VIBRATO-MOTIVATED ACOUSTIC FEATURES FOR SINGER IDENTIFICATION

Haizhou LI and Tin Lay NWE

Institute for Infocomm Research, Republic of Singapore
hli@i2r.a-star.edu.sg, tlnma@i2r.a-star.edu.sg

ABSTRACT

It is common that a singer develops a vibrato to personalize his/her singing style. In this paper, we explore the acoustic features that reflect vibrato information, to identify singers of popular music. We start with an enhanced vocal detection method that allows us to select vocal segments with high confidence. From the selected vocal segments, the cepstral coefficients which reflect the vibrato characteristics are computed. These coefficients are derived using cascaded bandpass filters spread according to the octave frequency scale. We employ the high level musical knowledge of song structure in singer modeling. Singer identification is validated on a database containing 84 popular songs in commercially available CD records from 12 singers. We achieve an average error rate of 16.2% in segment level identification.

1. INTRODUCTION

Singer identification (SingerID) is an important task in the area of music information retrieval. A singer identification system has three fundamental components: 1) A vocal detector that separates singing voice (vocal) from instrumental sound (non-vocal) in a song. 2) A feature extraction front-end that obtains singer feature parameters from the vocal segments. 3) A statistical classifier that identifies the singer given an unknown vocal segment.

The topic of vocal detection has recently been given much attention. One important effort recently made use of music semantics in addition to acoustic characteristics [1, 2]. In [1], song structure information and song specific vocal characteristics are integrated in vocal/non-vocal modelling. In [2], octave frequency scale is used to characterize the harmonic structure of vocal and non-vocal segments.

A large number of features have been explored for representing audio signals for SingerID. These include Mel Frequency Cepstral Coefficients (MFCC) [3], Linear Prediction Coefficients (LPC) [4], robust estimates of spectral envelopes [5] and octave frequency scale based cepstral coefficients [6]. The study in [7] uses Composite Transfer Function (CTF) which is derived from the instantaneous amplitude and frequency of the partials associated to vocal vibrato to identify singer. Mellody et al. [8] stated that temporal patterns in the vocal passage such as vibrato and note transitions are likely cues to singer identity and vocal quality. Sundberg [9] stated that vibrato develops more or less itself without thinking about it and without trying actively to acquire it. Vibrato rate is generally considered to be a constant within a singer. However, some singers are able to deliberately change their vibrato rate. Eric [10] found that individual singer's interpretation of a piece of music may affect the vibrato rate. In general, all these studies suggest that vibrato is a function of the style of singing of a particular singer, which can be controlled consciously or unconsciously. Inspired by these studies, in this paper, we investigate the effect of directly integrating vibrato information into acoustic features on SingerID performance. In [7], Wakefield derived Composite Transfer Function (CTF) based on vibrato information. We propose using several types of subband filters to integrate vibrato into acoustic features. The difference between the two studies lies in the ways we explore the vibrato characteristics.

The method in [6] uses octave frequency scale to derive the cepstral coefficients to identify singers. Apart from harmonic analysis, we find that this octaves frequency scale can be used to investigate the singer's vibrato characteristics as well.

Many common pattern classifiers such as HMM, neural networks (NN) and support vector machines (SVM) have been proposed for SingerID [4,3]. In those studies, a statistical model for each of the singers was created only using acoustic cues. Note that a popular song is composed according to a common structure that comprises of intro, verse, chorus, bridge and outro sections [2]. Each section of a song has its signal characteristics, which is uniquely interpreted by a singing voice and its music accompaniment [11]. We propose building multiple HMMs, also called Multi-Model HMM or MM-HMM, for a singer based on the structure of the songs, with each describing the singer's vibrato characteristics present in one section of the songs.

Our approach consists of three steps. First, the test song is segmented and classified into vocal and non-vocal segments. Then, we compute the confidence scores of the classification decision to confirm the vocal detection with a hypothesis test. Finally, we extract singer features from the vocal frames and evaluate them with the MM-HMM singer model.

The rest of the paper is organized as follows. In section II, we propose a method for vocal/non-vocal detection. In section III, we discuss in details the procedures for singer feature extraction from an audio signal. In section IV, we propose singer modelling based on sections of song structure. In section V, we present the experiment set-up and results. Finally, we conclude our study in section VI.

