

SINGING VOICE DETECTION IN MUSIC TRACKS USING DIRECT VOICE VIBRATO DETECTION

L. Regnier, G. Peeters

IRCAM

Sound Analysis/ Synthesis team, CNRS-STMS
1 place Stravinsky, 75004 Paris

ABSTRACT

In this paper we investigate the problem of locating singing voice in music tracks. As opposed to most existing methods for this task, we rely on the extraction of the characteristics specific to singing voice. In our approach we suppose that the singing voice is characterized by harmonicity, formants, vibrato and tremolo. In the present study we deal only with the vibrato and tremolo characteristics. For this, we first extract sinusoidal partials from the musical audio signal. The frequency modulation (vibrato) and amplitude modulation (tremolo) of each partial are then studied to determine if the partial corresponds to singing voice and hence the corresponding segment is supposed to contain singing voice. For this we estimate for each partial the rate (frequency of the modulations) and the extent (amplitude of modulation) of both vibrato and tremolo. A partial selection is then operated based on these values. A second criteria based on harmonicity is also introduced. Based on this, each segment can be labelled as singing or non-singing. Post-processing of the segmentation is then applied in order to remove short-duration segments. The proposed method is then evaluated on a large manually annotated test-set. The results of this evaluation are compared to the one obtained with a usual machine learning approach (MFCC and SFM modeling with GMM). The proposed method achieves very close results to the machine learning approach : 76.8% compared to 77.4% F-measure (frame classification). This result is very promising, since both approaches are orthogonal and can then be combined.

Index Terms— Singing voice detection, vibrato detection, voice segmentation, vibrato and tremolo parameters extraction, feature extraction.

1. INTRODUCTION

Singing voice melody is the most representative and memorable element of a song. Locating vocal segments in a track is the front-end of many applications including singer identification, singing voice separation, query-by-lyrics, query-by-humming and extraction of musical structure.

The problem of singing voice detection can be stated as follows : segment a song into vocal and non-vocal (i.e pure

instrumental or silent) parts. The main difficulty comes from the presence of musical instrument during most vocal segments.

We review here existing methods. Usual systems for partitioning a song into vocal and non-vocal segments start by extracting a set of audio features from the audio signal and then use them to classify frames using a threshold method or a statistical classifier. The result of this classification is then used to segment the track.

Features : According to [1], among the various features MFCCs (Mel Frequency Cepstral Coefficients) and their derivatives are the most appropriated features to detect singing voice ([2], [3]). Other features describing the spectral envelope are also used : LPC and their perceptual variants PLP [4] [5], Warped Linear Coefficients [6]. These features, traditionally used in speech processing are however not able to describe characteristics of the singing voice in presence of instrumental background. Energy [7], harmonic coefficient [8], perceptually motivated acoustic features [9] such as attack-decay, formants, singing-formant and vibrato have also been investigated.

Classifier : Different statistical classifiers such as GMM, HMM, ANN, MLP or SVM have been proposed to discriminate the classes using the various sets of features listed above.

Post-processing : Frame-based classifiers tend to provide noisy classification leading to an over-segmentation (short segments). In the opposite, human annotation tends to provide under-segmentation (long segments ignoring instrumental breaks or singer breathing). For these reasons, post processing is usually applied to the estimated segmentation. This post-processing can be achieved using median filtering or an HMM trained on segment duration as suggested in [10]. [3] introduces a method for deriving decision function and smoothing with an autoregressive moving average filter.

The structure of the paper is organized as follows. First, the characteristics of the singing voice are introduced in section 2. Section 3 provides an overview of the system used. Results obtained with this method are presented and compared with the ones obtained with a machine learning method in section 5. Section 6 concludes the paper.

2. CHARACTERISTICS OF SINGING VOICE

We introduce here the most representative characteristics of the singing voice.

2.1. Formants

Formants are the meaningful frequency components of speech and singing. Formants frequencies are determined by the shape of vocal tract. These frequencies (usually 3) allow us to identify vowels. A fourth formants called singing formant exists for singer and it ranges from 2kHz to 3kHz. Singing formant helps lyrics singer to stand above instrumental accompaniment. However, singing formant does not exist in many other type of singing.

2.2. Harmonicity

One of the most discriminative element to distingue singing voice from speech is harmonicity [8]. A sound is harmonic when its partials are located at multiples of the fundamental frequency. Singing voice is naturally harmonic due to resonances of the vocal tract. In addition, the number of harmonic partials can increase with singing technic.

