A Real Time Singing Voice Removal System Using DSP and Multichannel Audio Interface

Hyuntae Kim[1] , Taehoon Kim[2] and Jangsik Park[3]
[1]*Department of Multimedia Engineering, Dongeui University*
Gaya-dong, San 24, Busanjin-ku, Busan, 614-714, Korea
[2] *Department of Electronics Engineering, Busan, Dongeui Institute of Technology*
Yangji-ryo 54, Busanjin-ku, Busan, 614-715, Korea
[3] *Department of Electronics Engineering, Kyungsung University,*
Daeyeon3-dong, 110-1, Nam-gu, Busan, 608-736, Korea
htaekim@deu.ac.kr, kthn018@nate.com, jsipark@ks.ac.kr

Abstract

Separating technique for singing voice from music accompaniment is very useful in original sound type Karaoke instrument. We propose a real-time system to separate singing voice from music accompaniment for stereo recordings. Proposed algorithm consists of two stages. The first stage is a spectral change detector. The last stage is a selective vocal separation in frequency bins. Our system consists of a DSP board and a multichannel audio interface board. The DSP board is a TMS320C6713 DSK. And multichannel audio interface has six (three stereos) channels. The proposed vocal removal algorithms are embedded in DSP. Listening tests with extracted MR from proposed system show vocal separating and removal tasks successfully.

Keywords: *Singing voice remover, Real-time system, DSP board, Frequency domain processing*

1. 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 speeches 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 are voiced (about 90%) [2]. While speech has a larger amount of unvoiced sounds [3].

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 means separation of singing voice from music accompaniment is more difficult than speech separation [4].

There were a few attempts for vocal separation. Using MFCC (mel-frequency cepstral coefficients) and GMM (Gaussian mixture model) as a classifier were proposed [5]. In recent years, by observing the change of pitch information, how to separate singing voice were suggested [6].

In this paper, the input is first partitioned into spectrally homogeneous portions detecting significant spectral changes. Then, from energy comparison between differential signals between stereo AR (all recorded) signal and each channel signals in frequency domain, the presence or absence of singing voice were determined. The proposed systems were implemented in DSP board. And input AR stereo signal, extracted MR (music recorded) signal from DSP output, and extracted singing voice could be listened in real-time by multichannel audio interface cooperated with DSP board.

2. Proposed System

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*

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

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

2.2 Singing Voice Removal Algorithm

Proposed singing voice removal algorithm consists of power comparison between each channel of the stereo signal and inter-channels differential signal and then spectrally removed at each singing voice frequency bin when larger than threshold value respectively. The detail diagram is shown in Figure 1.

2.3 DSP Implementation

DSP codes set up using CCS (code composer studio) DSK v.3.1 with proposed singing voice removal algorithm. It is shown partly in Table 1, the parts for presence or absence of singing voice and removing them if present in each frequency bin. The completed codes downloaded in DSP board, TMS320C6713 DSK. And audio interface board designed with three stereo channel for a stereo AR (all recorded) signal, a stereo MR (music recorded) signal, and a separated stereo singing voice signal. It is interconnected with DSP board for real-time playback. It is shown in Figure 2.

