Spectro-temporal Modulation Based Singing Detection Combined with Pitch-based Grouping for Singing Voice Separation

Tse-En Lin[1], Chung-Chien Hsu[1], Yi-Cheng Chen[2], Jian-Hueng Chen[2] and Tai-Shih Chi[1]

[1]Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan.
[2]Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan.

cosandy.cm96g@nctu.edu.tw, hsu.chung.chien@gmail.com, cliff1196@cht.com.tw, cchdavis@cht.com.tw, tschi@mail.nctu.edu.tw

Abstract

A spectro-temporal modulation based singing voice detection cascaded with a Viterbi based pitch tracking algorithm is proposed in this paper for singing-voice separation from monaural recordings. To detect the singing voice, the spectro-temporal modulation energy related to voice harmonics is extracted using a spectro-temporal modulation analysis framework developed for the Fourier spectrogram. Separation of singing-voice from background music is conducted using a binary mask to group estimated harmonics of singing voice. The proposed system is evaluated using MIR-1K dataset and is shown outperforming three other binary-mask based systems in the vocal/music separation task.

Index Terms: singing voice detection, singing voice separation, spectro-temporal modulation, pitch tracking.

1. Introduction

Singing voice is indispensable for some practical music-related applications, such as lyric recognition [1], lyric synchronization [2] and singer identification [3]. Unfortunately, singing voice is often mixed with music in songs. Therefore, an effective vocal/music separation algorithm is crucial to relevant applications.

Previous research on vocal/music separation can be approximately divided into two categories. The approach of the first category involves a singing voice detection mechanism and a source separation algorithm. For instance, the singing voice was reconstructed from a mixture using adapted voice and music models in [4]. In [5], blind source separation algorithms, such as the independent component analysis (ICA) and the non-negative matrix factorization (NMF), were utilized in vocal-only regions. Inspired by auditory scene analysis (ASA) [6], a pitch extraction algorithm was used to extract harmonic components of singing voice [7][8]. In contrast, the approach of the second category does not require an independent voice detection mechanism. The characteristics of voice and music on spectrograms are exploited for separation [9][10].

In this paper, a novel singing voice detection mechanism is proposed and tested with a straightforward Viterbi-based pitch tracking algorithm for vocal/music separation. Pure spectral features, such as linear prediction coefficients (LPC) [11], and pure temporal features, such as the temporal modulation energy [12], have been shown effective in detecting singing voice. Unlike these conventional spectral or temporal features, joint spectro-temporal modulation features are used in our system. The original spectro-temporal modulation features extracted from an auditory model [13] have been successfully used in speech/nonspeech discrimination with a support vector machine (SVM) recognizer [14]. Thereafter, we have derived a similar spectro-temporal analysis framework for Fourier spectrograms [15] and successfully applied it to enhance speech [16]. Here, we use this framework to extract the spectro-temporal modulation energy from a Fourier spectrogram for detecting singing voice segments. The rational of our approach is that singing voice and background music possess different spectro-temporal modulations including frequency modulations (FM) and amplitude modulations (AM). Since a simple modulation-energy based feature is used, no complex recognizers, such as the SVM in [14], are needed in our system. The singing voice detection mechanism is tested in vocal/music separation simulations by cascading a Viterbi based pitch tracking algorithm, which picks the track with the minimal cost from inter-frame transitions of pitch candidates within detected vocal segments. Once the pitch track is obtained, a binary mask, which assigns harmonics to be one, is generated to separate singing-voice from background music.

The rest of the paper is organized as follows. Section 2 describes our system including singing voice detection, pitch estimation and vocal/music separation modules. Section 3 demonstrates experimental results of the proposed system and performance comparisons with other binary-mask based systems. Finally, conclusions and future work are discussed in Section 4.

2. Method

Three major steps were adopted in the previous works [7][8]. First, the vocal and music regions were detected based on extracted features. In [8], each vocal region was further divided into voiced and unvoiced frames. Second, a pitch extraction and tracking algorithm was used to estimate the pitch contour within singing voice segments. Finally, a time-frequency (T-F) mask was generated based on the estimated pitch to separate singing voice from the mixture. Our proposed overall system follows this ASA framework and has these three modules as well.

To have a fair comparison with [8], we adopt the same setting as in [7][8], i.e., analyzing the input sound mixture using 128 gammatone constant-Q filters whose center frequencies are quasi-logarithmically distributed between 80 and 5000 Hz [7]. Outputs of these gammatone filters are further divided into 40-ms frames with a 20-ms frame shift. This analysis stage decomposes a sound into T-F units, each of which refers to the output of a certain gammatone filter in a certain time frame. The vocal/music separation is then carried out on this T-F representation using a binary mask.