2.3. Vibrato and tremolo

One of the specificity of the singing voice is its natural vibrato. Vibrato is a musical effect used to add expression. Nowadays vibrato is used by most musical instruments to add vocal-like qualities to instrumental music.

The term ‘‘vibrato’’ is sometimes used inappropriately to qualify frequency modulation (or pitch vibrato) and amplitude modulation (or intensity vibrato). Vibrato only refers to the periodic variation of pitch, whereas tremolo is the term to be used for the periodic variation of intensity.

Only a few musical instruments can produce simultaneously both types of modulation. In wind and brass instruments the amplitude modulation is predominant. In string instruments the frequency modulation is predominant. In voice both modulations occur at the same time. This is due to the mechanical aspects of the voice production system [11].

Vibrato and tremolo can be described each by two elements : their frequencies (or vibrato/tremolo rate) and their amplitudes (vibrato/tremolo extent). For the voice, the average rate is around 6 Hz and increases exponentially over the duration of a note event [12]. The average extent ranges from 0.6 to 2 semitones for singers and from 0.2 to 0.35 semitones for string players [13].

We exploit this particularities (average rate, average extent and presence of both modulations) to discriminate voice between all musical instruments. We suppose that a partial corresponds to a singing sound if the extent values of its vibrato and tremolo are greater than thresholds.

3. PROPOSED METHOD

We present here a method to detect voice using only vibrato and tremolo parameters. Frame-analysis is first applied using a 40 ms window length with 20 ms hop size. At each time t , sinusoidal components are extracted. Each component s is characterized by its frequency $f_s(t)$, amplitude $a_s(t)$ and phase $\phi_s(t)$. These values are then interpolated over time to create sinusoidal tracks or partials $p_k(t)$ [14].

3.1. Selection of partials with vibrato

We consider a partial as vibrated if the extent around 6Hz, Δf_{p_k} is greater than a threshold (τ_{vib}). Thus, to determine the set of partial with vibrato P_{vib} we extract Δf_{p_k} for each partial p_k .

For each partial $p_k(t)$ the frequency values $f_{p_k}(t)$ are analyzed. The average frequency of $p_k(t)$ is denoted by $\mu_{f_{p_k}}$. Let us denote the Fourier transform of $f_{p_k}(t)$ by $F_{p_k}(\omega)$. The extent of the frequency modulation of a partial being proportional to its center frequency, we use relative values.

Thus, for a partial existing from time t_i to t_j the Fourier transform is given by :

$$F_{p_k}(f) = \sum_{t=t_i}^{t_j} (f_{p_k}(t) - \mu_{f_{p_k}}) e^{-2i\pi f \frac{t}{L}}$$

Where $L = t_j - t_i$.

The extent value in Hz is given by :

$$\Delta f_{p_k}(f) = \frac{F_{p_k}(f)}{L}$$

Its relative value is given by :

$$\Delta f_{rel p_k}(f) = \frac{\Delta f_{p_k}(f)}{\mu_{p_k}}$$

Finally, we determine the relative extent value around 6 Hz as follow :

$$\Delta f_{p_k} = \max_{f \in [4,8]} \Delta f_{rel p_k}(f).$$

To convert this value in cent (1 ton = 100 cents) we use the following formula :

$$\Delta_{cent} f_{p_k} = 1200 * \log 2(\Delta f_{p_k} + 1).$$

Thus we have :

$$p_k \in P_{vib} \Leftrightarrow \Delta f_{p_k} > \tau_{vib}$$

3.2. Selection of partial with tremolo

To determine the set of partial with tremolo (P_{trem}) we extract the extent value of tremolo using the values of amplitude a_{p_k} instead of f_{p_k} . We note Δa_{p_k} the extent relative value and P_{trem} the set of partials with tremolo. We have

$$p_k \in P_{trem} \Leftrightarrow \Delta a_{p_k} > \tau_{trem}.$$

3.3. Selection of singing partials

Let P_{voice} be the set of partials belonging to the voice. We introduce a second criteria based on the number of partials present at a time t .

