PREDOMINANT SINGING VOICE DETECTION IN POLYPHONIC MUSIC BASED ON PITCH CONTOUR CHARACTERISTICS

Author(s) Name(s)

Author Affiliation(s)

ABSTRACT

We apply state of the art features for vocal non-vocal dynamic. on extracted pitch contours instead of on frame-level. TODO

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

1. INTRODUCTION

The automatic detection of singing voice (SVD) in audio recordings is an important problem in audio signal processing and required step for other related tasks. When singing is the predominant source, the pitch trajectory can be naturally divided into short time regions, each representing a motif from main melody. A region has coherent characteristics: similar amplitude, pitch and timbre. This fact has been exploited as an intermediate step in an algorithm for predominant melody extraction [1]: Candidate pitch contours are generated by combining spectral bins with high harmonic salience. Each contour is then characterized by features of its pitch trajectory and salience. Based on the heuristics about the distribution of these features, contours of predominant voice are selected.

Recent work proposes to replace that heuristic contour selection by a discriminative classifier fitted on the features [2]. The authors show that a data–driven approach is beneficial over a heuristics one, since it allows to learn the contours characteristics from training data. Despite of improving overall melodic accuracy, the reported vocal false alarm is still relatively high (in the order of 40).

On the other hand, SVD approaches, which do not incorporate vocal melody extraction, usually distinguish voice by a frame-wise classifier trained on spectral features (for example MFCCs, spectral flatness) [3]. Recent a substantial improvement was achieved by including carefully parametrized features, which describe temporal dynamics of vocal timbre [4]. In particular, the authors showed that voicing false alarm rate was reduced.

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 timbral features to the classifier. 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 the features suggested by [4]. We focused on features that capture the fluctuating timbre of singing voice compared to other instruments.

Thanks to XYZ agency for funding.

2. METHOD

A block diagram 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. Additionally, contour timbre features are extracted from the isolated spectral components with origin from the contour's source: vocal or instrumental. Then using both pitch and timbre features a classifier is fit to discriminate the contour's source. Finally a single main vocal melody has to be decoded, because there might be regions, in which candidate contours are overlapping/simultaneous. We assume that singing voice appears only as 'lead instrument' (e.g. excluding backing voice), therefore final vocal regions are determined uniquely by the derived melodic line.

2.1. Contour creation

The contour creation step is adopted from [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 1) threshold factor of the salience of the highest peak in current frame τ_+ and 2) the threshold degree of variation below the mean salience per recording τ_σ. 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)