2.1. Singing voice detection

Previous studies adopted different kinds of spectral and temporal features for singing voice detection [11][12]. In contrast, joint spectro-temporal modulation features, which

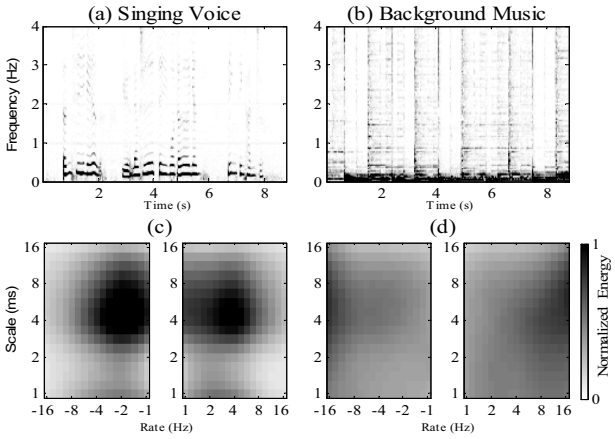

Figure 1: *Magnitude spectrograms and averaged spectro-temporal modulation energy profiles of singing voice and background music in the rate-scale domain.*

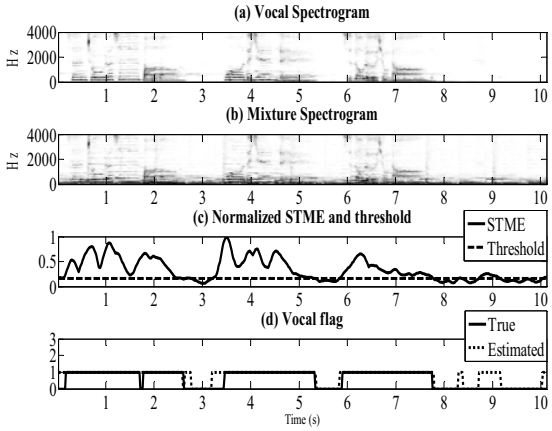

Figure 2: An example of the vocal/music discrimination at 0 dB SNR.

have been shown capable of encoding vibrato (frequency modulation, FM) and tremolo (amplitude modulation, AM) [15], are used in our system. The FM and AM are believed being two vital features for singing voice perception [17]. Extraction of modulation features and the vocal detection procedure are described below.

2.1.1. Spectro-temporal modulation feature

The spectro-temporal analysis framework in [15] is adopted for feature extraction. First, a Fourier spectrogram is calculated using the 1024-point short time Fourier transform (STFT) with the setting of a 40-ms frame length and a 20-ms frame shift. Then, a 2D spectro-temporal analysis on the magnitude spectrogram is carried out using a bank of 2D zero-phase spectro-temporal modulation filters (STMF) to decompose the magnitude spectrogram. This 2D analysis was inspired by spectro-temporal receptive fields recorded in the primary auditory cortex [13]. The frequency responses of the downward (with subscript "+") and the upward (with subscript "-") modulation filters can be written as:

$$STMF_{+}(\omega,\Omega) = \begin{cases} \left|\mathcal{F}\{h_{rate}(t)\} \otimes \mathcal{F}\{h_{scale}(f)\}\right|, & 0 \le \omega;\Omega \le \pi \\ 0 & , \text{ otherwise} \end{cases} \quad (1)$$

$$STMF_{-}(\omega,\Omega) = \begin{cases} \left|\mathcal{F}\{h_{rate}(t)\} \otimes \mathcal{F}\{h_{scale}(f)\}\right|, & -\pi \le \omega \le 0;0 \le \Omega \le \pi \\ 0 & , \text{ otherwise} \end{cases} \quad (2)$$

where $\mathcal{F}$ is the 1-D Fourier transform; $\otimes$ is the outer product and π indicates the half sampling frequencies of the discrete signal processing along the time and the frequency axes. The rate (ω in Hz) and the scale (Ω in ms) represent the Fourier domains of the time and the frequency axes, respectively. From above equations, the frequency responses of the downward and the upward modulation filters only have components in the first and second quadrant of the ω-Ω space, respectively. The h_{rate} and h_{scale} are derived from one-dimensional constant-Q gammatone filter with Q_{3dB} = 2. Detailed implementations can be accessed in [16].

