Singing Voice Separation from Monaural Recordings

Yipeng Li
Department of Computer Science and Engineering
The Ohio State University
liyip@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering
and the Center for Cognitive Science
The Ohio State University
dwang@cse.ohio-state.edu

Abstract

Separating singing voice from music accompaniment has wide applications in areas such as automatic lyrics recognition and alignment, singer identification, and music information retrieval. Compared to the extensive studies of speech separation, singing voice separation has been little explored. We propose a system to separate singing voice from music accompaniment from monaural recordings. The system has three stages. The singing voice detection stage partitions and classifies an input into vocal and non-vocal portions. Then the predominant pitch detection stage detects the pitch contour of the singing voice for vocal portions. Finally the separation stage uses the detected pitch contour to group the time-frequency segments of the singing voice. Quantitative results show that the system performs well in singing voice separation.

Keywords: Singing voice detection, predominant pitch detection, singing voice separation

1. Introduction

A successful singing voice separation system is useful in many areas such as automatic lyrics recognition and alignment, singer identification, and music information retrieval. In this paper, we focus on singing voice separation from monaural recordings. A monaural solution is indispensable in many cases, such as separating the singing for live recordings (non-studio recordings). The development of a successful monaural singing voice separation system could also enhance our understanding of how the human auditory system performs such tasks.

Although singing voice is produced by the speech organ, speech separation systems might not be directly applicable to singing voice separation. This is mainly because of the nature of other concurrent sounds. In a real acoustic environment, interfering sounds in most cases are uncorrelated with speech. For recorded songs, however, music accompaniment is correlated with singing voice since they are composed to be a coherent whole. This difference makes the separation of singing voice from music accompaniment potentially more challenging.

The perceptual work by Bregman [1] and others has inspired researchers to study *computational auditory scene analysis* (CASA). Compared to other sound separation approaches, such as spectral subtraction, CASA makes fewer assumptions about concurrent sounds therefore shows greater promise in singing voice separation from monaural recordings. Mellinger [2] proposed a CASA system which extracts onset and common frequency variation and uses them to group frequency partials from the same musical instrument. Godsmark and Brown [3] developed a CASA system which uses harmonicity and other auditory scene analysis principles in a blackboard architecture for music sound separation. Recently a speech separation system developed by Hu and Wang [4] exploits pitch and amplitude modulation to separate voiced speech from various kind of interference. Systematic evaluation shows that the system performs significantly better than previous systems.

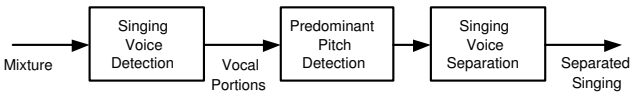

Figure 1. Schematic diagram of the proposed system

Since the Hu–Wang system allows the interference to be harmonic, it is possible to apply the system to singing voice separation. The accuracy of pitch detection is critical for the Hu–Wang system. However, as shown in [5] their pitch estimation is unreliable when singing voice is accompanied by music. This problem can be alleviated by a predominant pitch detection algorithm we proposed in [5], which detects more accurately the pitches of singing voice for different musical genres. Because their system works for voiced speech, it is necessary to have a mechanism to distinguish portions where singing voice is present from those where it is not. On the other hand, although their system is limited to voiced speech separation, this limitation is less severe for singing voice separation because unvoiced singing comprises a smaller percentage in terms of time and its contribution to the intelligibility of singing is less than that to the intelligibility of speech.

In this paper we propose a singing voice separation system which consists of three stages, as shown in Fig. 1. The first stage performs singing voice detection in which the input is partitioned and classified into vocal and non-vocal portions. Then vocal portions are used as input to the second stage for predominant pitch detection. In the last stage, detected pitch contours are used for singing voice separation where we extend the Hu–Wang system [4]. The output of the system is separated singing voice.

The remainder of this paper is organized as follows. Section 2 describes each stage of the system. Section 3 presents the evaluation and the last section concludes the paper.

