SINGING VOICE SEPARATION USING ADAPTIVE WINDOW HARMONIC SINUSOIDAL MODELING

Preeti Rao
I.I.T. Bombay
prao@ee.iitb.ac.in

Nagesh Nayak
SensiBol Audio Technologies
nageshsnayak@sensibol.com

Sharath Adavanne
SensiBol Audio Technologies
sharath@sensibol.com

ABSTRACT

This abstract describes our submission to the MIREX 2014 Singing Voice Separation task. Our system uses a support vector machine (SVM) classifier based singing voice detection block to first label vocal and instrumental in the polyphonic audio. This is followed by a predominant pitch detection algorithm in the vocal regions. The detected vocal pitch is used in the high-resolution harmonic sinusoidal modeling of the audio to isolate the frequency-time tracks corresponding to the singer's vocal harmonics. The singing voice is reconstructed from the estimated sinusoidal model parameters. The synthesized voice is then subtracted from the original polyphonic audio to obtain the separated instrumental track.

1. INTRODUCTION

Music soundtracks are created in recording studios by recording monophonic tracks from different music sources onto different channels. For a stereophonic soundtrack, the vocals are generally panned in the center. Subtracting one channel from the other in such soundtracks cancels out the vocals. Similarly, by identifying and isolating common spectral components between the two channels, the vocals can be reconstructed. However, use of this method fails when other instruments are panned in the center along with the vocals. Apart from this, a lot of popular music has been recorded during earlier times when only monophonic recording technology was available. In the absence of spatial cues, alternate methods for singing voice separation have to be devised for relying entirely on mono cues.

2. ALGORITHM DESCRIPTION

A singing voice detection module trained using an SVM classifier segments the polyphonic audio into "vocal" regions corresponding to regions where the singers' voice dominates and "instrumental" regions where the accompaniment dominates. A polyphonic pitch tracker is then used to estimate the pitch corresponding to the predominant source in only the vocal regions. The instrument regions are left unprocessed. Harmonic Sinusoidal Modeling is performed by picking peaks corresponding to the pitch and its corresponding harmonics in the spectrum and forming tracks between peaks in adjacent frames. These tracks are synthesized to obtain the vocal signal and then subtracted from the original polyphonic mixture to get the instrumental signal.

2.1 Singing Voice Detection (SVD)

The vocal regions, or singing spurts, in the audio are obtained by the classification of frame-level acoustic features that represent vocal characteristics. The features are extracted from harmonic sinusoidal modeling. Static and dynamic features capturing timbre and pitch behaviour are computed over fixed intervals [1]. An SVM classifier [2] is trained on a few hours of labeled data to achieve vocal-instrumental classification on the fixed duration data windows. Next, SVM classifier likelihood values and temporal continuity are applied to achieve the segmentation of vocal regions from the full audio in a manner that is biased towards vocal segments. While long instrumental interludes are detected correctly, shorter instrumental sections sometimes get absorbed within vocal segments.

2.2 Predominant f0 Estimation

A Two-Way Mismatch (TWM) based pitch tracker is used in this stage [3]. The TWM PDA falls under the category of harmonic matching PDAs that are based on the frequency domain matching of measured spectrum with an ideal harmonic spectrum**.** The TWM PDA is operated within the framework of dynamic programming-based (DP) smoothing. DP uses a combination of suitably defined local measurement and smoothness costs into a global cost, which is optimized over a continuous voiced segment. The output of this stage are the voice-pitch values for each vocal segment.

2.3 Harmonic Sinusoidal Modeling (HSM)

In this stage, the audio data is first windowed using a fixed duration Hamming window for analysis. For each analysis frame, sinusoidal components are identified by performing a DFT on the windowed signal followed by peak picking [4] in the magnitude spectrum of the signal. The phases of the signal are preserved for performing the synthesis.

The use of a fixed window length has its drawbacks. If the predominant pitch is low, use of a larger window ensures good frequency resolution for better separation of harmonics. However, a large window results in poor amplitudes and frequency estimates of sinusoids at higher harmonics for regions with large time-variations of pitch by the singer [5].

To overcome these issues, the sinusoidal components in each frame and each frequency range are chosen from a combination of three possible windows 10ms, 20ms and 40ms by taking into account the instantaneous pitch and inter-hop pitch differences. Using empirical thresholds on the instantaneous pitch and inter-hop pitch differences, the sinusoids are chosen from one out of the three windows. HSM [6] is then used along with the vocal pitch for identifying sinusoidal peaks lying along integer multiples of the pitch thus creating the tracks representing the vocal harmonics in the polyphonic audio track.

2.4 Synthesis

The isolated voice signal is synthesized using the parameters of the sinusoids estimated per frame in each HSM track. The synthesis involves linear interpolation of amplitudes and cubic phase interpolation of phases between sinusoids in the same track across successive frames [7]. This separated signal (voice) is subtracted from the original audio signal to obtain the separated instrumental signal.

3. REFERENCES

[1] V. Rao, C. Gupta and P. Rao, "Context-aware Features for Singing Voice Detection in Polyphonic Music," *Proc. of Adaptive Multimedia Retrieval*, 2011.

[2] C. Chang and C. Lin, "LIBSVM : a library for support vector machines". *ACM Trans. Intelligent Systems and Technology*, vol.2, no. 3, pp. 27:1-27, 2011.

[3] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music", *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 8, pp. 2145–2154, Nov. 2010.

[4] D. Griffin and J. Lim, "Multiband Excitation Vocoder," *IEEE Trans. Acoust., Speech and Sig. Process*., vol. 36, no. 8, pp. 1223–1235, 1988

[5] V. Rao, P. Gaddipati, and P. Rao, "Signal-Driven Window-Length Adaptation for Sinusoid Detection in Polyphonic Music," *IEEE Trans. Audio, Speech, Lang. Process.* , vol. 20, no. 1, pp. 342–348, 2012.

[6] Xavier Serra: *Musical Sound Modeling with Sinusoids plus Noise*, pages 91–122, Studies on New Music Research, Swets & Zeitlinger, 1997.

[7] R. McAulay and T. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoustics, Speech and Signal Process.*, vol 34, no. 4, pp. 744-754, 1986.