MIREX 2014 Submission for Singing Voice Separation

Frederick Yen
Master Program of SMIT
National Chiao-Tung University, Taiwan
fredyen1@gmail.com

Tai-Shih Chi
Dept. of Elec. & Comp. Engineering
National Chiao-Tung University, Taiwan
tschi@mail.nctu.edu.tw

ABSTRACT

In this submission, the music recordings are first transformed into auditory spectrograms. After extracting the spectral-temporal modulation contents of the time-frequency (T-F) units through a two-stage auditory model, we define modulation features pertaining to three categories in music audio signals: *vocal*, *harmonic*, and *percussive*. The T-F units are then clustered into three categories in a two-stage clustering process and the singing voice is synthesized from T-F units in the vocal category via time-frequency masking. This submission is an extended work from [1].

1. INTRODUCTION

Music instruments produce signals with various kinds of fluctuations such that they can be briefly categorized into two groups, *percussive* and *harmonic*. Signals produced by percussive instruments are more consistent along the spectral axis and by harmonic instruments are more consistent along the temporal axis with little or no fluctuations. These two categories occupy a large proportion of a spectrogram with mainly vertical and horizontal lines. To extend this sense into a more general form, the fluctuations can be viewed as a sum of sinusoid modulations along the spectral axis and the temporal axis. If a signal has nearly zero modulation along one of the two axes, its energy is smoothly distributed along that axis. Conversely, if a signal has a high frequency of modulation along one axis, then its energy becomes scattered along that axis. Therefore, if one can decipher the modulation status of a signal, one may be able to identify the instrument type of the signal.

Since modulations are important for music signal categorization, this modulation-decomposition auditory model is used as a pre-processing stage for singing voice separation in this paper. Our proposed unsupervised algorithm adapts this two-stage auditory model, which decodes the spectro-temporal modulations of a T-F unit, to extract modulation based features and performs two-stage singing voice separation under the CASA framework. A brief review of the auditory model is presented in Section 2. Section 3 describes the proposed method. Section 4 shows evaluation and results.

2. SPECTRO-TEMPORAL AUDITORY MODEL

A neuro-physiological auditory model is used to extract the modulation features. The model consists of an early cochlear (ear) module and a central auditory cortex (A1) module.

2.1 Cochlear Module

The input sound goes through 128 overlapping asymmetric constant-Q band-pass filters ($Q_{3dB} \gg 4$) whose center frequencies are uniformly distributed over 5.3 octaves with the 24 filters/octave frequency resolution. These constant-Q filters mimic the frequency selectivity of the cochlea. Outputs of these filters are then transformed through a non-linear compression stage, a lateral inhibitory network (LIN), and a half-wave rectifier cascaded with a low-pass filter. The non-linear compression stage models the saturation caused by inner hair cells, the LIN models the spectral masking effect, and the following stage serves as an envelope extractor to model the temporal dynamic reduction along the auditory pathway to the midbrain. Detailed descriptions of the cochlear module can be found in [2].

The output of the module is the auditory spectrogram, which represents the neuron activities along time and log-frequency axis. In this work, we bypass the non-linear compression stage by assuming input sounds are properly normalized without triggering the high-volume saturation effect of the inner hair cells.

2.2 Cortical Module

The second module simulates the neural responses of the auditory cortex (A1). The auditory spectrogram is analyzed by cortical neurons which are modeled by two-dimensional filters tuned to different spectro-temporal modulations. The rate parameter (in Hz) characterizes the velocity of local spectro-temporal envelope variation along the temporal axis. The scale parameter (in cycle/octave) characterizes the density of the local spectro-temporal envelope variation along the log-frequency axis. Furthermore, the cortical neurons are found sensitive to the direction of the spectro-temporal envelope. It is characterized by the sign of the rate para- meter in this model, with negative for the upward direction and positive for the downward direction. Detailed description of the cortical module is available in [3].

3. PROPOSED METHOD

3.1 Feature Extraction

Interpreting the instrument characteristics from the rate-scale perspective, several general properties can be drawn. Harmonic components can be usually regarded as having low rate and high scale modulations.

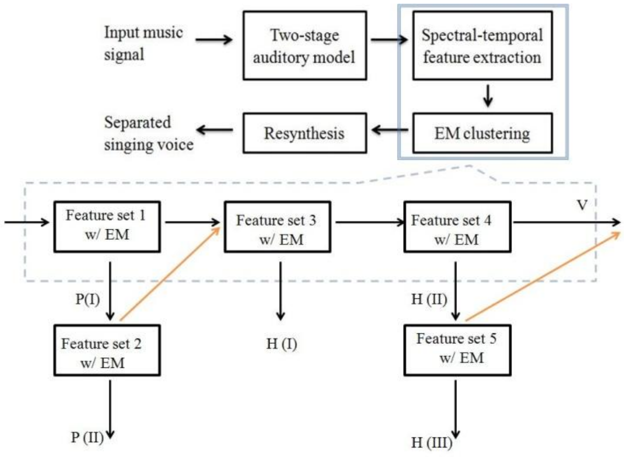


Figure 1. Block diagram of the submitted algorithm.

A schematic diagram of the proposed algorithm is shown in Figure 1.

It means that they have relatively slow energy change along time and rapid energy change along the log-frequency axis due to the harmonic structures. In contrast, percussive components typically show quick energy change along time and energy spreading along the whole log-frequency axis, such that they possess high rate and low scale modulations. Vocal components are often recognized as a mix version of the harmonic and percussive components with characteristics sometimes considered more similar to harmonics.

