SEPARATION OF VOICE AND MUSIC BY HARMONIC STRUCTURE STABILITY ANALYSIS

Yun-Gang Zhang[*]*, Chang-Shui Zhang*

Department of Automation
Tsinghua University, Beijing 100084, China
zyg00@mails.tsinghua.edu.cn, zcs@mail.tsinghua.edu.cn

ABSTRACT

Separation of voice and music is an interesting but difficult problem. It is useful for many other researches such as audio content analysis. In this paper, the difference between voice and music signals is carefully studied. It is proposed that the Harmonic Structure Stability is the key difference between them. A separation algorithm based on this theory is proposed. The main idea is to learn the average harmonic structure of the music, and then separate signals by using it to distinguish voice and music harmonic structures. Experimental results show that the algorithm can separate mixed signals and obtains not only a very high Signal-to-Noise Ratio (SNR) but also a rather good subjective audio quality.

1. INTRODUCTION

Separation of voice and music in a mixed signal is an important problem in music research. Here, *voice* means the singing voice in a song, *music* means the instrument accompaniments, i.e. *acappella* and *instrumental* in music terms, respectively. Separation of voice and music is helpful for many other music researches, such as Music Retrieval, Classification and Segmentation, Multi-pitch estimation, etc. [1, 2]. Signal separation is a difficult problem and no reliable methods are available for the general case.

However, voice and music are so different that a human can easily distinguish them. So, if it is possible to separate them by analyzing their difference? In this paper, the difference between voice and music is carefully studied and a new feature called Harmonic Structure Stability is proposed to represent this difference. A corresponding new algorithm is proposed to separate signals. The algorithm consists of four steps: preprocessing, harmonic structure extraction, music Average Harmonic Structure analysis, separation of signals. The main idea is to learn the average harmonic structure of the music, and then separate signals by using it to distinguish voice and music harmonic structures.

[*]This work is supported by the project (60475001) of the National Natural Science Foundation of China.

Music is a fast variation signal, it is difficult for traditional speech enhancement methods to enhance speech with music noise [3]. Compared to previous multi-pitch estimation methods, our method learns a model from the primary multi-pitch estimation results, and uses the model to improve the results. Gil-Jin and Te-Won proposed a probabilistic approach to single channel blind signal separation. The main idea is to exploit the inherent time structure of sound sources by learning a priori sets of basis filters [4]. In our approach, no training sets are needed. All information is directly learned from the mixed signal. Feng and etc. applied FastICA to extract singing and accompaniment from a mixture [5]. Vanroose used ICA to remove music background from speech by subtracting ICA components with the lowest entropy [6]. Compared to these approaches, our method preserves the harmonic structure in the separated signals and obtains a good subjective audio quality.

The rest of this paper is organized as follows: The essential difference between voice and music is analyzed in section two. The detail of the algorithm is described in section three. Experimental results are shown in section four. Finally, conclusion and discussions are given in section five.

2. HARMONIC STRUCTURE STABILITY ANALYSIS FOR VOICE AND MUSIC

The first task is to reveal the essential difference between voice and music signals. It is an interesting but challenging task. First, the frequency ranges of voice and music are overlapping. They can not be separated by frequency range analysis. Second, the human voice is a harmonic and non-harmonic mixture. Also, most music sounds are harmonic [7]. They can't be separated by harmony analysis. Third, the spectral peak tracks of both signals may stay at a certain note for a period of time [7]. They can't be separated by stable duration analysis. However, music signals are more ordered than voice. The entropy of music is much more constant in time than that of speech [8]. We try to find a way to define this difference for the purpose of separation.

Represent a monophonic sound signal $s(t)$, which may

be a voice or a music signal, by a sinusoidal model [9]:

$$s(t) = \sum_{r=1}^{R} A_r(t)\cos[\theta_r(t)] + e(t) \quad (1)$$

$A_r(t)$ and $\theta_r(t) = \int_0^t 2\pi r f_0(\tau)d\tau$ are the instantaneous amplitude and phase of the r^{th} harmonic, respectively, R is the maximal harmonic number, $f_0(\tau)$ is the fundamental frequency, $e(t)$ is the non-harmonic or noise component.

