A Study on Vocal Separation from Mixtured Music

Hyun-Tae Kim, Jang-Sik Park, *Member, KIMICS*

Abstract— **Recently, According to increasing interest to original sound Karaoke instrument, MIDI type karaoke manufacturer attempt to make more cheap method instead of original recoding method. Separating technique for singing voice from music accompaniment is very useful in such equipment. We propose a system to separate singing voice from music accompaniment for stereo recordings. Our system consists of three stages. The first stage is a spectral change detector. The second stage classifies an input into vocal and non vocal portions by using GMM classifier. The last stage is a selective frequency separation stage. The results of removed by listening test from the results for computer based extraction simulation, spectrogram results show separation task successfully. Listening test with extracted MR from proposed system show vocal separating and removal task successfully.**

Index Terms— **vocal remover, original sound Karaoke instrument, GMM, frequency domain processing.**

I. INTRODUCTION

ALTHOUGH speech separation has been extensively studied, few studies are devoted to separating singing voice from music accompaniment. Singing voice bears many similarities to speech. For example, they both consist of voiced and unvoiced sounds. But the differences between singing and speech are also significant. A well known difference is the presence of an additional formant, called the singing formant, in the frequency range of 2000–3000 Hz in operatic singing. This singing formant helps the voice of a singer to stand out from the accompaniment [1].

Another difference is related to the way singing and speech are uttered. During singing, a singer usually intentionally stretches the voiced sound and shrinks the unvoiced sound to match other musical instruments. This has a direct consequence. It alters the percentage of voiced and unvoiced sounds in singing. The large majority of sounds generated during singing is voiced (about 90%) [2] while speech has a larger amount of unvoiced sounds [3].

Manuscript received February 27, 2011; revised March 25, 2011; accepted April 8, 2011.

Hyun-Tae Kim is with the Department of Multimedia Engineering, Dongeui University, Busan, 614-714, Korea (Email: htaekim@deu.ac.kr)

Jang-Sik Park is with the Department of Electronics Engineering, Kyungsung University, Busan, 608-736, Korea (Email: jsipark@ks.ac.kr)

From the sound separation point of view, the most important difference between singing and speech is the nature of other concurrent sounds. In a real acoustic environment, speech is usually contaminated by interference that can be harmonic or nonharmonic, narrowband or broadband. Interference in most cases is independent of speech in the sense that the spectral contents of target speech and interference are uncorrelated. For recorded singing voice, however, it is almost always accompanied by musical instruments that in most cases are harmonic, broadband, and are correlated with singing since they are composed to be a coherent whole with the singing voice. This difference makes the separation of singing voice from music accompaniment potentially more challenging [4].

II. PROPOSED SYSTEM DESCRIPTION

Our system is illustrated in Fig. 1. The input to the system is a mixture of singing voice and music accompaniment. In the singing voice detection stage, the input is first partitioned into spectrally homogeneous portions by detecting significant spectral changes. Then each portion is classified, based on the overall likelihood, as a vocal portion in which singing voice is present, or a non-vocal portion in which singing voice is absent. Different features have been explored for singing voice detection [5-7]. Several studies have shown that MFCC (mel-frequency cepstral coefficients) is a good feature for sound classification, even for mixtures. Li et al. [8] compared different features in classifying a sound into seven classes and found that MFCC provides the best classification. And we choose GMM as the classifier for it has been widely and successfully used with MFCC for audio classification tasks [5], [8]. The last stage consist of spectral power comparison between each channel of the stereo signal and inter-channels difference and spectrally subtracted at each vocal frequency bin respectively. It is shown in Fig. 2.

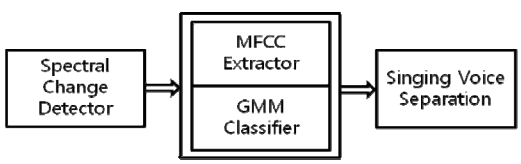

Fig. 1. Schematic diagram of the proposed system.

