or variances. Therefore, a certain kind of ordinary GMM learning results, *responsibility*, can be eventually used as unmixing coefficients $\hat{W}_j(t,f)$. For instance, a sample whose ILD and IPD values are close to the mean of a specific Gaussian is more likely to belong to it. In the case of Fig. 2, where means of two distributions are very similar, the distance to the common mean can also play a great role when GMM identifies responsibility: it is more possible that another sample whose ILD and IPD values are far from the common mean will be allocated to the Gaussian distribution with bigger variance in Fig. 2(b).

Table 1 summarizes the overall procedure for centered source separation using GMM on binaural cues. Note that this procedure can be easily expanded to the cases where spatial distributions of more than two sources are known. In addition, if the initialization was made with random values (1.(c) ii), it is necessary to identify which j is the index for the target source.

Table 2: Separation performances of hard decision with various ranges and GMM-based methods.

Song	Hard decision W/O GMM			GMM		
	Narrow	Optimal	Wide	Soft	Hard	Random (soft)
1	2.29	6.35	5.43	6.70	**6.95**	6.66
2	1.68	4.59	3.46	5.43	4.84	**5.44**
3	1.82	6.42	5.54	**6.54**	6.34	6.52
4	1.81	4.30	3.34	5.86	5.19	**5.89**
5	1.52	5.35	4.49	7.17	6.59	**7.18**
6	0.64	3.52	4.32	**4.72**	4.24	4.67
7	1.63	3.78	2.19	4.88	4.22	**4.89**
8	0.26	0.96	0.36	3.37	1.92	**3.41**
9	2.04	**7.68**	7.25	7.20	7.99	7.15
10	0.66	3.36	2.41	**4.26**	3.84	4.23
Average	1.44	4.63	3.88	**5.61**	5.21	5.60

4. EXPERIMENTAL RESULTS

We use 10 seconds-long excerpts of 10 commercially released Korean pop songs for test signals. Also, we use 13 other songs for training. All of them are stereophonic PCM wave signals with 44.1kHz sampling rate and 16bit encoding. Before the centered singing voice separation, drum sources were taken away using nonnegative matrix parcial co-factorization (NMPCF) algorithm as proposed in [10] [11]. Being windowed with sine squared function, 4096 samples of the signals are short-time Fourier transformed with 50% overlap. To assess the separation quality, we adopt signal-to-distortion ratio defined by,

$$\mathrm{SDR} := \frac{1}{C}\sum_{c_i} 10\log 10 \frac{\sum_t s^{(c_i)}(t)^2}{\sum_t (s^{(c_i)}(t) - \hat{s}^{(c_i)}(t))^2}. \qquad (6)$$

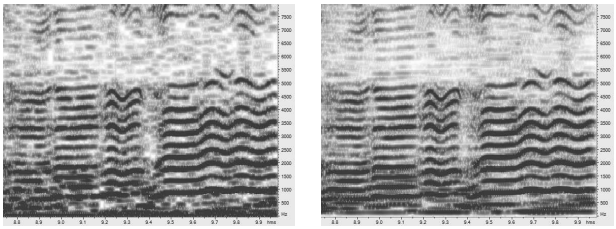
Equation (6) can be viewed as the same definition in [12] without allowing any possible deformation of the source, since the secured source signals are artificially filtered ones, right before the mixing process. On top of that, our goal is to separate out not only clean vocal signals, but all of their stereophonic sound effects. All of the training and test signals went through high pass filtering to cut off unnecessary low frequency parts under 140Hz.

For the hard decision tests, we empirically picked up the optimal α_j among various ILD and IPD ranges, namely $|\mathrm{ILD}| < 0.04$dB or $|\mathrm{IPD}| < 20^\circ$.

GMMs are individually learned for two subbands, under and over 8kHz. Therefore, the separation procedure in Table 2 is executed twice. Finer subband resolutions were not satisfying since the number of samples in each subband is not big enough to learn GMMs well. For the case of random initialization, resulting clusters are manually ordered by regarding the ones with smaller variances as the target source.

Table 2 shows separation performances. First of all, we can compare the optimal combination of the range parameter with exemplar narrower and wider ones, (0.01dB, 3°) and (0.32dB, 42°), respectively. Although the optimal combination provides the best results among the three, it is impossible in practice to know the optimal one a priori. Contrarily, the soft decision methods we proposed perform better than every hard decision case even in the case of random initialization. Besides, the good results with random initialization are also meaningful for us because they support the idea that there are two underlying Gaussians in feature domain of mixture music. With the learned GMMs, we can also choose not to use soft responsibilities; if we round off them to have 0 or 1, we can get hard decision results based on GMM. Although adopting hard decision after GMM degrades separation performances, it is still better than the ordinary hard decision method without GMM.

Fig. 3 further supports superiority of the proposed method. We can check that temporal discontinuity and peaky cells in the reconstructed spectrogram of singing voice, Fig. 3(a), disappear significantly in Fig. 3(b), that of reconstruction with soft decision. Compare them with the original source in Fig. 1(b).



(a) (b)

Fig. 3: Spectrograms of the reconstructed centered singing voice in song 7. (a) Spectrogram of the hard decision result without GMM. (b) Spectrogram of the soft decision result using GMM.

5. CONCLUSION

A delicate centered source separation method was introduced. Based on the assumption that the target source has a specific position in stereophonic sound field, such as centered singing voice, binaural cues of input mixture signals were clustered using GMM. Experimental results on the real-world commercial music showed improvement upon the ordinary hard decision method in separation performance. Also, we expect that the relatively lower complexity of the proposed method than that of complicated vocal source separation methods [6][8] can be an advantage when we implement a lightweight Karaoke application for hand-held devices while retaining acceptable separation quality.

6. ACKNOWLEDGEMENT

This research was supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2010.

7. REFERENCES

[1] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[2] J. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, 2008.

[3] I. Jang, J. Seo, and K. Kang, "Design of a file format for interactive music service," *ETRI Journal*, vol. 33, no. 1, pp. 128–131, 2011.

[4] *Information technology – Multimedia application format (MPEG-A) – Part12: Interactive music application format*, ISO/IEC IS 23 000-12, 2010.

[5] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *European Signal Processing Conference*, 2005.

[6] J. Durrieu, A. Ozerov, C. Fvotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proceedings of EUSIPCO*, 2009.

[7] D. Barry and B. Lawlor, "Sound source separation: Azimuth discrimination and resynthesis," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Naples, Italy, 2004.

[8] S. Sofianos, A. Ariyaeeinia, and R. Polfremann, "Towards effective singing voice extraction from stereophonic recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[9] C. M. Bishop, *Neural Networks for Pattern Recognition*, 1st ed. Oxford University Press, 1996.

[10] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[11] M. Kim, J. Yoo, K. Kang, and S. Choi, "Blind rhythmic source separation: Nonnegativity and repeatability," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[12] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.