Then the spectro-temporal modulation energy (STME) contour of the modulation filter tuned to (ω=4Hz, Ω=5ms) can be calculated as follows.

$$\begin{cases} RSE_{+}(t,\omega,\Omega) = \sum_{\forall f} \left|R_{+}(t,f,\omega,\Omega)\right| \\ RSE_{-}(t,\omega,\Omega) = \sum_{\forall f} \left|R_{-}(t,f,\omega,\Omega)\right| \\ STME(t) = \max(RSE_{+}(t,\omega=4,\Omega=5), RSE_{-}(t,\omega=4,\Omega=5)) \end{cases} \quad (3)$$

where $R_{+}(t,f,\omega,\Omega)$ and $R_{-}(t,f,\omega,\Omega)$ are the outputs of the downward and upward 2D modulation filter tuned to rate ω and scale Ω, respectively; and the rate-scale energy (RSE) profile are calculated by integrating the 4D output R(t, f, ω, Ω) along the frequency axis.

Sample spectrograms of a singing voice and background music are shown in Fig. 1(a) and (b). Fig. 1(c) and (d) show their corresponding averaged RSE profile (with all assessed rate ω and scale Ω) over all frames. Clearly, the modulation energy for singing voice is concentrated around ω=2~4 Hz in rate and Ω=5 ms in scale. On the other hand, the averaged RSE profile of music is all smeared due to the fact that many instruments with different modulations are present in the background music at any time instance. Based on Fig. 1, the output energy of the modulation filter tuned to (ω=4Hz, Ω=5ms) is selected as a vital feature for singing voice detection in our system.

2.1.2. Threshold setting

The extracted modulation energy is compared with a threshold for vocal/music discrimination. The assumption of a certain period of music-only segment being at the beginning of each song is not valid such that we cannot estimate an initial threshold from the beginning of the song and update the threshold frame by frame continuously. Therefore, we set an absolute threshold for each song as follows.

$$Th = P \cdot (\max(STME(t)) - \min(STME(t))) + \min(STME(t)) \quad (4)$$

where P is a scaling parameter. An example of the vocal/music discrimination at 0 dB SNR can be seen in Fig. 2. Fig. 2(c) shows the STME(t) from equation (3) and the absolute threshold. Fig. 2(d) shows the estimated and true vocal labels.

2.2. Pitch tracking

Pitch candidates in each frame are first determined using the average magnitude difference function (AMDF). The AMDF of each frame is formulated as follows.

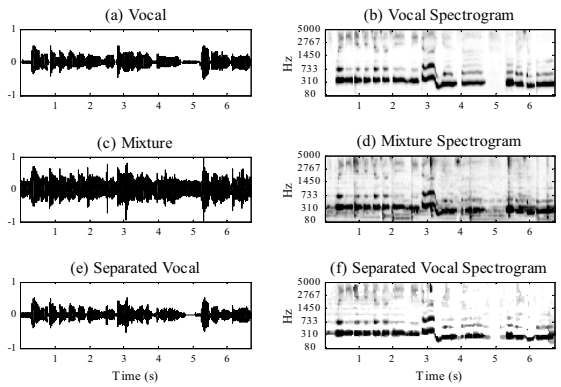

Figure 3: *Output example of the proposed vocal/music separation system at 0 dB SNR.*

$$AMDF(t,\tau)=\sum_{\forall f}\sum_{i=0}^{W-1}\left|Y(tS+i,f)-Y(tS+\tau+i,f)\right| \quad (5)$$

where Y is the output waveform of a gammatone filter; S is the number of sample in 20 ms (frame shift); W is the number of sample in 40 ms (frame length); t and f are the frame and frequency-bin indices; and τ is the lag parameter ranging from 1.67 to 12.5 ms corresponding to the possible pitch range from 80 to 600 Hz. If Y is periodic, the minima of AMDF encode the period. In our algorithm, six candidates per frame are extracted from the local minima of AMDF. After the candidate extraction, the Viterbi search algorithm is utilized to derive the most probable pitch contour across frames. The objective function (Obj) and inter-frame transition cost (Cost) are defined by following two equations.

$$Obj_t(c_j)=w_1*\min_{c_i}(Obj_{t\text{-}1}(c_i)+Cost_{t\text{-}1,t}(c_i,c_j))+AMDF(t,c_j) \quad (6)$$

$$Cost_{t\text{-}1,t}(c_i,c_j)=\begin{cases}\left|\log(\frac{c_j}{c_i})\right|, & \text{if current frame t is vocal}\\ 0, & \text{otherwise}\end{cases} \quad (7)$$

where c_i and c_j are the lag values of the i-th and j-th pitch candidates; and w_1 is a weighting parameter. The optimal path is found by minimizing $Obj_T(c_j)$ for every c_j (T denotes the index of the last frame).

