SINGING VOICE SEPARATION BY HARMONIC MODELING

Georgi Dzhambazov, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

1. INTRODUCTION

In this work we suggest how to separate singing voice from an audio mixture relying on an an approach for harmonic modeling. The idea of the harmonic model is to represent the spectral content of a predominant source with harmonic nature as a sum of the fundamental frequency and its partials. Harmonics are function of input detected pitch contours. Therefore, a pitfall for detecting singing voice might be pitch segments coming from soloing instruments with harmonic nature. In [3] are suggested a set of timbral features that help discriminate between regions of singing voice and musical instruments. We filter out instrumental regions by classifying the harmonic content, based on these features.

Although vocal modeling could be a conceptual part in some source separation (SS) approaches [1], in most of them, vocal detection (VD) is not done as an explicit step. Only recently it has been shown that VD can enhance the separated signal as a post-processing step to SS [2].

2. APPROACH

2.1 Pitch detection

The harmonic modeling requires as input the pitch series of the main source. We rely on the *Melodia* algorithm [4] to extract the pitch for time intervals with predominant melodic source (pitch contours). The methodology computes for each contour how harmonically salient it is, based on the saliences of each pitch value. Then contours from the main source are selected as having salience over a threshold relative to the average mean salience of all contours. We used the open-source implementation from the feature extraction framework *essentia* [1] , in which we have set the voicing threshold to 1.4, and guessUnvoiced=True which increased the recall of the vocal regions . A side effect of that is though, that some instrumental time intervals increase the false positives rate.

2.2 Harmonic modeling

The harmonic model [5] filters the spectral peaks corresponding to the first 30 harmonic partials of the singing

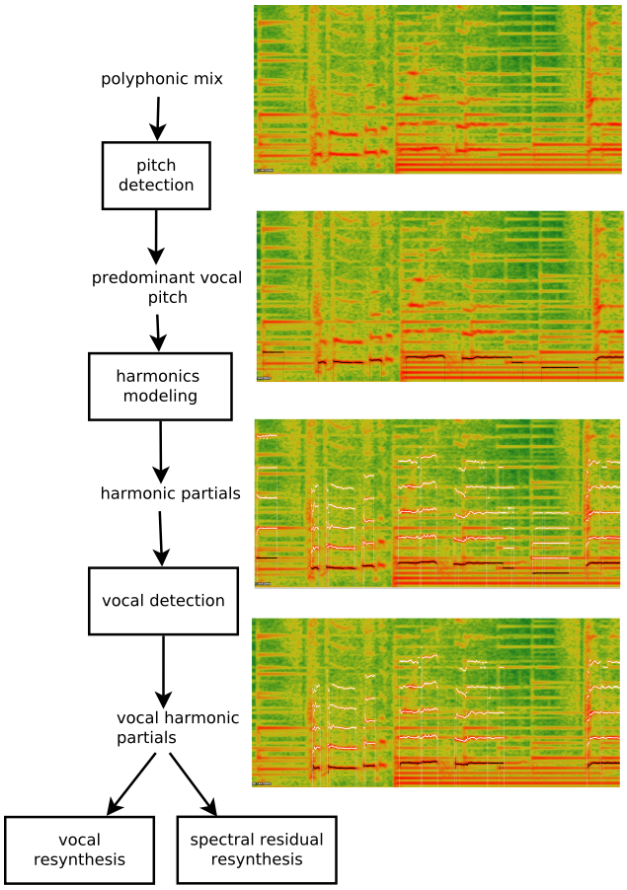

Figure 1. Overview of approach.

[1] http://essentia.upf.edu

	before VD	after VD
recall	0.83	0.76
false alarms	0.35	0.22

Table 1. Vocal detection (VD) results

voice.

$$Yh[k] = \sum_{r=1}^{R} A_r W[k - r\hat{f}_0]$$

2.3 Vocal detection

To filter out the regions from background instruments, we have trained a vocal classifier following the approach of [3]. Like them we trained a random forest classifier (100 trees in *scikit-learn*) but used only timbral features: 30 MFCCs and 'vocal variance' (the variance of the first 5 MFCCs). The motivation is that timbral-based vocal classification complements the vocal detection approach of *Melodia*, relying only on pitch-contour salience. A frame is classified as voiced or non-voiced based on majority voting of the trees. Detected non-vocal frames were muted from the estimated voice.

2.3.1 Training

[2] indicates that VD accuracy of the separated signal is higher when trained also on separated signal rather than on the original mix. We trained thus on harmonic partials extracted from the original mix using all avaialble audio of the *iKala* dataset of around 125 minutes [2]. The reference MIDI annotation is used as input to the harmonic model.

2.4 Resynthesis

The vocal source is resynthesized by means of a constant overlap add resynthesis with the *sms-tools* package [3]. Finally the background is derived by multuplying the original mix by a simple soft mask.

3. RESULTS

The proposed approach has been implemented in python [4]. Result on the test part of the iKala dataset are summarized in Table 1 [5].

Acknowledgements We are thankful to Bernhard Lehner for providing help with running the timbral feature extraction. This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583) and partly by the AGAUR research grant.

[2] http://mac.citi.sinica.edu.tw/ikala/description.html

[3] http://mtg.upf.edu/technologies/sms

[4] https://github.com/georgid/vocal-detection

[5] complete results are available on http://www.music-ir.org/mirex/wiki/2016:Singing_Voice_Separation_Results

	voice		accompaniment	
	mean	st dev	mean	st dev
NSDR	-2.281	3.534	0.395	1.470
SIR	6.562	9.778	1.984	9.805
SAR	2.394	4.562	2.708	2.661

Table 2. Results measured by source separation metrics: Normalized Signal-to-Distortion Ratio (NSRD), Signal-to-Interference Ratio (SIR), Signal-to-Artifacts Ratio (SAR)

4. REFERENCES

[1] Jean-Louis Durrieu, Bertrand David, and Gaël Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1180–1191, 2011.

[2] Bernhard Lehner and Gerhard Widmer. Monaural blind source separation in the context of vocal detection. In *présenté à 16th International Society for Music Information Retrieval Conference (ISMIR), At Malaga, Spain*, 2015.

[3] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner. On the reduction of false positives in singing voice detection. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7480–7484. IEEE, 2014.

[4] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[5] Xavier Serra. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Technical report, 1989.