2. VOCAL DETECTION AND VERIFICATION

The vocal detector detects the presence of vocals in the musical signal. We use subband based Log Frequency Power Coefficients (LFPC) [12] to form our feature vectors. We employ Multi-Model HMM (MM-HMM) classifier method and bootstrapping process [1] for vocal/non-vocal modelling and classification,

Vocal detector has more false alarm than false rejection. To exclude the misclassified vocal segments from the SingerID experiments, we perform hypothesis test [13] to measure the

confidence of the classification decision made by vocal detector. Then, the given test segment is rejected if the confidence measure is below a predetermined threshold. Otherwise, it is accepted to include in the SingerID experiments.

3. ACOUSTIC FEATURES

We propose to use octave frequency scale based subband filters to characterise vibrato of a singer. The output amplitudes of these filters contain the information of the vibrato. We then transform these subband energies into cepstral coefficients for statistical modelling.

3.1. Vibrato

Vibrato is a periodic, rather sinusoidal, modulation of pitch and amplitude of a musical tone [14]. The vibrato adds a special quality to the tone [9]. Figures 1(a), (b) and (c) show examples of vibrato excursions occurring at the tone D6 for three different singers. In Figures 1 (b) and (c), the vibrato excursions to the up and down of a note are not balanced compared to Figure 1(a). According to [15], such irregular vibrato excursions are very common in most of the tones. In addition, the figures show that vibrato extent differs from singer to singer. Some singers display a wider pitch fluctuation and a slower rate of vibrato which is referred to as 'wobble' [16]. Singers producing a vibrato with a narrower pitch modulation at a faster rate may be considered to have a 'bleat' [17]. In summary, three different characteristics of vibrato waveform should be integrated into the acoustic feature formulation. These are 1) regularity or irregularity in vibrato excursion, 2) two different vibrato types of 'wobble' and 'bleat', and 3) vibrato rate.

As a result of the modulation of pitch, the frequencies of all the overtones vary regularly and in sync with the pitch frequency modulation [9]. Therefore, we implement the subband filters with the center frequencies located at each of the musical notes to characterize the vibrato. The list of the frequencies of the musical notes can be found in [18]. Due to the fact that singing voice contains high frequency harmonics [4], our subband filters span up to 7 octaves (16kHz). According to [19] and [9], minimum range of the vibrato extent is ± 0.5 semitone and maximum range of the vibrato is ± 1 semitone. Our subband filters are implemented with a bandwidth of ± 1.5 semitone from each note since vibrato extent can increase more than ± 1 semitone when a singer raises his/her vocal loudness [9]. We employ three different types of subband filters: triangular, parabolic and cascaded subband filters, to capture vibrato information from acoustic signal.

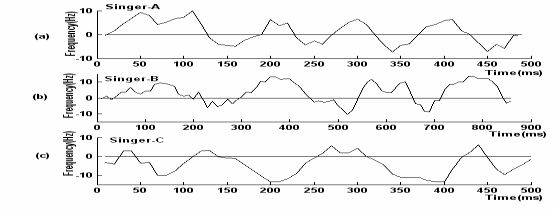

Figure 1. Vibrato waveforms of 3 singers at note D6, 1174.6Hz .

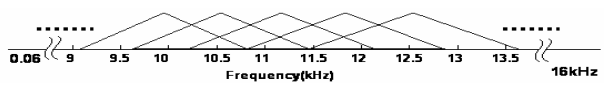

Figure 2. A bank of triangular bandpass filters

3.2. Triangular subband filters

The triangular bandpass filters and subband outputs are shown in Figures 2 and 3 respectively. These filters are symmetric and center at the musical notes. The more the vibrato fluctuation occurs, the less the filter outputs. The output amplitude is maximum when there is no fluctuation. These filters are able to capture whether the vibrato fluctuation is wide (wobble) or narrow (bleat). However, these filters cannot judge whether the vibrato fluctuates left or right because of their symmetrical shape. In order to capture the regularity or irregularity in vibrato excursion, the filter output has to reflect the information of left or right vibrato fluctuation. To account for the left/right direction of vibrato fluctuations, we propose parabolic subband filters in the following section.

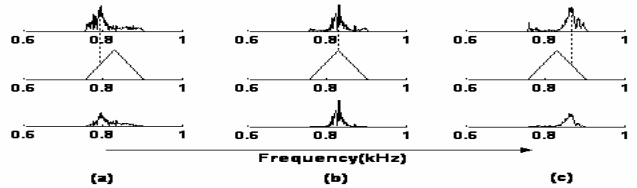

Figure 3. Vibrato fluctuations and triangular bandpass filtering observed at the note G#5, 830.6Hz. (a) vibrato fluctuates left (b) no fluctuation (c) vibrato fluctuates right. The upper panel shows spectrum partial. The middle panel presents the frequency response of the tri triangular bandpass filters. The lower panel demonstrates the output of the triangular bandpass filters.