2.3. Vocal/music separation and synthesis

Based on the estimated pitch contour, a binary mask is generated for the mixture spectrogram to segregate the singing voice. The binary mask for a certain T-F unit is set to one if the following condition is met.

$$\frac{AMDF_{tf}(\tau_t)-\min_{\tau}(AMDF_{tf}(\tau))}{\max_{\tau}(AMDF_{tf}(\tau))-\min_{\tau}(AMDF_{tf}(\tau))}\le\theta \quad (8)$$

where $AMDF_{tf}$ is the AMDF of the waveform in the frame t and the frequency-bin f; τ_t is the lag corresponding to the estimated pitch value in the frame t; and θ is set to 0.5 empirically. Finally, the singing voice is synthesized using the method in [18]. Fig. 3 shows an output example of the proposed system. The waveforms and spectrograms of the original vocal, the mixture and the separated vocal are demonstrated.

Table 1: *Performance (GNSDR) using different P and w_1*

		-5 dB	0 dB	5 dB	Average
P=0.05	w_1=0	2.40	3.61	3.31	3.12
	w_1=0.1	2.31	3.71	3.42	**3.15**
	w_1=1	0.39	2.72	2.91	2.01
	w_1=10	-1.23	0.41	0.31	-0.17
P=0.1	w_1=0	2.35	3.49	2.67	2.84
	w_1=0.1	2.27	3.61	2.79	2.89
	w_1=1	0.34	2.64	2.32	1.77
	w_1=10	-1.29	0.34	-0.12	-0.36
P=0.15	w_1=0	2.18	3.12	1.51	2.27
	w_1=0.1	2.12	3.25	1.62	2.33
	w_1=1	0.19	2.32	1.25	1.25
	w_1=10	-1.43	0.04	-0.93	-0.77

3. Evaluation

Our system is evaluated under different parameter settings and compared with three other systems [8][9][10]. Performance of our system under two ideal conditions with given true vocal position (IVP: short for "Ideal case with true Vocal Position"), and true pitch value plus true vocal position (IPV: short for "Ideal case with true Pitch Value") is also derived. In addition, the upper bound of separation performance using the ideal binary mask (IBM) [19], where the original vocal and music waveforms are assumed available, is given for comparisons.

3.1. Dataset

The MIR-1K dataset [7], which is commonly used in vocal/music separation tasks, is used in our evaluations. It contains 1000 song clips sampled at 16K Hz and with durations of 4 to 13 seconds. The singing voice and background music were recorded separately. Annotations, including lyrics, pitch contours, ground truths of vocal, music and unvoiced frames, are provided. In our tests, the singing voice was mixed with music at 5, 0, and -5 dB SNRs.

3.2. Performance measure

The global normalized signal to distortion ratio (GNSDR), which was proposed in [4] and used in [8][9][10], was selected as the performance measure. The GNSDR is derived as follows.

First, the signal to distortion ratio (SDR) is defined as:

$$SDR(s,\hat{s})=10\log_{10}\frac{\langle s,\hat{s}\rangle^2}{\|s\|^2\|\hat{s}\|^2-\langle s,\hat{s}\rangle^2} \quad (9)$$

where s and $\hat{s}$ are original and estimated singing voice signals; and $\langle\cdot\rangle$ and $\|\cdot\|$ denote the inner product and L_2 norm. Then, the NSDR is derived as:

$$NSDR(s,x,\hat{s})=SDR(s,\hat{s})-SDR(s,x) \quad (10)$$

where x is the mixture of the singing voice and background music. Finally, the GNSDR is a weighted NSDR by

$$GNSDR(s,x,\hat{s})=\frac{\sum_{i=1}^{N}NSDR(s_i,x_i,\hat{s}_i)\cdot L_i}{\sum_{i=1}^{N}L_i} \quad (11)$$

where L_i is the duration of the i-th song and N is the total number of song clips.

