SINGING VOICE DETECTION BY TIMBRE-BASED PITCH CONTOUR CLASSIFICATION

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 polyphonic 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 segments, each representing a motif from main melody. One melodic segment usually has pitch and intensity that change smoothly over time. A well-performing predominant melody extraction method [1] exploits the notion of such segments 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 frmo predominant voice are selected, based on heuristics about the distribution of these features.

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 MedleyDB [ref], but Bosch et al. [] reported improvements on symphonic music, since the heuristic rules from [salamon] 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). Furthermore, the timbre of the voice in a contour changes smoothly and is thus relatively coherent. In both [] and [] this fact is not 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) [3]. Recent classification approach was based on the observation that temporal fluctuations of singing voice timbre discriminate it from other instruments [4]. The authors did a careful selection of the parameters of both conventional static features and fluctuation ones. In particular, a 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 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 introduced by [4]. .

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. Then contour timbre features are extracted from the isolated spectral components with origin from the contour's source: vocal or instrumental. Next using both pitch and timbre features a classifier is trained to discriminate the contour's source. Finally a single vocal melody line has to be decoded, because there might be regions, in which candidate contours are overlapping. 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)

Thanks to XYZ agency for funding.

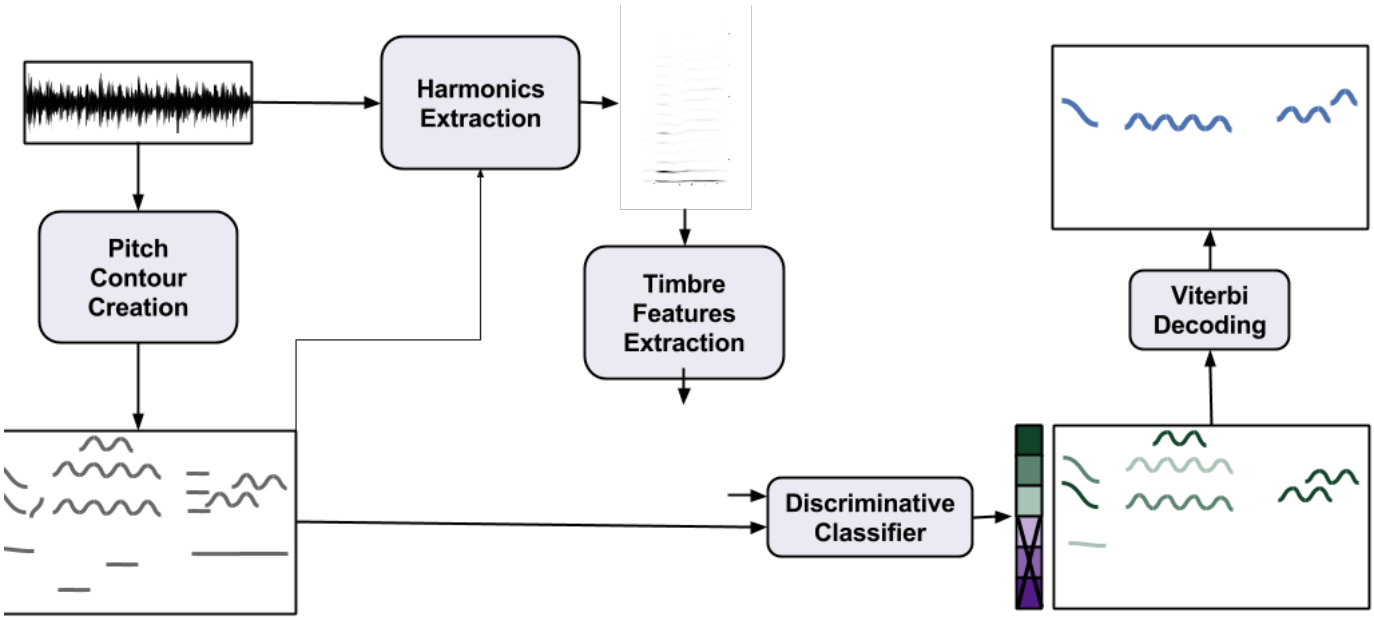

Fig. 1. Block diagram of the proposed system

- Salience (mean, st dev and sum)
- Vibrato (presence, rate and extent)

2.2. Extraction of timbre features

The features_SAL do not describe the timbre of the contour's source. 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 [5]. 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 spectral envelope instead of the polyphonic spectrum. Similarly, [6] trained timbral features per contour, extracted from harmonic components. Maybe cite here Marxer , too, which paper?.

Extraction of harmonic partials

The harmonic sinusoidal model of [Serra] selects the spectral content, generated by a source with 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 estimate 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 [ref]. 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 SVD problem [5, 4]. Due to the application of Discre Cosine Transform 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 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. should I write here formula of OA? This is necessary as labeling each contour manually is very labori-

ous task. We have though increased the proposed threshold Θ because we want contours from background instruments to be included as positive examples, which is important for reducing false alarm rate.

2.3.1. *Vocal Melody Decoding*

After the classification stage, we discard the contours with a likelihood of less than a threshold β, 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 relatively predominant singing voice. We experiment with different parameters of vocal variance and fluctogram, as well as different contour sets S as input. Then these parameters are kept unchanged for a comparison with other methods on a dataset with more challenging singing material in the second experiment.

Metrics

We report voicing recall, false alarm and F1-measure. Melody extraction metrics are reported for the sake of completeness. All evaluation metrics were computed using mir_eval [**?**].

3.1. Experiment 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) .

Changing the contour salience parameters $\tau_+ = 0.9$ and $\tau_\sigma = 0.9$, used in [2], can increase the number of input contorus, e.g. recall of baseline. Setting $\tau_+ = 0.7$ and $\tau_\sigma = 1.3$ resulted in a huge amount of short non-vocal contours and vocal contours being underrepresented. 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 [4]. Figure **??** lists voicing metrics for different feature parameters of contours with

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.

feature set	min_duration	VR	VFA	V F-1
features_SAL 1.3 0.7	100	92.	42.	
features_SAL	200	88.9	44.7	70.1
+VV (300_100_300)		87.4	35.9	72.1
+VV (300_50_300)		87.9	37.6	70.9
+fluctogram (200_20_200)		86.9	31.8	74.1
+fluctogram (300_20_300)		86.8	32.8	73.1

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

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, since there are many full tracks with large unvoiced portions.

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] 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.

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