3.3. Parabolic subband filters

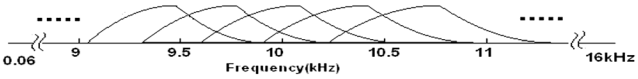

Figure 4. A bank of parabolic bandpass filters

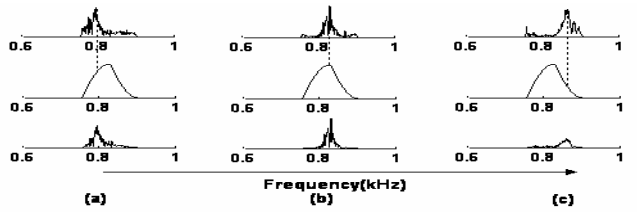

Figure 5. Vibrato fluctuations and parabolic bandpass filtering observed at the note G#5, 830.6Hz. (a) vibrato fluctuates left (b) no fluctuation (c) vibrato fluctuates right. The upper panel shows spectrum partial. The middle panel presents the frequency response of the parabolic bandpass filters. The lower panel demonstrates the output of the parabolic bandpass filters. The parabolic bandpass filters reflect the direction of vibrato fluctuations in the lower panel by the contrast of output amplitude level.

A bank of parabolic subband filters and subband outputs are shown in Figures 4 and 5 respectively. Asymmetric shape of the parabolic subband filters allows us to detect the direction of the vibrato fluctuation and hence informs us of the regularity or irregularity in vibrato excursion. As in triangular subband filters, these filters also capture 'wobble' or 'bleat' vibrato types.

The two filters: triangular and parabolic subband filters compare

the amplitude variations at the filter output to capture the vibrato by assuming amplitudes of partials are constant. However, in vocal vibrato, there is frequency as well as amplitude variation of the partials [7]. Thus, the amplitude variations observed at the output of these filters will reflect variations associated to both of the frequency and amplitude of the partials. To account only for the frequency variations, we proposed the cascaded subband filters which will be discussed in the following section.

3.4. Cascaded subband filters

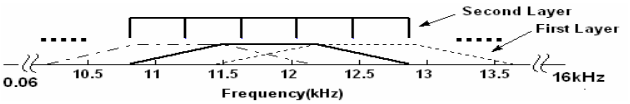

Figure 6. A bank of cascaded subband filters

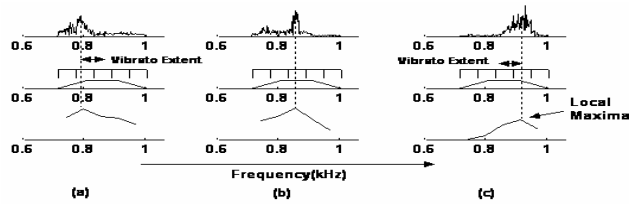

Figure 7. Vibrato fluctuations and cascaded bandpass filtering observed at the note G#5, 830.6Hz. (a) vibrato fluctuates left (b) no fluctuation (c) vibrato fluctuates right. The upper panel shows spectrum partial. The middle panel presents the frequency response of the cascaded bandpass filters. The lower panel demonstrates the instantaneous amplitude output of the cascaded bandpass filters. With the output from the cascaded bandpass filters, we are able to track the local maxima to derive the vibrato extent.

The proposed cascaded subband filters and subband outputs are shown in Figures 6 and 7 respectively. The filter has two cascaded layers of subbands. The first layer has overlapped trapezoidal filters. The second layer has 5 non-overlapped rectangular filters of equal bandwidths for each trapezoidal subband. Trapezoidal filters are tapered between ± 0.5 semitone to ± 1.5 semitone. The vibrato fluctuations are observed by tracking the local maxima in the instantaneous amplitude output of the subbands in the second layer as shown in the lower panel of Figure 7. Local maxima tell the position of the partial. The distance between the center frequency of the corresponding filter and the note informs the vibrato extent. The tapered and overlapped trapezoidal filters in the first layer allow vibrato fluctuations of adjacent notes observed at the output of the subbands in the second layer to be 'continuous'.

The advantage of cascaded subband filter over the triangular and parabolic filters is this filter is robust to amplitude variations associated with vocal vibrato. As a result, the cascaded subband filters are expected to capture irregularities or regularities in vibrato excursion and 'wobble' or 'bleat' vibrato types more effectively than triangular or parabolic subband filters.