2. System description

2.1. Singing voice detection

The goal of this stage is to partition the input into vocal portions in which singing voice is present and non-vocal portions in which singing voice is absent. Our strategy is

based on the observation that, when a new sound enters a mixture, it usually introduces significant spectral changes. Therefore the possible instance when a sound enters can be determined by identifying significant spectral changes. This idea is more suitable for singing voice detection since, in order to conform with the rhythmic structure of a song, a voice is more likely to join the accompaniment at beat times when strong spectral perturbation occurs. Therefore in this stage we first use a spectral change detector to partition the input into spectrally homogeneous portions and then pool the information within a portion for classification.

The spectral change detector used in this stage is proposed by Duxbury et al. [6]. It calculates the Euclidian distance in the complex domain between the expected spectral value and the observed one in each frame. Significant spectral changes are indicated as local peaks in the distance values. After the input is partitioned into portions, each portion is classified into vocal or non-vocal according to the overall likelihood. Formally let $\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_M\}$ be a set of feature vectors of a portion with M frames. Let $\log p(\boldsymbol{X}|c_v)$ and $\log p(\boldsymbol{X}|c_{nv})$ represent the log likelihood of an observed feature vector $\boldsymbol{X}$ being in the vocal class c_v and the non-vocal class c_{nv}, respectively. Then a portion is classified as vocal if:

$$\sum_{j=1}^{M} \log p(\boldsymbol{X}_j|c_v) > \sum_{j=1}^{M} \log p(\boldsymbol{X}_j|c_{nv}) \qquad (1)$$

Since mel-frequency spectral coefficients (MFCC) and Gaussian mixture models (GMM) are widely used in audio classification tasks, we choose MFCCs as the features and GMMs as the classifiers for likelihood evaluation.

2.2. Predominant pitch detection

In this stage, a predominant pitch detection algorithm proposed in [5] is used to detect the pitch contour of singing voice for vocal portions. The algorithm first decomposes a vocal portion into its frequency components with a 128-channel gammatone filterbank. A normalized correlogram is then computed for each channel and each frame to obtain periodicity information. The peaks in the normalized correlogram contain the periodicity information of the input. However, due to the presence of music accompaniment, some peaks may give misleading information. To alleviate the problem, channel and peak selection are applied to all channels to extract reliable periodicity information. The algorithm uses Hidden Markov Model (HMM) to describe the pitch generation process. In each frame the observation probability of a pitch hypothesis is calculated by integrating the periodicity information across all frequency channels. The transition probability between two consecutive frames is determined by training. In order to reduce the interference of other harmonic sounds from accompaniment, the HMM tracks up to 2 predominant pitch contours simultaneously. Finally the Viterbi algorithm is used to find the most likely sequence of pitch hypotheses and the first pitch contour of this optimal sequence is considered as the pitch contour of the singing voice. More details of the algorithm can be found in [5, 7].

2.3. Singing voice separation

In this stage, the voiced speech separation algorithm developed by Hu and Wang [4] is extended for singing voice separation for vocal portions. The singing voice separation algorithm first passes the input, i.e., a vocal portion, through an auditory periphery which is a 128-channel gammatone filterbank. The output of each channel is further divided into 16–ms time frames with 50% overlap. In this way, the input is decomposed into a time-frequency (T-F) map, each element of which is called a T-F unit. For each T-F unit, the following features are extracted: energy, autocorrelation, cross-channel correlation, and cross-channel envelope correlation.

Next, the algorithm forms segments by merging contiguous T-F units based on temporal continuity and cross-channel correlation. Only those T-F units whose energy and cross-channel correlation both high are considered. Neighboring units, either in time or frequency, are merged into segments iteratively.

By comparing the local periodicity information indicated in the autocorrelation of a T-F unit to the estimated periodicity of the singing voice in the same frame, the T-F unit is labeled as either singing voice dominant or accompaniment dominant. We use the pitch contour of the singing voice obtained in the second stage to label each T-F unit. More specifically, a T-F unit in frequency channel c and time frame m is labeled as singing dominant if:

$$\frac{A(c, m, \tau_S(m))}{A(c, m, \tau_P(c, m))} > \theta_T \qquad (2)$$

where $A(c, m, \tau)$ is the autocorrelation of the unit with the time lag indicated by τ, and θ_T is a threshold. $\tau_S(m)$ is the time lag corresponding to the estimated pitch period in frame m while $\tau_P(c, m)$ is the time lag corresponding to the global maximum of $A(c, m, \tau)$ within the plausible pitch range from 80 to 500 Hz. This periodicity criterion works well for T-F units where harmonics are resolved — a harmonic is resolved if it activates a dedicated auditory filter. For filters responding to multiple harmonics, their responses are amplitude-modulated. As a result, the time lag of the global maximum of $A(c, m, \tau)$ of those filters within the pitch range might not correspond to the pitch period.