2.1 Spectral change detector

We use a simple spectral change detector proposed by Duxbury *et al.* [3]. This detector calculates the Euclidian distance $\eta(m)$ in the complex domain between the expected spectral value and the observed one in a frame

$$\eta(m) = \sum_k (\left| \hat{S}_k(m) - S_k(m) \right|) \tag{1}$$

where $S_k(m)$ is the observed spectral value at frame m and frequency bin k. $\hat{S}_k(m)$ is the expected spectral value of the same frame and the same bin, calculated by

$$\hat{S}_k(m) = |S_k(m-1)| \hat{\phi}_k(m) \tag{2}$$

where $|S_k(m-1)|$ is the spectral magnitude of the previous frame at bin k. $\hat{\phi}_k(m)$ is the expected phase which can be calculated as the sum of the phase of previous frame and the phase difference between the previous two frames

$$\hat{\phi}_k(m) = \tilde{\varphi}_k(m-1) + (\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2)) \tag{3}$$

where $\tilde{\varphi}_k(m-1)$ and $\tilde{\varphi}_k(m-2)$ are the unwrapped phases for frame m-1 and frame m-2, respectively. $\eta(m)$ is calculated for each frame.

A local peak in $\eta(m)$ indicates a spectral change, which can either be that the spectral contents of a sound are changing or a new sound is entering the scene. To accommodate the dynamic range of the spectral change as well as spectral fluctuations, we apply weighted dynamic threshold to identify the instances of significant spectral changes. Specifically, a frame m will be recognized as an instance of significant spectral change if $\eta(m)$ is a local peak, and $\eta(m)$ is greater than the weighted median value in a window of size H

$$\eta(m) > C_i median(\eta(m - \frac{H}{2}), \cdots, \eta(m + \frac{H}{2})) \tag{4}$$

where C is the weighting factor. Finally, two instances are merged if the enclosed interval is less than T_{min}; specifically, if two significant spectral changes occur within T_{min}, only the one with the larger spectral change value $\eta(m)$ is retained.

2.2 Vocal and non-vocal classifier

After the input is partitioned, we pool the information in a whole portion to obtain more reliable classification. A portion is classified as vocal if the overall likelihood of the vocal class is greater than that of the non-vocal class. Fomally let $\{X1;X2;\ :::;XM\}$ be a set of feature vectors for a portion with M frames. Let log $p(X|cv)$ and log $p(X|cnv)$ represent the log likelihood of an observed feature vector X being in the vocal class cv and the non-vocal class cnv, respectively. Then a portion is classified as vocal if :

$$\sum_{j=1}^{M} log p(X_j | c_v) > \sum_{j=1}^{M} log p(X_j | c_{nv}) \tag{5}$$

We choose MFCC as the feature vector and the GMM as the classifier since they have been widely and successfully used for audio classification tasks. A Gaussian mixture model with K components, each having a diagonal covariance matrix, is used to model the MFCC distribution of the two classes: c_v and c_{nv}.

GMM is implemented with HTK toolkit, which was developed from Cambridge University in England. The version of HTK for this is HTK 3.2. GMM setup process with artificially separated vocal from non-vocal data is shown in Fig. 2. The right side of dotted line shows the corresponding function for left side of the line in HTK toolkit.

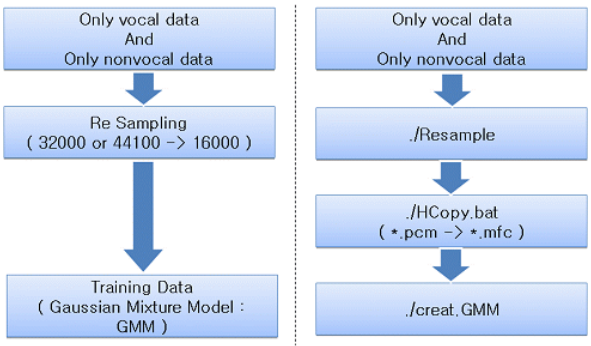

Fig. 2. GMM training process for vocal and non-vocal classification.

After GMM setup processing, vocal and non-vocal test process is followed as described in Fig. 3.

Fig. 3. GMM testing process for vocal and non-vocal classification.

2.3 Spectral power comparison

The last stage consists of spectral power comparison between each channel of the stereo signal and inter-channels difference and spectrally subtracted at each vocal frequency bin respectively. The detailed diagram is shown in Fig. 4. And the detailed procedure of the last stage is summarized by Table **I.**

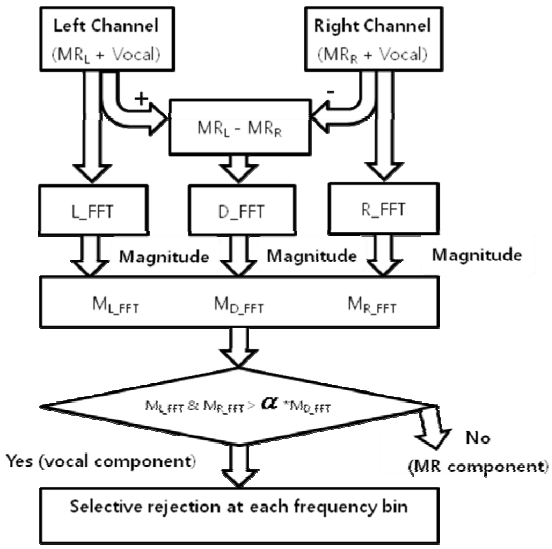

Fig. 4. Detailed diagram of the last stage.