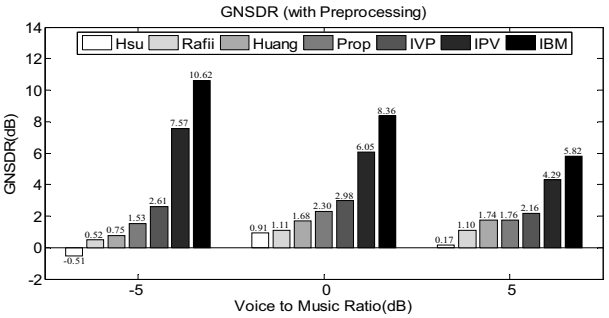

Figure 4: *Performance of the proposed system (Prop), three existing systems (Hsu [8], Rafii [9], and Huang [10]) and three ideal cases under various SNRs.*

3.3. Experiment

There are two parameters, P and w_1, in our system. The parameter P controls the threshold for vocal/music discrimination, i.e., it settles the trade-off between the hit rate of singing voice and the hit rate of music. The parameter w_1 determines the smoothness of the estimated pitch contour. To select optimal parameters for our system, simulations were carried out for different parameter settings. Table I shows GNSDR of our system for all parameter combinations ($P \in \{0.05, 0.1, 0.15\}$ and $w_1 \in \{0, 0.1, 1, 10\}$) at three SNRs using the MIR-1K dataset. Clearly, P=0.05 and w_1=0.1 gives the highest average GNSDR.

Separation results of the proposed system (Prop) with P=0.05 and w_1=0.1 and three other systems (Hsu [8], Rafii [9] and Huang [10]) are shown in Fig. 4. Performance of two ideal cases of our system (IVP and IPV) and of using IBM is also demonstrated for comparisons. Under all test conditions, our proposed system outperforms the three other systems using GNSDR measure. Meanwhile, one can observe that the performance difference between IPV and IVP is larger than the difference between IVP and Prop under all test conditions. This suggests that the performance bottleneck of our system is created by the pitch extraction algorithm.

4. Conclusion and Discussions

In this paper, we propose a novel feature for singing voice detection. This feature is derived from a spectro-temporal analysis framework, motivated by observed spectro-temporal receptive fields of neurons in the primary auditory cortex, for the Fourier spectrogram. Next, we utilize a Viterbi-based algorithm to extract a pitch contour in voice frames and separate the singing voice from background music using a binary mask based on the estimated pitch. Experimental results show that our system performs better than three other binary-mask based systems in vocal/music separation tasks.

One potential future work is to replace the simple threshold mechanism with a more complicated classifier for singing voice detection. However, the performance gain will be rather limited according to IVP results shown in Fig. 4. A major improvement will come from a better pitch estimation algorithm. Developing a robust pitch extraction algorithm and comparing performance with other not-mask-based state-of-the-art systems, such as the one in [20], will be pursued in the future.

5. Acknowledgements

This research is supported by National Science Council, R.O.C. under Grant NSC 101-2220-E-009-065 and Chunghwa Telecom Co., Ltd.

6. References

[1] C. K. Wang, R. Y. Lyu, and Y. C. Chiang, "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker," in *Proc. EUROSPEECH*, pp. 1197–1200, 2003.

[2] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "LyricSynchronizer: Automatic Synchronization System between Musical Audio Signals and Lyrics," *IEEE J. Select. Top. Signal Process.*, vol. 5, no. 6, pp. 1252–1261, 2011.

[3] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer Identification in Polyphonic Music Using Vocal Separation and Pattern Recognition Methods," in *Proc. ISMIR*, pp. 375–378, 2007.

[4] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. WASPAA*, pp. 90–93, 2005.

[5] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. ISMIR*, pp. 337–344, 2005.

[6] A. S. Bregman, *Auditory Scene Analysis: the perceptual organization of sound,* MIT Press, 1990.

[7] Y. Li and D. L. Wang, "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, 2007.

[8] C.-L. Hsu and J.-S. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, 2010.

[9] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *Proc. ICASSP*, pp. 221–224, 2011.

[10] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-Voice Separation From Monaural Recordings Using Robust Principal Component Analysis," in *Proc. ICASSP*, pp. 57–60, 2012.

[11] A. L. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," in *Proc. WASPAA*, pp. 119–122, 2001.

[12] C. Wu and G. Liang, "Robust singing detection in speech/music discriminator design," in *Proc. ICASSP*, pp. 865–868, 2001.

[13] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.

[14] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no.3, pp. 920–930, 2006.

[15] T.-S. Chi and C.-C. Hsu, "Multiband analysis and synthesis of spectro-temporal modulations of Fourier spectrogram," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. EL190–EL196, 2011.

[16] C.-C. Hsu, T.-E. Lin, J.-H. Chen, and T.-S. Chi, "Spectro-temporal subband wiener filter for speech enhancement," in *Proc. ICASSP*, pp. 4001–4004, 2012.

[17] J. Sundberg, "The Perception of Singing," in *The Psychology of Music*, D. Deutsch, Ed., second ed, 1998.

[18] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

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

[20] J.-L. Durrieu, G. Richard, B. David, and C. Fevotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 564-575, 2010.