3.5. Cepstral coefficient computation

A music signal is divided into frames of 20ms with 13ms overlapping. Each frame is multiplied by a Hamming window. Then, the audio frame is passed through a bank of 96 triangular subband filters spaced in octave scale from 65Hz to 16kHz and the log energy of each band is calculated. Finally, a total of 9 Octave Frequency Cepstral Coefficients using triangular subband filters ($OFCC_{TRI}$) are computed from log energies using Discrete Cosine Transform [20] for each audio frame. To integrate the information of vibrato rate, we augment the 9 coefficients with time derivatives or delta parameters from the $OFCC_{TRI}$ of two neighboring frames. We then replaced the triangular subband filters with parabolic and cascaded subband filters to compute the $OFCC_{PRA}$ and $OFCC_{CASCADE}$ coefficients respectively.

4. STATISTICAL SINGER MODELING

Most SingerID efforts have been devoted to different statistical classifiers [4, 3]. However, to our knowledge, none of the studies has taken the song structure information into account in the modelling. According to [1], vocal and non-vocal segments display variation of signal characteristics based on the structure of the song. Hence, we employ Multi-Model HMM (MM-HMM) classifier formulation method [1] for singer modelling. To capture the intrinsic signal characteristics, we propose singer modelling, according to the type of the section of a song, with three different HMM models $\Lambda_s = \{\lambda_s^V, \lambda_s^I, \lambda_s^{CBO}\}$. Figure 8 shows the general pop song structure and creation of HMM models based on the structure of the song.

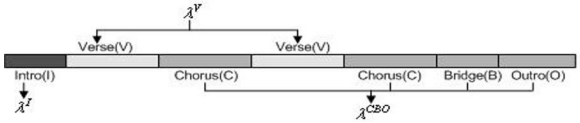

Figure 8. Creation of three HMM models for each singer

5. EXPERIMENTS

Our experimental database consists of 7 popular solo songs from each of 12 singers. Solo songs are sung by only one singer with musical accompaniment. A total of 84 songs are included in the database. Four songs of each singer are allocated to the training set and the remaining 3 songs of each singer to the test set. We obtain a training set of 48 songs and a test set of 36 songs.

Table 1: Average error rates of vocal/ non-vocal detection on test database (N= Non-vocal, V=Vocal, Avg=Average)

Error Rate (%)	N	V	Avg
MM-HMM	27.2	7.5	17.3
MM-HMM + Hypothesis test	7.7	9.7	8.7

5.1. Vocal/non-vocal detection

We first train the vocal and non-vocal HMM models using the manually annotated songs, then perform classification between the vocal and the non-vocal segments. Vocal/ non-vocal classification decisions are made on every subsegment of 1 second. The classification error rate is reported in the second row of Table 1. We further examine each of the classification results by the additional hypothesis test as discussed in Section 2. The hypothesis test threw out 3.0% of the segments. The results are reported in the third row of Table 1. It is observed that the hypothesis test greatly improves the performance of vocal detection achieving a 49.7% relative error reduction.

5.2. Singer identification

Several experiments are conducted to evaluate the effectiveness of vibrato based features. We use the continuous density HMM with four states and two Gaussian mixtures per state for all HMM models in our experiments. Using the training database, we train the three HMM models $\Lambda_s = \{\lambda_s^V, \lambda_s^I, \lambda_s^{CBO}\}$ for each singer of 12 singers. The process of SingerID is carried out using the vocal segments which are selected by the hypothesis test. Then, the vibrato based acoustic features are computed. During identification, a SingerID decision is made on every 5 to 10 seconds test segment. We conduct experiment to compare the SingerID performance of five feature types, namely, $OFCC_{CASCADE}$ (F1), $OFCC_{PRA}$ (F2), $OFCC_{TRI}$ (F3), MFCC (F4) [21], and LPCC (F5) [20].

Table 2. Average error rates (ER) for 12 singers using vocal segments (Case I) and using instrumental segments (Case II)

	F1	F2	F3	F4	F5
ER (%)-CaseI	16.2	17.6	18.7	21.1	21.5
ER (%)-CaseII	59.6	56.8	63.1	62.8	55.4

The second row of Table 2 shows that the $OFCC_{CASCADE}$ feature, with an average error rate of 16.2%, outperforms all other features. It is observed that the cascaded subband filters capture well the vibrato characteristics by 8.0% relative, or 1.4% absolute error rate reduction over the parabolic bandpass filters. The vibrato-motivated acoustic features ($OFCC_{CASCADE}$, $OFCC_{PRA}$, and $OFCC_{TRI}$) in general give better results than the traditional features (MFCC and LPCC).