TABLE I
THE DETAILED PROCEDURES OF THE LAST STAGE

No. of procedures	Details
Step 1	Compute $MR_L - MR_R$ in time domain
Step 2	Transform each channel of the stereo signal in time domain into frequency domain by FFT
Step 3	Co Compute magnitudes of each channel and $MR_L - MR_R$ channel in frequency domain
Step 4	Implement spectral power comparison between each channel of the stereo signal and inter channels difference
Step 5	Reject selectively at each vocal frequency bin in stereo channel

III. COMPUTER SIMULATION AND RESULTS

First, we test the first and second stage of the proposed system referred in fig. 1 with a ballad song. Second, we experiment total process of the proposed system with 30 songs from a variety of music genre by famous singers. Almost of the songs are Korean songs, but just 4 songs are English songs.

Fig. 5 shows the classification result for a clip of ballad music. The singing voice is shown in Fig. 5(a), it is mixed with music accompaniment. In Fig. 5(b), the spectrogram of the mixture is plotted. The vertical lines in Fig. 5(c) show the instances of significant spectral changes identified by our spectral change detector. The input is over-partitioned to some extent, but the beat times and the time instances when the singing voice enters are well captured except at times between 2 and 2.3 s.

Fig. 5(d) shows artificially classified vocal and non-vocal portions. Fig. 5(e) shows the final classification, which matches the reference labeling indicated in Fig. 5(d) well except at around 1.7 s for a very short non-vocal portion.

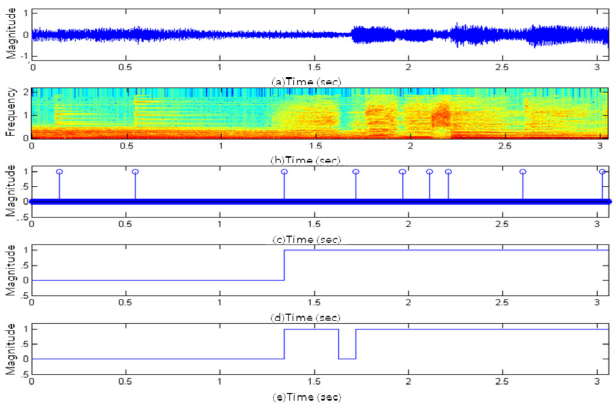

Fig. 5. Singing voice detection for a clip of ballad music. (a) The waveform of the singing voice and the accompaniment. (b) The spectrogram of the mixture. Brighter area indicates stronger energy. The vertical lines in (c) indicate the spectral change moments identified by the spectral change detector. (d) Artificially classified frame-level classification of the clip. A high value indicates the frame is classified as vocal and a low value as non-vocal. (e) The final classification using the spectral change detection and the overall likelihood.

TABLE Ⅱ
DESCRIPTION OF THE RATINGS USED IN THE MOS

MOS	Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

In order to evaluate the performance of proposed system a Mean Opinion Score (MOS) has to be performed with 10 listeners. Listener groups consisted of a professor, two graduate students and seven undergraduate students. And their major or interesting fields is audio signal processing. Before listening tests, we make standard signal with vocal removal quality

associated with MOS level for each music genre. Tests were performed with standard signals firstly, and then 30 songs were tested. Also, the tests were done for each listener separately. 5 levels of the MOS shown in Table Ⅱ.

The results of listening test for a listener are shown in Table Ⅲ. The results were averaged for the ten listeners and given in Table Ⅳ. From the results, hip-hop, rock and pop music tend to worse than trot and ballad. This is because the energy of the MR parts has a similar to vocal at each frequency bin.

The MOS scores indicate that the proposed system removed vocal slightly well and distorted background music minimally.