Divide $s(t)$ into overlapped frames and calculate f_0^l and A_r^l by detecting peaks in the magnitude spectrum. $A_r^l = 0$, if there doesn't exist the r^{th} harmonic. $l = 1, \ldots, L$ is the index of the frame. f_0^l and $[A_1^l, \ldots, A_R^l]$ describe the position and amplitudes of harmonics. Normalize A_r^l by multiplying a factor $\rho^l = C/A_1^l$ (C is an arbitrary constant) to ensure separation procedure will not be influenced by the amplitude. Then, translate the amplitudes into a log scale, because the human ear has a logarithmic sensitivity to sound as frequency varying. Define Harmonic Structure Coefficient as equation (2). As we know, the timbre of a sound is mostly controlled by the number of harmonics and the ratio of their amplitudes. So $\mathbf{B}^l = [B_1^l, \ldots, B_R^l]$, which is free from the fundamental frequency and amplitude, exactly represents the timbre of a sound. In this paper, these coefficients are used to represent the harmonic structure of the sound. Average Harmonic Structure and Harmonic Structure Stability are defined as follows to model music signals and measure the stability of harmonic structures.

- Harmonic Structure Coefficient:

$$B_i^l = \log(\rho^l A_i^l)/\log(\rho^l A_1^l), i = 1, \ldots, R \quad (2)$$

- Average Harmonic Structure (**AHS**): $\bar{\mathbf{B}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{B}^l$

- Harmonic Structure Stability (**HSS**):

$$HSS = \frac{1}{R}\sum_{l=1}^{L}\left\|\mathbf{B}^l - \bar{\mathbf{B}}\right\|^2 = \frac{1}{R}\sum_{r=1}^{R}\sum_{l=1}^{L}\left(\mathrm{B}_r^l - \bar{\mathrm{B}}_r\right)^2 \quad (3)$$

AHS and HSS (see Fig. 1) are the mean and variance of $\mathbf{B}^l$. Since timbres of most instruments are stable, so $\mathbf{B}^l$ varies little in different frames in a music signal and AHS is a good model to represent music signals. While, during singing a song, the vibration channel varies much, so $\mathbf{B}^l$ varies much in a voice signal. So, HSSs of music signals are small and HSSs of voice signals are big. This characteristic is useful for voice and music separation.

3. SEPARATION ALGORITHM

Suppose we have a signal mixture consisting of one voice and one music, in which both voice and music signals are monophonic. The separation algorithm consists of four steps: preprocessing, extraction of harmonic structures, music AHS analysis, separation of signals.

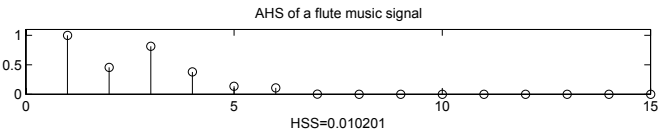


Fig. 1. AHS and HSS of a flute music signal.

In preprocessing step, the mean and energy of the input signal are normalized. In the second step, the pitch estimation algorithm of Terhardt's [10] is extended and used to extract harmonic structures. This algorithm is suitable for estimation both fundamental frequency and all it's harmonics. In Terhardt's algorithm, in each frame, all spectral peaks exceeding a given threshold are detected. The frequencies of these peaks are $[f_1, \ldots, f_K]$, K is the number of peaks. For a fundamental frequency candidate f, count the number of f_i which satisfies the following condition:

$$floor[(1+d)f_i/f] \geq (1-d)f_i/f \quad (4)$$

$floor(x)$ denotes the greatest integer less than or equal to x. This condition means whether $r_i f \cdot (1-d) \leq f_i \leq r_i f \cdot (1+d)$. If the condition is fulfilled, f_i is the frequency of the r_i^{th} harmonic component when fundamental frequency is f. For each fundamental frequency candidate f, the coincidence number is calculated and $\hat{f}$ corresponding to the largest coincidence number is selected as the estimated fundamental frequency.

The original algorithm is extended in the following way: First, in peak detection procedure, not all peaks exceeding the given threshold are detected, only significant ones are selected by a edge detecting procedure. This is very important for eliminating noise and achieving high performance in next steps. Second, not only fundamental frequency but also all its harmonics are extracted, then $\mathbf{B}$ can be calculated. Third, the original optimality criteria is to select $\hat{f}$ corresponding to the largest coincidence number. This criteria is not stable when the signal is polyphonic, because harmonic components of different sources may influence each other. A new optimality criteria is define as follows:

$$d = \frac{1}{n}\sum_{i=1, f_i\ coincident\ with\ f}^{K}\frac{|r_i - f_i/f|}{r_i} \quad (5)$$

Then, $\hat{f}$ corresponding to the smallest d is selected as the estimated fundamental frequency. The new criteria measures the precision of coincidence. Generally speaking, for a fundamental frequency candidate f, harmonic components of the same source are more probably to have a high coincidence precision than those of a different source. So the new

criteria is helpful for separation of harmonic structures of different sources. Fourth, in the original algorithm, only one pitch was detected in each frame. In our situation, the sound is a mixture and is polyphonic. So, all pitches which corresponding d below a given threshold are extracted.