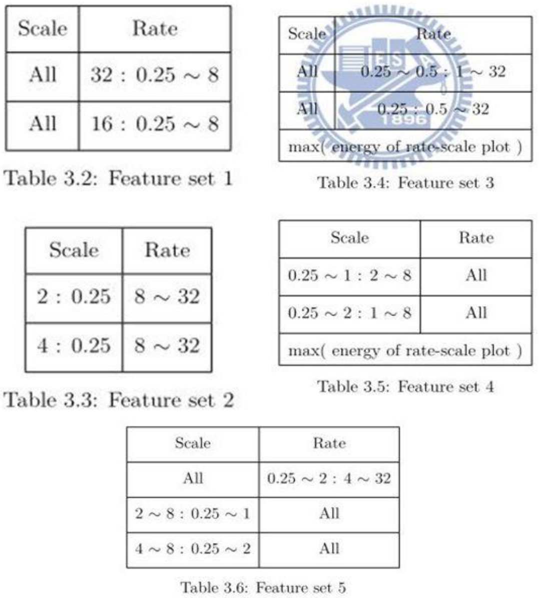
Given an auditory spectrogram transformed from an input music signal, the rate-scale plots of the T-F units are generated. As a pre-process, in order to prevent extracting trivial data from nearly inaudible T-F units of the auditory spectrogram, we leave out the T-F units that have energy less than 1% of the maximum energy of the whole auditory spectrogram. With the rest of the T-F units, we obtain the rate-scale plot of each unit and proceed to the feature extraction stage.

For each rate-scale plot, the total energies of the negative and positive rate side are compared. The side with greater energy is determined as the dominant plot. From the dominant plot, we extract features from the feature sets shown in Tables 3.2 ~ 3.6.

3.2 Unsupervised Clustering

After feature extraction in each stage, unsupervised clustering is followed. The spectrogram is divided into three parts consisting of channel 1 to channel 60, channel 46 to channel 75, and channel 61 to channel 128, respectively, with overlaps of 15 channels.

The clustering step is performed using the EM algorithm to group data into two unlabelled clusters. The EM algorithm assigns a probability set to each T-F unit showing its likelihood of belonging to each cluster. Each of the three sub-spectrograms is clustered into two groups.



Scale	Rate
All	32 : 0.25 ~ 8
All	16 : 0.25 ~ 8

Table 3.2: Feature set 1

Scale	Rate
All	0.25 ~ 0.5 : 1 ~ 32
All	0.25 : 0.5 ~ 32
max(energy of rate-scale plot)	

Table 3.4: Feature set 3

Scale	Rate
2 : 0.25	8 ~ 32
4 : 0.25	8 ~ 32

Table 3.3: Feature set 2

Scale	Rate
0.25 ~ 1 : 2 ~ 8	All
0.25 ~ 2 : 1 ~ 8	All
max(energy of rate-scale plot)	

Table 3.5: Feature set 4

Scale	Rate
All	0.25 ~ 2 : 4 ~ 32
2 ~ 8 : 0.25 ~ 1	All
4 ~ 8 : 0.25 ~ 2	All

Table 3.6: Feature set 5

Table 3.2~3.6. Feature sets 1~5.

Total of six groups are generated and merged back into two whole spectrograms by comparing the correlations of the overlapped channels between different groups. With no prior information about the labels of the two whole spectrograms, criterions at each stage are used to select the vocal spectrogram. The vocal spectrogram is then synthesized to an estimated signal using the auditory model toolbox [7].

4. EVALUATION RESULTS

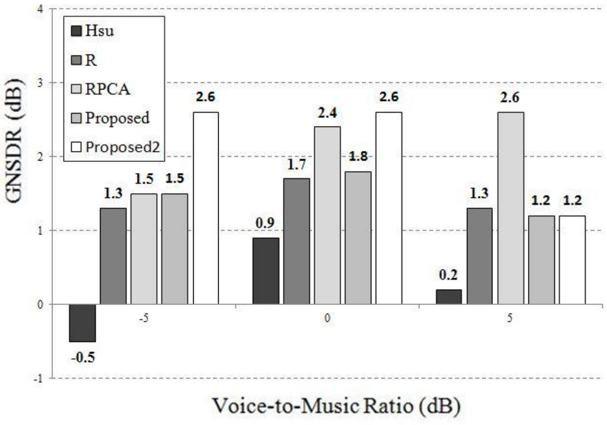


Figure 5. GNSDR comparison at voice-to-music ratio of -5, 0, and 5 dB with existing methods.

The MIR-1K [4] is used as the evaluation dataset. The SDR ratios [5] are computed by the BSS Eval toolbox v3.0 [6]. We compute the GNSDR to compare with other proposed algorithms listed in [1].

From Figure 5, we can observe that this submission (proposed2) has the highest performance in 0 and -5 dB SNR conditions. In the 5 dB SNR condition, the performance of proposed2 is comparable to the performance of REPET.

5. REFERENCES

[1] F. Yen, Y. Luo, and T. Chi, “Singing Voice Separation using Spectro-Temporal Modulations,” *to be presented at the Int. Soc. for Music Inform. Retrieval Conf.*, 2014.

[2] 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.

[3] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, Vol. 106, No. 5, pp. 2719-2732, 1999.

[4] 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. on Audio, Speech, and Language Process.*, Vol. 18, No. 2, pp. 310-319, 2010.

[5] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Process.*, Vol. 14, No. 4, pp. 1462-1469, 2006.

[6] http://bass-db.gforge.inria.fr/bss_eval/

[7] http://www.isr.umd.edu/Labs/NSL/nsl.html