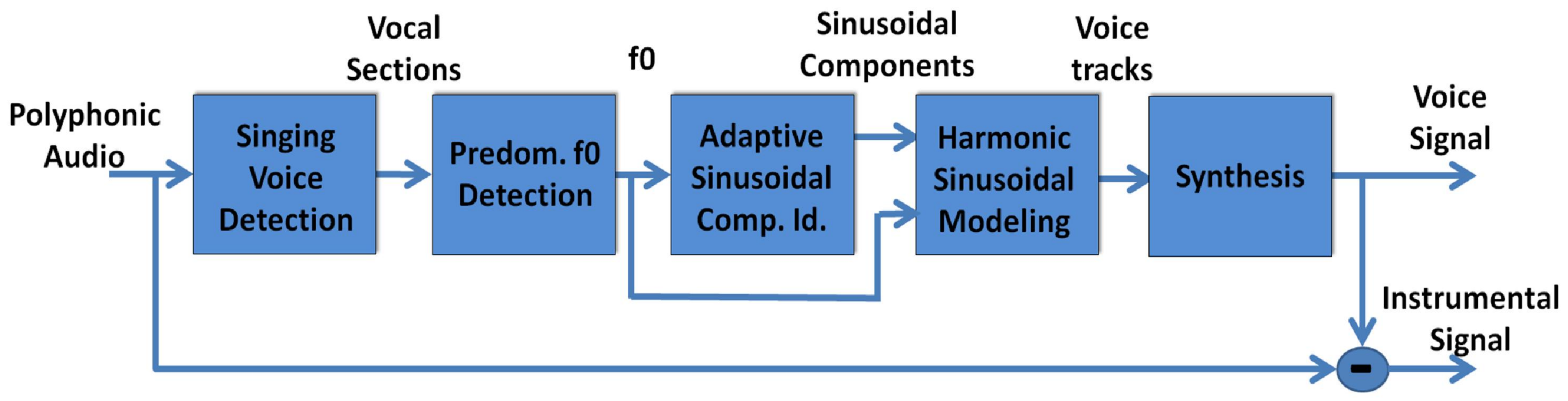
Singing Voice Separation using Adaptive Window Harmonic Sinusoidal Modeling MIREX 2014

Preeti Rao, Nagesh Nayak, and Sharath Adavanne
prao@ee.iitb.ac.in,{nagesh,sharath}@sensibol.com

Introduction

Our system uses a support vector machine (SVM) classifier based singing voice detection block to first label vocal and instrumental regions 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. Singing Voice Detection (SVD)

- Frame-level acoustic features that represent vocal characteristics are extracted from harmonic sinusoidal modeling.
- Static and dynamic features capturing timbre and pitch behaviour are computed over fixed intervals [1].
- SVM classifier trained on a few hours of labeled data used to achieve vocal-instrumental classification on the fixed duration data windows. Classifier is biased towards identifying vocal segments correctly.
- The output of this stage are the vocal and instrumental regions in the audio.

2. Predominant f0 Estimation

- A Two-Way Mismatch (TWM) PDA based on the frequency domain matching of measured spectrum with an ideal harmonic spectrum is used [2].
- 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.

3. Adaptive Sinusoidal Components Identification

- Sinusoidal components are identified by performing a DFT on the windowed signal followed by peak picking in the magnitude spectrum of the signal.
- Phases of the signal are preserved for performing the synthesis.
- The use of a fixed analysis window 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 [3].
- 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, sinusoids are chosen from one of the three windows.

4. Harmonic Sinusoidal Modeling (HSM)

- HSM [4] is used along with the vocal pitch for identifying sinusoidal peaks lying along integer multiples of the pitch thus creating tracks representing only the vocal harmonics in the polyphonic audio track.

5. 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 [5].
- The separated signal (voice) is subtracted from the original audio signal to obtain the separated instrumental signal.

MIREX 2014 GNSDR Summary Results

Algorithm	Voice GNSDR (dB)	Music GNSDR (dB)	Runtime (hh)
GW1	2.8861	5.2549	24
HKHS1	-1.3988	0.3483	06
HKHS2	-1.9413	0.5239	06
HKHS3	-2.4807	0.1414	06
IIY1	4.2190	7.7893	02
IIY2	4.4764	7.8661	02
JL1	4.1564	5.6304	01
LFR1	0.6499	3.0867	03
RNA1	**3.6915**	**7.3153**	**06**
RP1	2.8602	5.0306	01
YC1	-0.8202	-3.1150	13

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] 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, Nov. 2010.
[3] 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, 2012.
[4] Xavier Serra: *Musical Sound Modeling with Sinusoids plus Noise*, pages 91–122, Studies on New Music Research, Swets & Zeitlinger, 1997.
[5] R. McAulay and T. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoustics, Speech and Signal Process.*, vol 34, no. 4, 1986.