TABLE Ⅲ
A RESULT OF LISTENING MOS TEST FOR A LISTENER

title	singer	genre	Score				
			unsatisfactory	poor	fair	good	excellent
Cry	Viva soul	Hip-Hop			O		
when I stare in your eyes	Lee, Jung	Ballard				O	
Look at the sky!	Kang, Chan	Ballard				O	
Just heard the name	Kang, Hyunju	Ballard				O	
The words "I love you"	Kim, Dongryul	Ballard				O	
Something's gotta give	Dynamic Duo	Hip-Hop			O		
Obstinate person	Drunken Tiger	Hip-Hop			O		
Let's have dop	Vaive	Ballard				O	
Slow down feat	Buga Kings	Hip-Hop				O	
Love	Bu Hwal	Rock			O		
At the street	Sung Sikyung	Ballard				O	
what is love?	Yangpa(Onion)	Ballard				O	
Love is then all the way	Oh, Jonghyeok	Ballard				O	
Love two	Yun, Dohyun Band	Rock			O		
Y	Free style	Rap/Ballard			O		
Street life	DJ DOC	Rap			O		
Two lanes bridge	Cha, Taehyun	Trot				O	
Goodness!	Jang, Yunjung	Trot					O
Zan zara	Jang, Yunjung	Trot					O
Fall down	Seo, Jukyung	Trot				O	
Dulie	Nam, Jin	Trot				O	
Woman who erases makeup	Kang, Jin	Trot				O	
One million rose	Sim, Subong	Trot				O	
Please give back youth	Na, Huna	Trot				O	
Love Song	Sara Bareilles	Pop			O		
Paralyzer	Finger Eleven	Pop				O	
See You Again	Miley Cyrus	Pop			O		
Tattoo	Jordin Sparks	Pop		O			
Day by day	Big Bang	Dance				O	
Please don't go! don't go	Brown Eyes	Ballard				O	
		Total	0	1	9	18	2
		%	0	3.3	30	60	6.7

TABLE Ⅳ
AVERAGE MOS FOR 10 LISTENERS OF THE PROPOSED SYSTEM

Listener	Average MOS
A	3.7
B	3.5
C	3.4
D	3.2
E	3.5
F	3.3
G	3.4
H	3.5
I	3.4
J	3.5
Total average	3.44

IV. CONCLUSIONS

We propose a system to separate singing voice from recorded music. Our system consists of three stages. The first stage is a spectral change detector. The second stage classifies an input into vocal and non vocal portions by using GMM classifier. The last stage consists of spectral power comparison between each channel of the stereo signal and inter-channels difference and spectrally subtracted at each vocal frequency bin respectively. By PC based MOS test, proposed system removed vocal slightly well and distorted background music minimally.

ACKNOWLEDGMENT

This work was supported by Dong-eui University Grant (2010AA199).

REFERENCES

[1] J. Sundberg, "The acoustics of the singing voice," *Scientific American*, pp. 82–91, Mar. 1977.

[2] Y. E. Kim, "Singing voice analysis/synthesis," Ph.D. dissertation, MIT, Media Lab, 2003.

[3] D. L. Wang, "Feature-based speech segregation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. New York: IEEE Press (dual imprint with Wiley), 2006, to appear.

[4] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *Proc. 6th Conf. Digital Audio Effect (DAFx-03)*, London, U.K., 2003.

[5] A. L. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2001, pp. 119–122.

[6] A. L. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *Proceedings of AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[7] N. C. Maddage, C. Xu, and Y. Wang, "A SVM-based classification approach to musical audio," in *Proc. ISMIR*, 2003.

[8] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, pp. 533–544, 2002.

[9] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Lett.*, vol. 22, pp. 533–544, 2002.

[10] M. Slaney, "Auditory Toolbox for MATLAB," Jan. 1999 [Online]. Available: http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/

[11] K. Murphy, "HMM Toolbox for MATLAB," Jun. 2005 [Online]. Available: http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

Hyun-Tae Kim received the B.S., the M.S. and the Ph.D. degree in the Electronics Eng. from Pusan National University, Korea in 1989, 1995 and 2000, respectively. He joined the Dongeui University in Korea as professor in the Multimedia Engineering Department since March 2002. He was a visiting professor at Georgia Institute of Tech. in USA at 2008.

Jang-Sik Park received the B.S., the M.S. and the Ph.D. degree in the Electronics Eng. from Pusan National University, Korea in 1992, 1994 and 1999, respectively. He joined the Kyungsung University in Korea as professor in the Electronics Engineering Department since March 2011.