# SINGING VOICE DETECTION BY TIMBRE-BASED PITCH CONTOUR CLASSIFICATION

*Georgi Dzhambazov, Juan Jose Bosch*

Music Technology Group, Universitat Pompeu Fabra, Barcelona

## ABSTRACT

In this work we tackle the problem of detecting regions with predominant singing voice in polyphonic music recordings. We consider a baseline an approach that classifies predominant pitch contours into vocal and instrumestal ones based on their relative melodic salience. Considering that singing voice fluctuates more frequently in pitch and timbre, the pitch contours are characterized additionaly by dynamic features, describing these fluctuations. The accuracy is evaluated both on benchmark datasets for western music and a newly compiled dataset from Turkish Makam singing. We show that the introduction of the new features helps reduce the false positive rate.

***Index Terms***— One, two, three, four, five

## 1. INTRODUCTION

The automatic detection of singing voice (SVD) in polyphonic music recordings is an important problem in audio signal processing and required primary step for other related tasks. When singing is the predominant source, the pitch trajectory can be naturally divided into short time segments, each representing a motif from main melody. One melodic segment usually has pitch and intensity that change smoothly over time. The predominant melody extraction method of [1] exploits this notion to form pitch contours: A candidate pitch contour is created by combining spectral bins with continuously changing harmonic salience. Each contour is then characterized by features of its pitch trajectory. Contours from predominant voice are selected, based on heuristics about the distribution of these features. We will refer to this method as MELODIA.

Bittner et al. [2] proposed to replace such heuristic contour selection by a discriminative classifier trained on the mentioned features. The overall melody extraction accuracy did not improve when tested on the MedleyDB dataset [**?**], but Bosch et al. [3] reported improvements on symphonic music, since the heuristic rules from MELODIA were not appropriate for such data. In both works though the reported vocal false alarm rate is still relatively high (in the order of 40).

Thanks to XYZ agency for funding.

In general, the timbre of the voice within a contour changes smoothly and is thus relatively coherent. In both [2] and [3] no timbre features are considered.

On the other hand, SVD approaches, which do not aim at estimating the pitch of the singing voice, usually distinguish voice by a frame-wise classifier trained on static spectral features (such as MFCCs or spectral flatness) [4]. Recent classification approach was based on the observation that temporal fluctuations of singing voice timbre discriminate it from other instruments [5]. The authors did a careful selection of the parameters of both conventional static timbre features and these describing timbral and pitch fluctuations. As a result, the baseline was improved by reducing voicing false alarm rate.

In this work we tackle the problem of SVD in the case when voice is the source of a predominant melody. We extend the contour-classification based approach of [2] by adding features to the classifier, which describe timbral temporal fluctuations. We investigated that one reason for the high false alarm of [2] is that contours from background instruments, misclassified as voice, have salience statistics relatively similar to vocal contours. To address these, we exploit some of the features, introduced by [5]. .

## 2. METHOD

An overview of the proposed method can be seen in Figure 1. First, candidate vocal pitch contours are extracted from the polyphonic mix along with a set of features describing the contour's pitch characteristics. Then fluctuation features are extracted from the isolated harmonic partials per contour, each with either vocal or instrumental origin. Next using both the features, derived during contour creation, and the fluctuation features a classifier is trained to discriminate the contour's source. The contours with low probability of being vocal are discarded. Finally, a single vocal melody line is decoded by means of Viterbi algorithm, because there might be regions, in which candidate contours are overlapping. We assume that singing voice in the dataset occurs only as 'lead instrument' (e.g. that voice as accompaniment is not present), therefore final vocal regions are determined uniquely by the derived melodic contour.