To deal with the problem of unresolved harmonics, the algorithm extracts the amplitude modulation (AM) rate for each unit and compares the AM rate with the estimated pitch period. More specifically, a normalized envelope of a T-F unit is first extracted. Then a single sinusoid with the same period as the estimated pitch period is constructed. To compare the sinusoid with the normalized envelope, the phase of the sinusoid is adjusted such that the square error between these two signals is minimized. After the phase is determined, the T-F unit is labeled as singing dominant if the envelope can be well described by the obtained sinusoid. This AM criterion is formally defined as:

$$\frac{\sum_{n=0}^{N-1} [\hat{r}(c, n) - cos(\frac{2\pi n}{\tau_S(m) f_S} + \phi_{cm})]^2}{\sum_{n=0}^{N-1} \hat{r}^2(c, n)} < \theta_{AM} \qquad (3)$$

where $\hat{r}(c, n)$ is the normalized envelope. ϕ_{cm} represents the phase minimizing the square error and f_S is the sampling frequency of the input. n is the time index and N is the length of the envelope. θ_{AM} is a threshold. For units labeled as singing dominant by the AM criterion, additional segments are generated based on temporal continuity and cross-channel envelope correlation.

Table 1. Classification accuracy for different methods (% frames)

	-5 dB	0 dB	5 dB	10 dB
proposed method	**80.3**	**85.0**	**90.2**	**91.1**
frame-level classification	71.3	77.4	81.7	83.8
HMM	79.0	83.5	87.5	88.8

Table 2. Predominant pitch detection error rates with actual classification (%)

	-5 dB	0 dB	5 dB	10 dB
Proposed	**44.2**	**31.7**	**24.3**	**21.6**
Klapuri	55.5	41.7	31.7	26.5
Wu et al.	55.1	39.0	29.0	22.4

In the final step of the separation algorithm, segments where a majority of T-F units is labeled as singing dominant are grouped to form the foreground stream, which corresponds to the singing voice. From the segments in the foreground stream the singing voice can be obtained by resynthesis. For more details of the Hu–Wang system, see [4].

3. Evaluation

We extracted 10 songs sampled at 16 kHz from karaoke CDs for singing voice detection. These CDs are recorded with multiplex technology. With proper de-multiplexing software, clean singing voice and accompaniment can be extracted. We further extracted 25 clips from the 10 songs for singing voice pitch detection and separation. These clips include rock and country music. We refer to the energy ratio of singing voice to accompaniment as signal to noise ratio (SNR) as in speech separation studies.

Fig. 2 shows the output of each stage of the proposed system for a clip of rock music. Fig. 2(a) is the waveform of the clean singing voice. The thick lines above the waveform indicate reference vocal portions obtained by applying an energy-based silence detector on the clean singing voice. The mixture in which the singing voice and the accompaniment are mixed in 0 dB is shown in Fig. 2(b). Fig. 2(c) gives the result of singing voice detection on the mixture. A high value indicates the frame is classified as vocal and a low value as non-vocal. Fig. 2(d) gives the result of predominant pitch detection on the detected vocal portions. The detected pitches are plotted as dots against the reference pitch contours which are plotted as solid lines. The output of the singing voice separation stage is plotted in Fig. 2(e). As can be seen, the separated singing voice matches the clean singing voice well.

For singing voice detection, we trained the classifiers with samples mixed in 0 and 10 dB and tested them in 4 different SNRs using 10-fold cross validation. The average classification accuracies (percentage of frames) are shown in the first row of Table 1. For comparison purposes, the results of frame-level classification (each frame is a portion) and HMM similar to the one used in [8] are also shown in Table 1. As can be seen, the proposed singing voice detection method performs better for all SNRs.