The selection is made as follow : Let p_k be a partial with a vibrato and a tremolo existing between time t_i and t_j :

$$p_k \in P_{vib} \cap P_{trem} | p_k(t) \neq 0, \forall t \in [t_i, t_j].$$

Then p_k belongs to the voice if it exists another partial between t_i and t_j which also has a vibrato and a tremolo.

$$p_k \in P_{voice} \Leftrightarrow \exists p_l \in P_{vib} \cap P_{trem}, \exists t \in [t_i, t_j] | p_l(t) \neq 0$$

3.4. Post-processing

To be closer to human segmentation we decide to eliminate non-vocal segments with a duration lower than 1 second. Thus a short non-vocal segment located between two vocals segments will be classify as vocal as show on figure 1 .

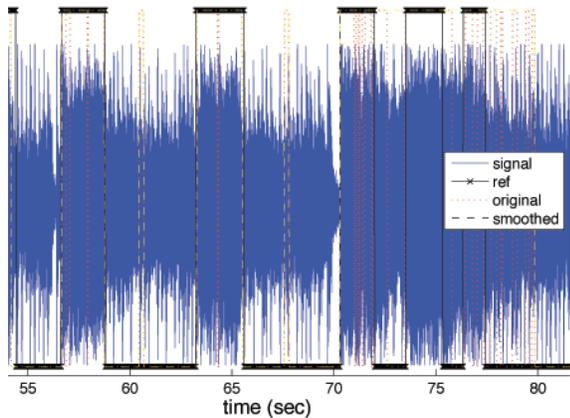


Fig. 1. Segmentation into singing/non singing part before and after post-processing on signal : “aline.wav”

4. EXPERIMENTAL RESULTS

4.1. Test set

The song database used to carry out this study is the one used in [10]. This database is composed of 90 songs selected for their variety in artists, languages, tempi and music genre. They constitute a representative sampling of commercial music. The sampling frequency of the songs is 44.1 kHz, stereo channel and 16 bits per sample.

Each file was annotated manually into singing and non-singing sections by the same person to provide the ground truth-data. Establishing where a singing segment starts and ends with certainty is problematic. The segmentation was done as a human would expect i.e small instrumental breaks of short duration were not labeled. The database is split into

training (58 songs) and test (32 songs) sets. The whole set is well balanced since 50.3% of frames are of singing segments and 49.7% of non singing segments. We note that all files from the database contains singing and no one of them is a Cappella music.

Note that we are assuming that the signal is known to consist only of music and that the problem is locating singing within it. We are not concerned with the problem of distinguishing between vocal and regular speech, nor music and speech.

4.2. Results

For this task we only consider the results given for the class “singing segment”. The classification accuracy is calculated on all the frames of the test set with the F-measure. Figure 2 shows the precision and recall obtained when one threshold is fixed and the second one vary. For both thresholds,

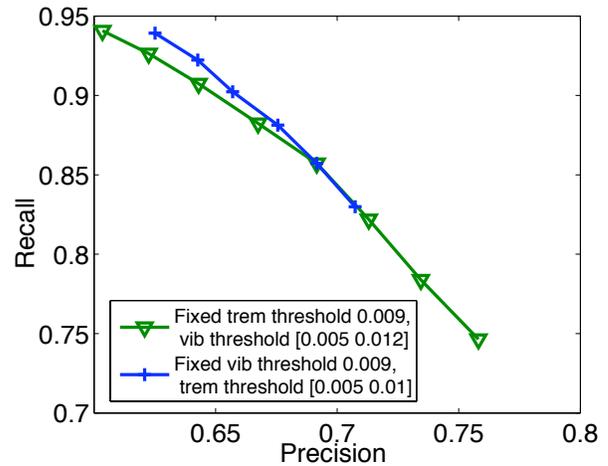


Fig. 2. P.R curves of singing voice identification using threshold method

increasing a given threshold increase precision while decreasing recall. Thus we can chose the parameters depending on the task. For singer identification it is necessary to get a high precision so high thresholds would fit. In general we would choose threshold to obtain a good precision and a good recall at the same time.

The thresholds τ_{vib} and τ_{trem} are learnt on the training set. We obtain the best classification with both thresholds equal to 0.009 which is equivalent to 15.5 cents vibrato and tremolo extent. These results are shown on table 1 .