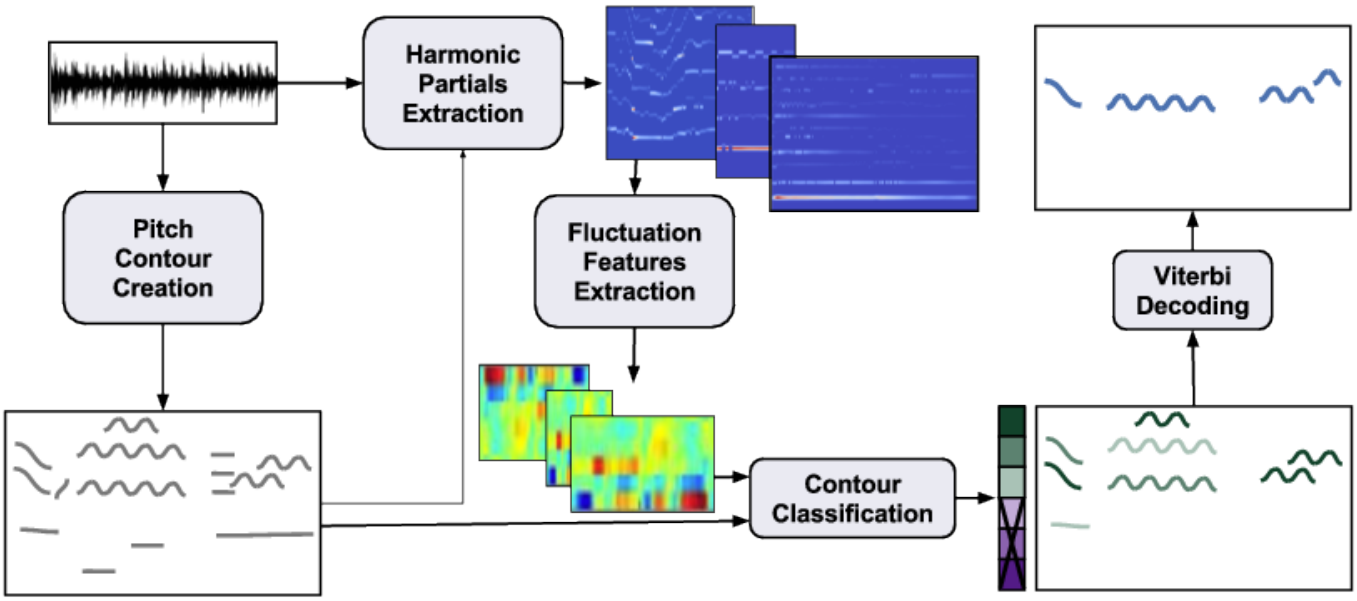


**Fig. 1**. Overview of the proposed method. Contour classification is based on a vector combined from the features, derived during contour creation, and the fluctuation features. The classifier yields a vocal probability for each contour (depicted with different colour). The contours with low probability are discarded and Viterbi decoding is

### 2.1. Contour creation

The contour creation step is adopted without modifications from MELODIA[1]. In a prepossessing step, peaks are extracted from the polyphonic spectrum, based on their harmonic salience. A set of contours $S$ is created by grouping sequences with salient peaks, which are continuous in time and pitch. The number of extracted contours depends on several parameters. We have varied one of them: the threshold on the extent of variation below the mean salience per recording $\tau_\sigma$. To obtain a higher recall of baseline we increased $\tau_\sigma = 1.1$ (default $\tau_\sigma = 0.9$, used in [2]).

Finally, created contours are assigned a number of 1-dimensional feature statistics, which we will refer to as features_SAL:

- Duration (from 0.1 to 2-3)
- Pitch (mean and st dev)
- Salience (mean, st dev and sum)
- Vibrato (presence, rate and extent)

### 2.2. Extraction of fluctuation features

The features_SAL do not represent the temporal variation of timbre of each contour. A challenge of capturing well timbral dynamics for the classification of instruments in polyphonic music was found to be the isolation of individual instrumental spectra [6]. Taking advantage of the pitch curve, for each candidate contour, we extract harmonic partials over time using harmonic sinusoidal modeling (see Figure 1). Harmonics represent ideally the spectral content of the contour's source. We then extract features from the interpolated harmonic spectrum $\bar{X}_t$ instead of the polyphonic spectrum $X_t$. Similarly, [7] trained timbral features per contour, extracted from harmonic components. The authors used features decribing both static timbre and temporal timbre fluctuations.

