A Simple Music/Voice Separation Method based on the Extraction of the Repeating Musical Structure

Zafar RAFII & Bryan PARDO

Northwestern University, EECS Department, Evanston, IL, USA.

zafarrafii@u.northwestern.edu • pardo@northwestern.edu • http://music.cs.northwestern.edu/

Introduction

Repetition "is the basis of music as an art" (Schenker, 1954). This is especially true for popular songs, generally characterized by an underlying repeating musical structure over which the singer performs varying lyrics. Based on this simple observation, we propose to extract the repeating musical background from the non-repeating musical foreground. The basic idea is to identify the periodically repeating audio segments, compare them to a repeating segment model, and extract the energy corresponding to the repeating patterns. The result is a simple but effective music/voice separation system.

Method

▶ Step 1: Identify the repeating period of the structure

- Compute the "beat spectrum" $\mathbf{b}$ from the spectrogram $\mathbf{V}$.

$$B(i,j) = \frac{1}{m-j+1} \sum_{k=1}^{m-j+1} V(i,k)^2 \, V(i,k+j-1)^2$$
$$b(j) = \frac{1}{n} \sum_{i=1}^{n} B(i,j) \quad (1)$$
$$\text{for } i = 1 \dots n \text{ and } j = 1 \dots m$$

- Identify the peak with the largest magnitude and longest period $\mathbf{p}$.

▶ Step 2: Compute the repeating segment model

- Segment the spectrogram $\mathbf{V}$ at period rate $\mathbf{p}$.
- Compute the repeating model $\overline{\mathbf{V}}$ as the mean of the segments in $\mathbf{V}$.

$$\overline{V}(i,l) = \left(\prod_{k=1}^{r} V(i, l+(k-1)\,p) \right)^{\frac{1}{r}} \quad (2)$$
$$\text{for } i = 1 \dots n \text{ and } l = 1 \dots p$$

▶ Step 3: Build the repeating binary time-frequency mask

- Compute the mean-scaled spectrogram $\widetilde{\mathbf{V}}$ using the model $\overline{\mathbf{V}}$.

$$\widetilde{V}(i, l+(k-1)\,p) = \left| \log \left(\frac{V(i, l+(k-1)\,p)}{\overline{V}(i,l)} \right) \right| \quad (3)$$
$$\text{for } i = 1 \dots n, \; l = 1 \dots p \text{ and } k = 1 \dots r$$

- Build the binary time-frequency mask $\mathbf{M}$ by assigning time-frequency bins in $\widetilde{\mathbf{V}}$ below a tolerance factor $\mathbf{t}$ to the repeating structure.

$$M(i,j) = \begin{cases} 1 & \text{if } \widetilde{V}(i,j) \leq t \\ 0 & \text{otherwise} \end{cases} \quad (4)$$
$$\text{for } i = 1 \dots n \text{ and } j = 1 \dots m$$

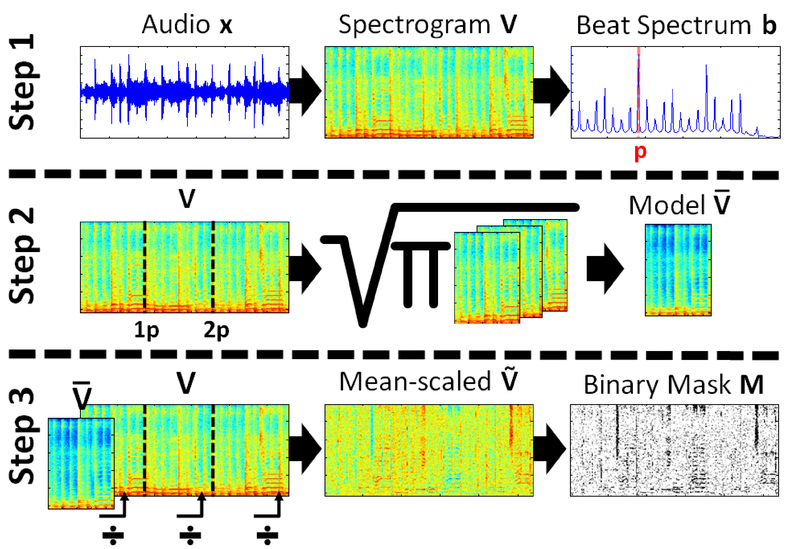

Figure: Overview of the repeating background/non-repeating foreground separation.