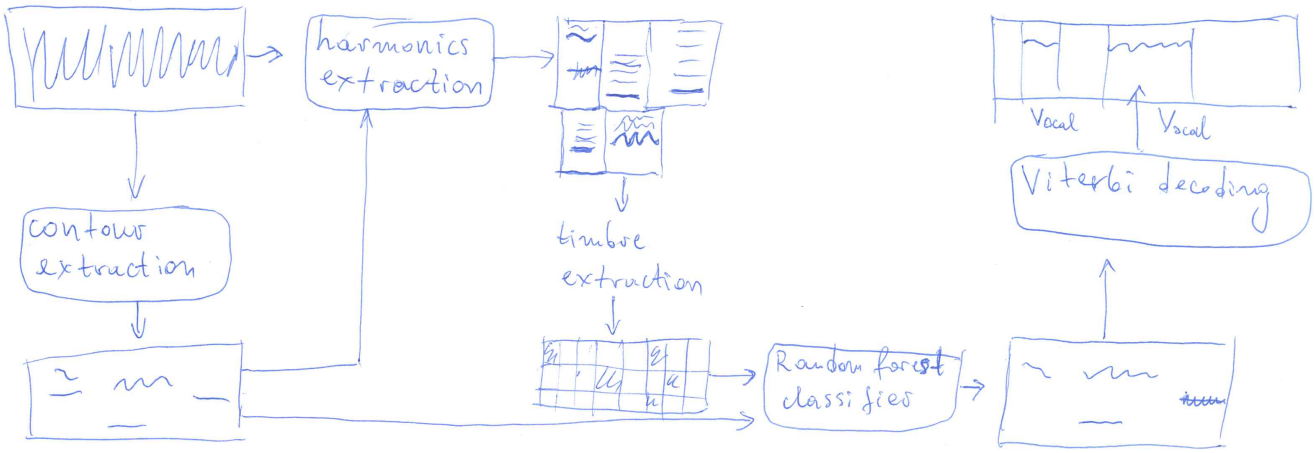

Fig. 1. Block diagram of the proposed system

2.2. Extraction of timbre features

The features_SAL do not describe the timbre of the contour's source. Taking advantage of the pitch curve, for each candidate contour, we extract harmonic partials over time using harmonic sinusoidal modeling (see Figure 1). This yields ideally spectral content of only the contour's source. A contribution of this work is the evaluation of the setting of feature extraction from this pseudo mono-source signal representation. Similarly [5] trained timbral features per contour, extracted from harmonic components. What is our advantage to him ? Maybe cite here Marxer , too, which paper?.

Extraction of harmonic partials

The harmonic sinusoidal model of [Serra] filters the spectral components corresponding to a source with fundamental frequency. Parabolic interpolation is used for locating the peaks corresponding to the the harmonics $f^n \approx h_n f^0$, where h_n is the harmonic index. This gives a more accurate estimate of the peak locations and amplitudes. We estimate the first 30 harmonics in order to exploit the instability of the vocal pitch at higher harmonics as compared to other instruments.

2.2.1. *Vocal Variance*

Singing voice can be distinguished for its continuous timbre: When singers articulate actual words, the variation of the spectral envelope is more pronounced than for most other instruments [ref]. To approximate the spectral shape, MFCCs have been widely utilized, because they relate well to the changes of the vocal tract shape. Features that describe spectral dynamics over time proved to contribute to the SVD problem [6, 4]. Due to the application of DCT as a last step of computing MFCCs, the first few coefficients represent the slow variations of spectrum [Rabiner and Huang]. We utilize the variance of the first 5 MFCCs adopting the frame and hop size of 800 and 200 ms, suggested by [4].

2.2.2. *Fluctogram*

—

2.3. Contour classification

We opted for a Random Forest Classifier, because it was shown to outperform a generative classifier [2] and has been applied also in [4]. Prior to training, contour labeling is done by thresholding the amount of overlap of the extracted contours S with ground-truth annotation. should I write here formula of OA? This is necessary as labeling each contour manually is very laborious task. We have though increased the proposed threshold Θ because we want contours from background instruments to be included as positive examples, whic is important for reducing false alarm rate.

feature set	VR	VFA	RPA	RCA	OA	
features_SAL	0.92	0.42	0.79	0.84	0.73	
+VV						

Fig. 2. Mean metrics from 5 random splits of training and test data on the iKala dataset.

Fig. 3. VR and VFA on iKala as a function of overlap size

2.3.1. *Vocal Melody Decoding*

A threshold β on the probability of the discriminative model is learned from an evaluation set. Thresholding is necessary to discard regions of time with no predominant nature. Thus regions with no predominant pitch are considered as non-vocal already at this stage. The final melody line decoding is done by Viterbi decoding. This way from simultaneous candidate contours only one is preferred at a given point in time. [DISCARD some detailes hrere. I have the feeling I am repeating too much of Bittner.]

3. EXPERIMENTS

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

We report voicing recall, and false alarm rate. Melody extraction metrics are reported for the sake of completeness. All evaluation metrics were computed using mir_eval [**?**].

3.1. Adding timbre features to contour classification

iKala ... proportion of vocal/non-vocal.

The complete feature set consists of features_SAL and the median of the vocal variance(VV) and fluctogram. A list of feature importances is shown in Figure...

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

3.2. 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 []. describe medleyDB: Unlike iKala, it has more more diverse singers and diverse accompaniment.

Method	VR	VFA	RPA	RCA	OA	
BIT						
LEH						
TC						

Table 1. Comparison of the proposed timbral classification TC with BIT[2] and LEH [4] on the vocal subset of MedleyDB

If contradictory frames are assigned a contour, their voicing is decided in the context of the whole contour. This is a more informed decision because a vocal contour has coherent salience.

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] 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.

[4] 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)*. IEEE, 2014, pp. 7480–7484.

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

[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.