After harmonic structure extraction, a data set of harmonic structures is obtained. As the analysis in section two, music harmonic structures in different frames are similar to each other, while voice harmonic structures have a large variation. So, in the data set all music harmonic structures form a cluster with very high density while voice harmonic structures scatter around like background noise.

In the third step, harmonic structures outside three standard deviations are removed from the data set. This operation removes most of the voice harmonic structures. Then the average harmonic structure of music can be calculated.

In separation step, in each frame of the mixed signal, if there exists a harmonic structure similar to the music AHS, a music harmonic structure is detected. Otherwise, there is no music structure. Music harmonic structures in all frames are extracted to reconstruct the music signal and then removed from the mixture. The rest of the mixture after removing music harmonic structures is the separated voice signal.

The procedure of music harmonic structure detection is detailed as follows. Given the music AHS $[\bar{B}_1, \ldots, \bar{B}_R]$ and a fundamental frequency candidate f, a possible music harmonic structure is predicted. $[f, 2f, \ldots, Rf]$ and $[\bar{B}_1, \ldots, \bar{B}_R]$ are the frequencies and harmonic structure coefficients of it. Find the closest peak in the magnitude spectrum for each predicted harmonic component. Suppose $[f_1, \ldots, f_R]$ and $[B_1, \ldots, B_R]$ are the frequencies and harmonic structure coefficients of these peaks (measured peaks). Using formula 6 to calculate the distance between the predicted harmonic structure and the measured peaks.

The first part of formula 6 is a modified version of Two-Way Mismatch measure defined by Maher and Beauchamp, which measures the frequency difference between predicted peaks and measured peaks [11]. Where $\Delta f_r = |f_r - r \cdot f|$. The second part of formula 6 measures the shape difference between the two, a is a normalize coefficient. Note that, only harmonic components with none zero harmonic structure coefficients are considered. Let $\hat{f}$ indicate the fundamental frequency candidate corresponding to the smallest distance between the predicted peaks and the actual spectral peaks. If $D(\hat{f})$ is smaller than a threshold T_d, a music harmonic structure is detected. Otherwise there is no music harmonic structure in the frame. If a music harmonic structure is detected, the corresponding measured peaks in the spectrum are extracted. And music signal is reconstructed by IFFT. The voice signal is reconstructed by taking IFFT on the rest spectrum. Smoothing between frames is needed to eliminate click noise between frames.

$$D(f) = \sum_{r=1,\bar{B}_r>0,B_r>0}^{R} \{\Delta f_r \cdot (rf)^{-p} + \frac{\bar{B}_r}{\bar{B}_{\max}} \times q\Delta f_r \cdot (rf)^{-p}\} + a \sum_{r=1,\bar{B}_r>0,B_r>0}^{R} (\frac{\bar{B}_r}{\bar{B}_{\max}})(\bar{B}_r - B_r)^2 \quad (6)$$

4. EXPERIMENTAL RESULTS

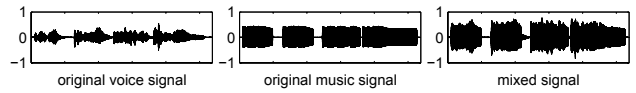
We had calculated HSSs on both the Iowa music instrument database and a singing voice database. The music database contains 21 kinds of instruments and total 765 samples. The sample rate is set to 44.1K to contain more high frequency harmonics. The singing voice database contains 50 segments coming from 10 acappellas including male and female singing voice. Each segment is 30 seconds in length. The mean of the HSSs of harmonic music samples is 0.04. The mean of the HSSs of singing samples is 0.18. So, HSS represents the essential difference between voice and music.

Fig. 2 are two examples of separation experiments. Table 1 are corresponding SNRs. All the files can be downloaded from http://www.au.tsinghua.edu.cn/szll/bodao/zhangchangshui/bigeye/member/zyghtm/music.htm. It can be seen that the mixtures are well separated and very high SNRs are obtained. The distance between music harmonic structures and the corresponding music AHS is small (the mean distances is 0.01 and 0.006 in experiment 1 and 2, respectively), and the distance between voice harmonic structures and the music AHS is bigger (the mean distances is 0.1 and 0.13 in experiment 1 and 2, respectively). So, the music AHS is a good model for music signal representation and for voice and music separation. The separation procedure is based on harmonic structure analysis, it makes the separated signals with a rather good subjective audio quality because the harmonic structure is reserved. This is an important advantage of the proposed algorithm. All the parameters are determined experimentally. $d = 0.04, p = 1, q = 4, a = 0.01$ in all experiments, which are parameters in equation 4, 6. T_d, used in separation step, is 0.15 and 0.35 in experiment 1 and 2 , respectively.