Next, we would like to see, how the $OFCC_{CASCADE}$ feature works in the identification of singers in absence of singing vibrato. We conduct the experiments using only non-vocal (instrumental) segments. The motivation here is to see whether singing vibrato really helps $OFCC_{CASCADE}$ feature to perform better than others. The third row of Table 2 reports the results. Without surprise, the $OFCC_{CASCADE}$ feature does not work as well as others when using only instrumental segments. Generally speaking, the vibrato-motivated features ($OFCC_{CASCADE}$, $OFCC_{PRA}$, and $OFCC_{TRI}$) are more sensitive to the presence of singing voice than the traditional features (LPCC and MFCC).

We conduct further experiments using baseline HMM training method. SingerID error rate of baseline HMM method is 17.6%. The result shows proposed MM-HMM training method perform better than baseline HMM training method.

6. CONCLUSIONS

We have presented an approach for automatic singer identification of pop songs. Our contributions to the SingerID task can be summarized as follows: 1) we propose a method for vocal detection with verification; 2) we propose to use vibrato to represent the singer's characteristics through the exploration of several acoustic features; 3) we propose a method for multi-model HMM singer modeling based on the song structure. Experimental evaluations have shown the superiority of the proposed musical acoustic features over conventional methods. We conclude that the vibrato-motivated singer feature is effective, and that the MM-HMM model has successfully made use of song structure information in SingerID.

7. REFERENCES

[1] T. L. Nwe and Y. Wang, "Automatic Detection of Vocal Segments in Popular Songs", Proceedings of 5th International Conference on Music Information Retrieval (ISMIR), pp. 138-145, 2004.

[2] N. C. Maddage, C. Xu, M. S. Kankanhalli and X. Shao, "Content-Based Music Structure Analysis with Applications to Music Semantics Understanding", Proceedings of 12th Annual ACM International Conference on Multimedia, 2004.

[3] B. Whitman, G. Flake, and S. Lawrence, "Artist Detection in Music with Minnowmatch" IEEE Workshop on Neural Networks for Signal Processing, pp. 559-568, 2001.

[4] T. Zhang, "System and Method for Automatic Singer Identification", IEEE International Conference on Multimedia and Expo, 2003.

[5] M. A. Bartsch, and G. H. Wakefield, "Singing Voice Identification Using Spectral Envelope Estimation" TEEE Trans. ASSP, vol. 12, pp.100-109, 2004.

[6] N. C. Maddage, C. Xu and Y. Wang, "Singer Identification Based on Vocal and Instrumental Models" Proc. ICPR Conf., pp. 375 – 378, 2004

[7] G. H. Wakefield and M. A. Bartsch, "Where's Caruso? Singer Identification by Listener and Machine" Cambridge Music Processing Colloquium, Cambridge, England, 2003.

[8] M. Mellody, F. Herseth, and G. H. Wakefield, "Modal Distribution Analysis, Synthesis, and Perception of A Soprano's Sung Vowels" Journal of Voice, vol. 15, pp. 469-482, 2001.

[9] J. Sundberg, The Science of Singing Voice, Northern Illinois University Press , 1987.

[10] P. Eric, "Measurements of The Vibrato Rate of Ten Singers", JASA, vol. 96, pp. 1979-1984, 1997.

[11] I. Waugh, Song Structure. Music Tech Magazine, 2003.

[12] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Stress Classification Using Subband Based Features", IEICE Trans. vol. E86-D, no.3, pp. 565-573, 2003.

[13] R. A. Sukkar and L. Chin-Hui, "Vocabulary Independent Discriminative Utterance Verification for Non-keyword Rejection in Subword Based Speech Recognition", IEEE Trans. ASSP, vol. 4, pp. 420-429, 1996.

[14] R. Timmers, and P. Desain, Vibrato: Questions and Answers from Musicians and Science. Proc. Int. Conf. On Music Perception And Cognition, England, 2000.

[15] J. Bretos and J. Sundberg, "Measurements of Vibrato Parameters in Long Sustained Crescendo Notes As Sung by Ten Sopranos", Journal of Voice, vol. 17, pp. 343-352, 2003.

[16] L. J. David, Understanding Vibrato: Vocal Principles that Encourage Development, http://www.voiceteacher.com/vibrato.html

[17] C. Dromey, N. Carter and A. Hopkin, "Vibrato Rate Adjustment" Journal of Voice, vol. 17, pp. 168-178, 2003.

[18] F. A. Everest, The Master Handbook of Acoustics, New York, Mcgraw-Hill, 2001.

[19] C. Seashore, Psychology of Music, New York: Mcgraw-Hill; 1938.

[20] L. R. Rabiner, and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, N.J, 1993.

[21] C. Becchetti, and L. P. Ricotti, Speech Recognition Theory and C++ Implementation. New York: John Wiley & Sons, 1998.