For predominant pitch detection, we calculate the error rates (at the frame level) of pitch detection for the first two stages. Note that in this case the singing voice detection

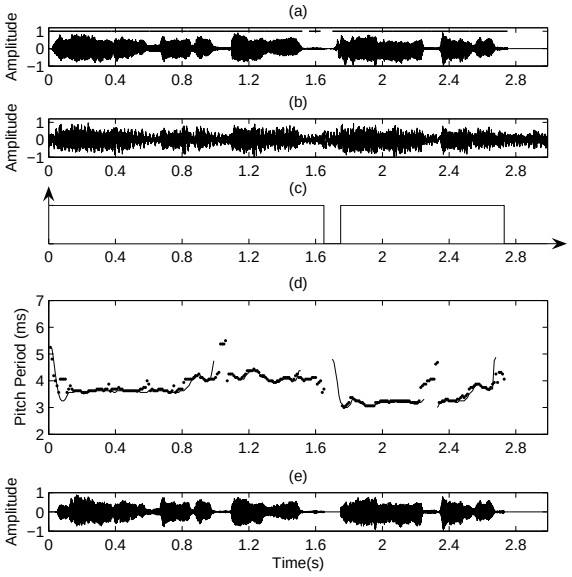

Figure 2. Output of each stage for a clip of rock music. (a) The singing voice. (b) The mixture. (c) The output of the singing voice detection stage. Vocal portions are indicated by high values and non-vocal portions by low values. (d) The output of the predominant pitch detection stage. Dots indicate the detected pitches and the solid lines indicate the reference pitches. (e) The output of the singing voice separation stage.

stage also contributes to the pitch detection errors. The reference pitches are calculated using Praat [9]. An error occurs if a detected pitch is not within 10% of the reference pitch. The error rates of the proposed method as well as those of two other methods are shown in Table 2. Klapuri's algorithm [10] performs multipitch detection. We implemented the algorithm and chose the first detected pitch as the predominant one. The obtained pitch sequence was smoothed to improve the pitch detection accuracy. The performance of the original algorithm developed by Wu et al. [7] is also listed in Table 2. As can be seen, for all SNRs, our method has lower pitch detection error rates.

An important aspect of evaluating sound separation systems is the criterion, which is directly related to the computational goal of a system. For musical applications, the perceptual quality of the separated sound is emphasized in some cases. However, perceptual quality is subjective and hard to quantify. Here we adopt the notion of ideal binary mask proposed in [4]: a T-F unit in the mask is assigned 1 if the energy of the target source in the unit is stronger than that of other concurrent sounds, and 0 otherwise. This notion is grounded on the well-established auditory masking phenomenon [11]. For more discussion of the ideal binary mask, the interested reader is referred to [12]. With clean singing voice and accompaniment available, the ideal binary mask can be readily constructed. Our informal listening experiments show that the quality of singing voice resynthesized from the ideal binary mask is close to that of the original one when SNR is high and it degrades gradually with decreasing SNR. Therefore we suggest to use the ideal binary mask as the computational goal for singing voice sep-

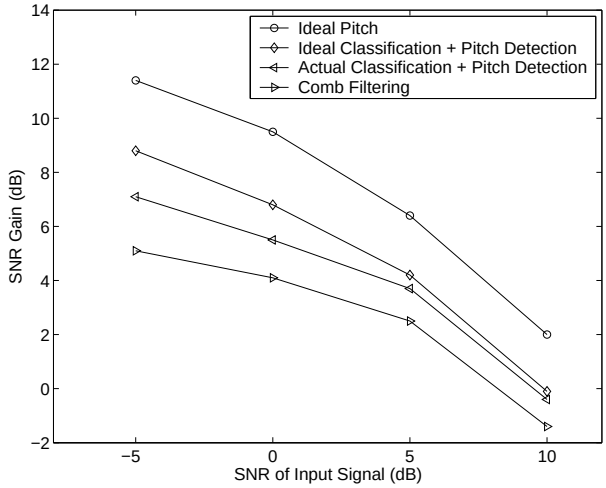

Figure 3. SNR gain comparison.

aration.