Fig. 2 also shows speech enhancement results obtained by a speech enhancement software which tries to estimate the spectrum of noise in the pause of speech and enhance the speech by spectral subtraction [3]. Detecting pauses in speech with music background and enhancing speech with music noise are both very difficult problems, so tradition speech enhancement techniques can't work here.

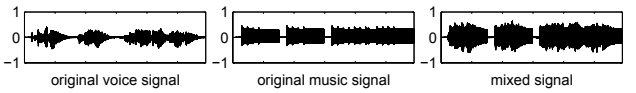
5. CONCLUSION AND DISCUSSION

In this paper, the essential difference between voice and music signals is analyzed. Harmonic structure stability and av-



(a) Experiment1:The original voice and music signal,and the mixed signal

(b) Experiment1:The separated signals and the speech enhancement result



(c) Experiment2:The original voice and music signal,and the mixed signal

(d) Experiment2:The separated signals and the speech enhancement result

Fig. 2. Experimental results.

Table 1. SNR results (DB): snr_v and snr_m are the SNRs of voice and music signals in the mixed signal. snr'_e is the SNR of speech enhancement result. snr'_v and snr'_m are the SNRs of the separated voice and music signals.

	snr_v	snr_m	snr'_e	snr'_v	snr'_m	Total inc.
Experiment1	-7.9	7.9	-1.1	6.7	10.8	17.5
Experiment2	-5.2	5.2	-1.5	6.6	10.0	16.6

erage harmonic structure are defined to represent this difference and separate signals. Experimental results show a good performance of this method. The difference between voice and music signals is significant. So the algorithm based on analysis and modeling this difference is a promising way to solve many problems in music and speech research.

The proposed method has applications in many research areas. It is useful for melody extraction and then makes audio retrieval become much easier. In our algorithm, not only harmonic structures but also corresponding fundamental frequencies are extracted. So, the algorithm is also a new multi-pitch estimation method. It analyzes the primary multi-pitch estimation results and learns a model to represent music signals and improve multi-pitch estimation results. More importantly, pitches of different sources can be distinguished by the AHS model. This advantage is significant for automatic transcription.

There are still some limitations. First, for non-harmonic instruments, such as some drums, the proposed algorithm doesn't work. Some rhythm tracking algorithms can be used instead to separate drum sounds. Fortunately, most instrument sounds are harmonic. Second, when there exists more than one music in the mixture, the algorithm should be extended to learned the AHS for every instrument. Then the separation can be done in the similar way. Third, for some instruments, the timbre in the onset duration is somewhat different from which in the stable duration. Also different performing methods (pizz. or arco) produces different timbres. In these cases, the music harmonic structures will form several clusters not one. Then a GMM model instead of an average harmonic structure model (actually a point model) should be used as to represent the music.

6. REFERENCES

[1] J. S. Downie, "Music information retrieval," *Annual Review of Information Science and Technology*, vol. 37, pp. 295–340, 2003.

[2] Anssi Klapuri, "Automatic transcription of music," M.S. thesis, Tampere University of Technology, Finland, 1998.

[3] Serguei Koval, Mikhail Stolbov, and Mikhail Khitrov, "Broadband noise cancellation systems: new approach to working performance optimization," in *EUROSPEECH'99*, 1999, pp. 2607–2610.

[4] Gil-Jin Jang and Te-Won Lee, "A probabilistic approach to single channel blind signal separation," in *NIPS*, 2003.

[5] Yazhong Feng, Yueting Zhuang, and Yunhe Pan, "Popular music retrieval by independent component analysis," in *ISMIR*, 2002, pp. 281–282.

[6] Peter Vanroose, "Blind source separation of speech and background music for improved speech recognition," in *The 24th Symposium on Information Theory*, May 2003, pp. 103–108.

[7] Tong Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 4, pp. 441–457, May 2001.

[8] J. Pinquier, J. Rouas, and R. Andre-Obrecht, "Robust speech / music classification in audio documents," in *ICSLP*, 2002, pp. 2005–2008.

[9] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Popea, A. Picialli, and G. De Poli, Eds. Swets & Zeitlinger Publishers, 1997.

[10] E. Terhardt, "Calculating virtual pitch," *Hearing Res.*, vol. 1, pp. 155–182, 1979.

[11] R. C. Maher and J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994.