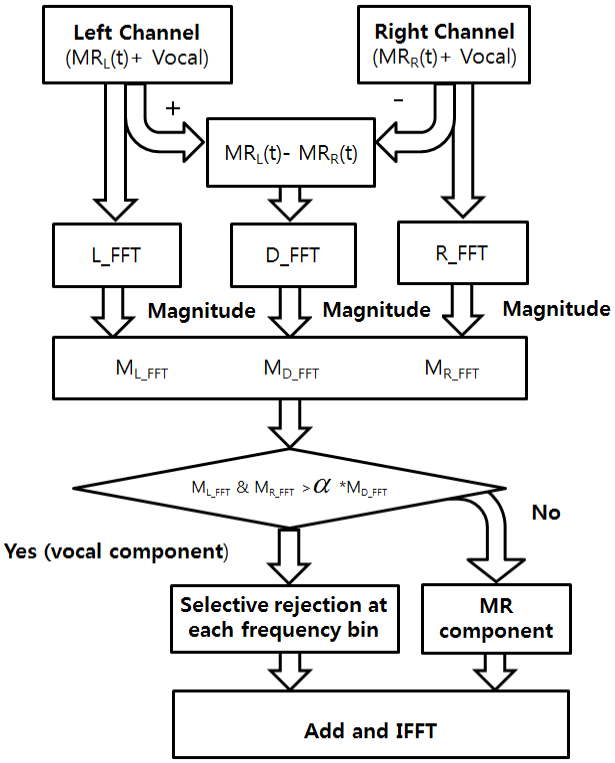

Figure1. The detail diagram of the proposed singing voice removal algorithm

Table 1. A part of DSP codes for proposed algorithm

```
// Remove if larger than threshold value!
    for(i=0;i<FFT_N;i++) {
                if(diff_in_pw[i] < x_L_pw[i]) {
                            if(diff_in_pw[i] < x_R_pw[i]) {
// If left channel signal and right channel signal larger than differential signal,
                            vocal_L[2*i] = x_L[2*i];
                            vocal_L[2*i+1] = x_L[2*i+1];
                            vocal_R[2*i] = x_R[2*i];
                            vocal_R[2*i+1] = x_R[2*i+1];

// Make 0 to remove!
                            x_L[2*i] = 0;
                            x_L[2*i+1] = 0;
                            x_R[2*i] = 0;
                            x_R[2*i+1] = 0;
                            }
                    }
            }
```

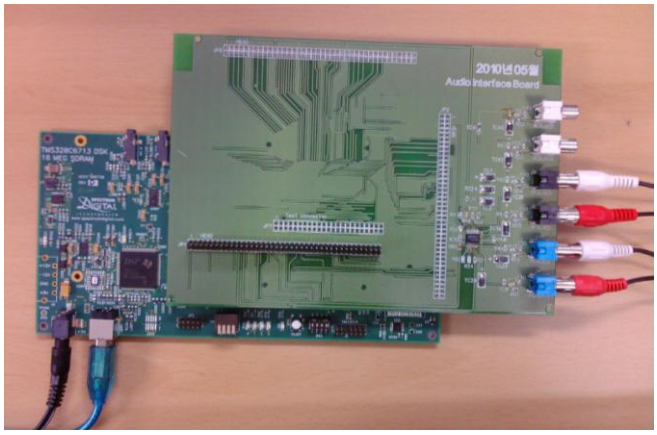

Figure 2. An implemented singing voice remover using DSP-TMS320C6713

3. Experiment Result

We experiment with 30 songs from a variety of music genre by famous singers for the proposed system test. Almost of the songs are Korean songs, but just 4 songs are English songs. 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. Ratings of the MOS are shown in Table 2.

The results were averaged for the ten listeners and given in Table 3. 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 in those genres.

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

Table 2. 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

Table 3. Average MOS results 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

4. Conclusion

In this paper, a real-time system to separate singing voice from music accompaniment for stereo recordings was proposed. Proposed algorithm consists of two stages. The first stage is a spectral change detector. The last stage is a selective vocal separation in frequency bins. Our system consists of a DSP board and a multichannel audio interface board. From energy comparison between differential signals between stereo AR (all recorded) signal and each channel signals in frequency domain, the presence or absence of singing voice were determined. The proposed vocal removal algorithms are embedded in DSP. Listening test with extracted MR from proposed system show vocal separating and removal task successfully.

Acknowledgements

This work was supported by Dong-eui University Grant. (2011AA201).

References

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

[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, Wang, D. L. and Brown, G. J., Eds. New York: IEEE Press (dual imprint with Wiley) **(2006)**.

[4] Y. Li and D. L. Wang, "Separation of Singing Voice from Music Accompaniment for Monaural Recordings", In: IEEE Transaction on audio, speech, and language processing, Vol. 15, No. 4, **(2007)** May.

[5] B. Jounghoon and K. Hanseok, "Spectral Subtraction Using Spectral Harmonics for Robust Speech Recognition in Car Environments. In: ICCS 2003, LNCS 2660, pp.1109-1116 **(2003)**.

[6] C. L. Hsu and J. S. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset", In: IEEE Transactions on audio, speech, and language processing, Vol. 18, Issue 2 **(2010)**.

Authors

Hyuntae 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.

Jangsik 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.

Taehoon Kim received the B.S., the M.S. and the Ph.D. degree in the Electronics Eng. from Pusan National University, Korea in 1995, 1997 and 2002, respectively. He joined the Dongeui Institute of Technology in Korea as professor in the Electronics Engineering Department since March 2012.