	Before filtering	After Filtering
Fmeasure	66.54%	76.83%
Recall	61.69%	83.57%
Precision	70.56%	71.09%

Table 1. F-measure for $\tau_{vib} = \tau_{trem} = 0.009$

We compare this results with the ones obtained with a

learning machine approach. The features chosen are : MFCC, Δ MFCC, $\Delta\Delta$ MFCC , SFM (Spectral Flatness Measure), Δ SFM and $\Delta\Delta$ SFM. These feature are normalized by their inter-quantile-range (IQR). Feature with IQR equal to zero are deleted. The best 40 features are selected using a feature selection algorithm, in our case : Inertia Ratio Maximization with Feature Space Projection (IRMFSP) [15]. Then, a linear discriminant analysis is applied on each class. Finally, each class is modeled with a height-component GMM. The results are given in table 2 . Results of both methods are very close.

	Before filtering	After Filtering
Fmeasure	75.81%	77.4%
Recall	75.8%	77.5%
Precision	75.80%	77.4%

Table 2. F-measure : Features MFCC, Classifier GMM

We notice that the proposed method gives a better recall whereas the learning approach is more precise.

5. CONCLUSION

Conventional singing detection methods ignore the characteristics of singing signal. We presented here a method to detect vocal segments within an audio track based only on vibrato and tremolo parameters of partials. Using this, a partial is said to belong to singing sound or not using a simple threshold method. For a large test set the best classification achieved was 76.8% (F-measure). The threshold method has been compared to a machine learning approach using MFCCs and their derivatives as features and a GMM as statistical classifier leading to 77.4% F-measure. It is surprising that our simple approach leads to very close results to the more sophisticated machine-learning approach. Since both approaches are orthogonal and since they both lead to good results, it could be interesting to combine them. Future work will therefore concentrate on replacing the threshold method by a machine learning approach which could also model the spectral shape representation provided by the MFCCs.

6. ACKNOWLEDGEMENT

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. We thank M. Ramona for sharing his annotated audio collection. We also thank our colleague D. Tardieu who contributed to the contents and approach of this paper.

7. REFERENCES

- [1] M. Rocamora and P. Herrera, "Comparing audio descriptors for singing voice detection in music audio files," *SBCM - Brazilian Symposium on Computer Music 07*, 2007.
- [2] W.H. Tsai, H.M. Wang, and D. Rodgers, "Automatic singer identification of popular music recordings via estimation and modeling of solo vocal signal," in *Eighth European Conference on Speech Communication and Technology*. ISCA, 2003.
- [3] H. Lukashevich, M. Gruhne, and C. Dittmar, "Effective singing voice detection in popular music using arma filtering," *Workshop on Digital Audio Effects (DAFx'07)*, 2007.
- [4] AL Berenzweig and DPW Ellis, "Locating singing voice segments within music signals," *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp. 119–122, 2001.
- [5] A. Berenzweig, D.P.W. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," *AES 22nd International Conference*, 2002.
- [6] Youngmoo E. Kim and Brian Whitman, "Singer identification in popular music recordings using voice coding features," in *ISMIR*, 2002.
- [7] T. Zhang, "Automatic singer identification," *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1, 2003.
- [8] W. Chou and L. Gu, "Robust singing detection in speech/music discriminator design," *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2, 2001.
- [9] Tin Lay Nwe and Haizhou Li, "Singing voice detection using perceptually-motivated features.," in *ACM Multimedia*, Rainer Lienhart, Anand R. Prasad, Alan Hanjalic, Sunghyun Choi, Brian P. Bailey, and Nicu Sebe, Eds. 2007, pp. 309–312, ACM.
- [10] M. Ramona and B. Richard, G. David, "Vocal detection in music with support vector machine," *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'08). IEEE International Conference on*, pp. 1885 – 1888, March 2008.
- [11] V. Verfaillie, C. Guastavino, and P. Depalle, "Perceptual evaluation of vibrato models," *Proceedings of Conference on Interdisciplinary Musicology*, 2005.
- [12] E. Prame, "Measurements of the vibrato rate of ten singers," *The Journal of the Acoustical Society of America*, vol. 96, pp. 1979, 1994.
- [13] R. Timmers and P. Desain, "Vibrato : Questions and answers from musicians and science," *Proc. ICMPC*, 2000.
- [14] Axel Roebel, "Frequency-slope estimation and its application to parameter estimation for non-stationary sinusoids," *Computer Music Journal*, 2008.
- [15] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization," *115th AES convention, New York, USA, October*, 2003.