#### Extraction of harmonic partials

The harmonic sinusoidal model of [8] can represent the spectral content, generated by a source with given fundamental frequency. The spectral peaks are computed at the location of harmonics $f^n \approx h_n f^0$ , where $h_n$ is the harmonic index. Parabolic interpolation refines the exact frequency locations. We estimated $\bar{X}_t$ with a relatively huge number of harmonics (30), in order to exploit the instability of the singing voice at high harmonics as compared to other instruments.

#### *2.2.1. Vocal Variance*

Singing voice can be distinguished for its relatively frequent timbre fluctuation: When singers vary their vocal tract, the variation of the spectral envelope is more pronounced than for most other instruments [9]. This is especially true when singers articulate actual words. To approximate the spectral shape, MFCCs have been widely utilized, because they relate

well to changes of the vocal tract shape. Features that describe spectral dynamics over time proved to contribute to the solution of the SVD problem [6, 5]. Due to the application of Discre Cosine Transform as a last step of computing MFCCs, the first few coefficients represent the slow variations of the spectrum [**?**]. We utilize the variance of the first 5 MFCCs adopting the frame and hop size of 800 and 200 ms, suggested by [5].

#### *2.2.2. Fluctogram*

The fluctogram is designed to detect sub-semitone fluctuations of the pitch contour and its related harmonic partials [5]. As a preprocessing step the spectrum is converted to a semitone cent scale with resolution one bin per 10 cents. The idea of the fluctogram is to compare the spectrum of a time frame $X_t$ to the subsequent one $X_{t+1}$ by cross-correlation. The fluctuation (in bins) is given by the index of the maximum correlation when $X_{t+1}$ is shifted ±n bins. To track only contours in a local spectral region, the spectrum is divided into 17 2-octave-wide bands and fluctuation is calculated independently for each band. Since the harmonic spectrum $\bar{X}_t$ represents the spectral content of one melodic source, we applied the cross-correlation directly on cent-scale-warped $\bar{X}_t$ for each contour with lowest bin at the contour's inital f0.

### 2.3. Contour classification

We need to discriminate between contours that have as a source singing voice and those with origin from background instruments. We opted for a Random Forest Classifier, because it was shown to outperform a generative classifier [2]. Prior to training, contour labeling is done by thresholding the amount of overlap of the extracted contours $S$ with ground-truth annotation. This is necessary as labeling each contour manually is very laborious task. We have adopted the proposed threshold $\Theta$=0.5.

#### *2.3.1. Vocal Melody Decoding*

After the classification stage, we discard the contours with a likelihood of less than a threshold $\beta$, learned from an evaluation set. Thus time regions with no vocal contours are considered as non-vocal already at this stage. To select one from simultaneous candidate contours, we employ Viterbi decoding using the contour's likelihood score as emission probability, and a transition matrix to encourage continuity in pitch space.

## 3. EXPERIMENTS

We present two experiments. First we train and evaluate our feature set on a dataset with predominant singing voice. We experiment with different parameters of vocal variance and fluctogram, as well as different contour sets S as input. Then in a second experiment, these parameters are kept unchanged for a comparison with other methods on a dataset with more challenging singing material.

## Metrics

We report voicing recall (VR), false alarm (VFA) and F1-measure (VF1). Melody extraction metrics Raw Pitch Accuracy (RPA), Raw Chroma Accuracy (RCA) and Overall Accuracy (OA) are reported for the sake of completeness. All evaluation metrics were computed using mir_eval [10].