The performance of the system can be quantified by calculating the SNR before and after the separation using the singing voice resynthesized from the ideal binary mask as the ground truth [4]:

$$SNR = 10 \log 10[\frac{\sum_n I^2(n)}{\sum_n (I(n) - O(n))^2}] \qquad (4)$$

where $I(n)$ is the ground truth. In calculating the SNR after separation, $O(n)$ is the output of the system. In calculating the SNR before separation, $O(n)$ is the mixture resynthesized from an all-one mask, which compensates for the distortion introduced in the resynthesis.

We evaluate the performance of the proposed system for 4 different SNRs: -5, 0, 5, and 10 dB. Fig. 3 shows the SNR gains after separation for different cases. The SNR gains using the ideal pitch as input to the separation stage are shown as the line on the top. This gives the ceiling performance of our pitch-based separation system. The second line from the top gives the SNR gains using reference vocal portions (ideal classification) and pitch detection. As the predominant pitch detection stage introduces errors to the system, the gains are lower than that using ideal pitch. The third line from the top shows the SNR gains of the system, i.e., using actual classification and pitch detection. As the singing voice detection stage also makes errors, the performance is further decreased. Although the SNR after separation of the proposed system for the 10 dB case is not improved, the system achieves SNR improvements of 7.1, 5.5, and 3.7 dB for the input SNR of -5, 0, and 5 dB, respectively. This demonstrates that the proposed system works better for low SNR situations. We also compare the proposed separation system with a standard comb filtering method [13], which extracts the spectral components at the multiples of a given pitch. As shown in the bottom line in Fig. 3, the performance of the comb filtering method is consistently worse than that of the proposed system. Since the classification stage rejects energy from the accompaniment, this stage alone is expected to contribute to the SNR gains. Quantitatively the SNR gains from the classification stage alone are 1.4, 1.0, 1.1, and 0.2 dB for -5, 0, 5, and 10 dB cases, respectively. Therefore the SNR gains are mainly contributed from the pitch-based separation. Demos of singing voice separation can be found at http://www.cse.ohio-state.edu/~liyip/Research/Publication/2006/singing_demo.htm.

4. Conclusion

In this paper, we have proposed a monaural system to separate singing voice from music accompaniment. Our system first detects vocal portions and then applies predominant pitch detection to each vocal portion to obtain the pitch contour of singing voice. Finally the system uses detected pitch contours to separate the singing voice from music accompaniment by extending a voiced speech separation system. Quantitative evaluation of the system shows that it performs well for singing voice separation, especially in low SNR conditions.

5. Acknowledgments

This research was supported in part by an AFOSR grant (F49620-04-1-0027) and an NSF grant (IIS-0534707). We thank G. Hu for many useful discussions.

References

[1] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, Cambridge, MA, 1990.

[2] David K. Mellinger, *Event Formation and Separation in Musical Sound*, Ph.D. thesis, Stanford University, Department of Computer Science, 1991.

[3] D. Godsmark and G. J. Brown, "A blackborad architecture for computational auditory scene analysis," *Speech Communication*, vol. 27, no. 4, pp. 351–366, 1999.

[4] Guoning Hu and DeLiang Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[5] Yipeng Li and DeLiang Wang, "Detecting pitch of singing voice in polyphonic audio," in *Proc. IEEE ICASSP*, 2005, vol. 3, pp. 17–20.

[6] Chris Duxbury, Juan Pablo Bello, Mike Davies, and Mark Sandler, "Complex domain onset detection for musical signals," in *Proc. of the 6th Conference on Digital Audio Effect (DAFx-03)*, London, U.K., 2003.

[7] Mingyang Wu, DeLiang Wang, and Guy J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech Audio Processing*, vol. 11, pp. 229–241, 2003.

[8] Adam L. Berenzweig and Daniel P. W. Ellis, "Locating singing voice segments within music signals," in *Proc. IEEE WASPAA*, 2001, pp. 119–122.

[9] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer, version 4.0.26," (http://www.fon.hum.uva.nl/praat), 2002.

[10] A.P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Transactions on Speech Audio Processing*, vol. 11, pp. 204–816, 2003.

[11] Brian C. J. Moore, *An Introduction to the Psychology of Hearing*, fifth edition, Academic Press, London, U.K., 2003.

[12] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 181–197. Kluwer Academic, Boston, MA, 2005.

[13] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, New York, 1993.