Evaluation

▶ Dataset: MIR-1K Dataset[1]

- 1000 song clips, recorded at 16 kHz, from 4 to 13 sec
- clips from 110 karaoke Chinese pop songs performed by amateurs
- includes manual annotations of the pitch contours, indices of the vocal/non-vocal frames, and indices and types for unvoiced frames

▶ Competing method: Hsu *et al.*'s music/voice separation[2]

- Vocals separation using pitch-based inference (best automatic version)
- + detection and separation of unvoiced vocal frames
- + spectral subtraction method to enhance voiced vocals separation

▶ Mixing process and separation measure (see Bars):

- 3 sets of mixtures: 3 voice-to-music mixing ratios (-5, 0, 5 dB)
- Global Normalized Signal-to-Noise Ratio (GNSDR) for each set

▶ Potential enhancements for our system (see Box plots):

- Use of an optimal period $\mathbf{p}$
- Use of an optimal tolerance $\mathbf{t}$
- Use of the index information of the vocal frames

[1] http://sites.google.com/site/unvoicedsoundseparation/mir-1k

[2] Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310-319, February 2010.

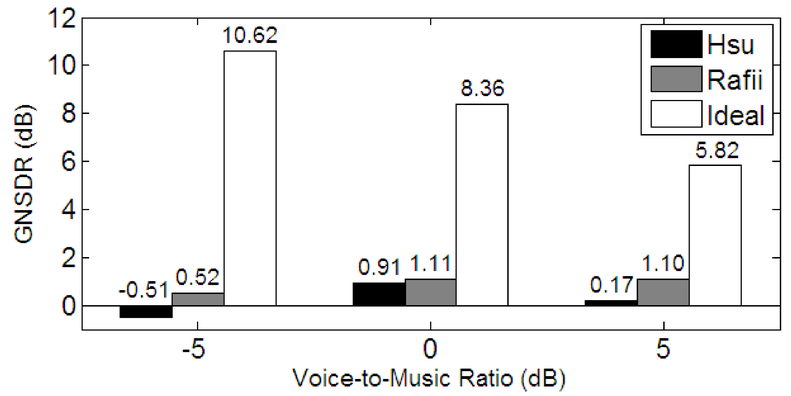

Figure: Comparison of the global separation performance between the best automatic version of Hsu *et al.*'s music/voice separation method (black), our automatic method (gray) and the ideal binary mask (white), averaged over all the mixtures for 3 different voice-to-music ratios. Higher values are better.

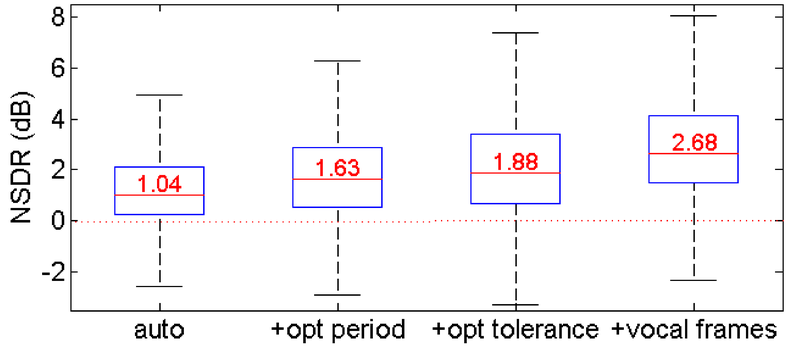

Figure: Separation performance of our automatic music/voice separation method and its successive potential enhancements (+optimal period, +optimal tolerance, +vocal frames) for all the mixtures at voice-to-music ratio of 0 dB. Higher values are better.

Conclusion

We have proposed a novel and promising separation method based on the extraction of the repeating patterns in the musical structure. Evaluation on a dataset of 1,000 song clips showed that this method can be successfully applied for music/voice separation. Unlike other music/voice separation approaches, this method does not depend on particular features, rely on complex frameworks, or require prior training. Because it is only based on self-similarity, it has the advantage of being simple, fast, blind, and thus completely automatable.

This work was supported by NSF grant numbers IIS-0643752 and IIS-0757544.