### 3.1. Experiment 1: Adding fluctuation features to contour classification

First we ran experiments on the *iKala* dataset that have predominant singing voice, designed for vocal source separation [11]. It has a vocal proportion of around 90% of the audio with total length of 125 minutes.

The complete feature set consists of features_SAL, the median of the vocal variance (VV) and median of the fluctogram variance. Experiments were conducted with $\tau_\sigma = (0.9, 1.1)$, resulting in around 6000 vocal contours.

We varied minimal contour duration, because the vocal variance does not make sense for too short contours. Also, to assure enough MFCC frames per contour, we decreased the MFCCs hopsize of 200 ms, proposed in [5]. Figure 1 lists voicing metrics for different feature parameters of contours.

It can be seen that voicing influences slightly accuracy of the extracted melody. It has to be noted that some errors come from training the non-vocal class on contours from non-main melody, which are essentially vocal with octave errors.

### 3.2. Experiment 2: Comparison with related approaches

For the sake of comparison with previous work on contour classification and voicing, we evaluated our approach on the vocal part of *medleyDB* [**?**]. *MedleyDB* presents a higher diversity of singers and genres, and more realistic data in comparison to *iKala*, because in *MedleyDB* there are full tracks with large unvoiced portions

| feature set | $\tau_\sigma$ | min_duration | VR | VFA | VF1 | RPA | RCA | OA |
|---|---|---|---|---|---|---|---|---|
| features_SAL | 1.1 | 100 | 83.3 | 35.5 | | 72.5 | 77.4 | 70.7 |
| features_SAL | 0.9 | | 44.03 | 0.9 | | 40.8 | 45.7 | 58.0 |
| +VV (300_100_300) | | | | | | | | |
| features_SAL | 0.9 | 200 | 88.9 | 44.7 | 81.2 | 76.5 | 81.4 | 69.8 |
| +VV (300_100_300) | | | 89.0 | 39.6 | 84.4 | 77.1 | 81.7 | 72.4 |
| +VV (300_50_300) | | | 87.9 | 37.6 | 84.1 | 76.6 | 80.8 | 72.7 |
| +fluct | | | 89.2 | 40.5 | 84.2 | 77.3 | 82.0 | 72.0 |
| +fluct+VV (300_100_300) | | | 87.9 | 36.2 | 84.7 | 76.2 | 80.8 | 73.1 |

**Table 1**. Mean metrics from 5 random splits of training and test data on the iKala dataset.

**Table 2**. VR and VFA on iKala as a function of overlap size

## 4. REFERENCES


[1] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[2] Rachel M Bittner, Justin Salamon, Slim Essid, and Juan P Bello, "Melody extraction by contour classification," in *Proc. ISMIR*, pp. 500–506.

[3] J. Bosch, R. M. Bittner, J. Salamon, and E. Gómez, "A comparison of melody extraction methods based on source-filter modelling," in *17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, New York, Aug. 2016.

[4] Martın Rocamora and Perfecto Herrera, "Comparing audio descriptors for singing voice detection in music audio files," in *Brazilian Symposium on Computer Music, 11th. San Pablo, Brazil*, 2007, vol. 26, p. 27.

[5] Bernhard Lehner, Gerhard Widmer, and Sebastian Bock, "A low-latency, real-time-capable singing voice detection method with lstm recurrent neural networks," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 21–25.

[6] Vishweshwara Rao, Chitralekha Gupta, and Preeti Rao, "Context-aware features for singing voice detection in polyphonic music," in *International Workshop on Adaptive Multimedia Retrieval*. Springer, 2011, pp. 43–57.

[7] Vignesh Ishwar, "Pitch estimation of the predominant vocal melody from heterophonic music audio recordings," 2014.

[8] Xavier Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Tech. Rep., 1989.

[9] Johan Sundberg and Thomas D Rossing, "The science of singing voice," *the Journal of the Acoustical Society of America*, vol. 87, no. 1, pp. 462–463, 1990.

[10] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[11] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 718–722.