ON FUSION OF TIMBRE-MOTIVATED FEATURES FOR SINGING VOICE DETECTION AND SINGER IDENTIFICATION

Tin Lay NWE and Haizhou LI

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

ABSTRACT

Timbre is the quality of sound which allows the ear to distinguish between musical sounds. In this paper, we study timbre effects in identification of singing voice segments in popular songs. Firstly, we identify between singing voice and instrumental segments in a song. Then, singing voice segments are further categorized according to their singer identity. Timbre-motivated effects are formulated by fusion of systems that use the features from vibrato, harmonic information and other features extracted using Mel and Log frequency scale filter banks. Statistical methods to select singing voice segments with high confidence measure are proposed for better performance in singer identification process. The experiments conducted on a database of 214 popular songs show that the proposed approach is effective.

Index Terms— Singing voice, singer identification, timbre, vibrato, harmonic

1. INTRODUCTION

Singing voice detection (SingVD) and Singer identification (SingerID) are the two important problems in the area of music information retrieval (MIR). The problem of SingVD is formulated as follows: given a song, classify each segment of the song as to whether it is purely instrumental (nonvocal) segment, or a mixture of vocals with/without background instrumental (vocal) segment. For the problem of SingerID, it is formulated as identifying singer of vocal segments using a classifier such as Hidden Markov Models (HMM). Although considerable progress has been made, it remains a challenge to automate the SingVD and SingerID systems with high precision.

Some recent works suggest several features for SingVD systems. Some methods originate from speech recognition. For example, Mel Frequency Cepstral Coefficients (MFCC) [1], Linear Prediction Coefficients (LPC) [1], perceptual linear prediction coefficients [2], energy function [3] and the average zero-crossing rate [3]. Others benefit from the research in music analysis, such as spectral flux [3], and relative subband energy [1]. All these features are considered to be general spectral features for speech and audio recognition.

A large number of features have been explored for SingerID. These include Mel frequency cepstral coefficients (MFCC) [4], linear prediction coefficients (LPC) [3], robust estimates of spectral envelopes [5] and octave frequency scale based cepstral coefficients [6].

Recent studies have started looking into perceptual features which are able to appreciate the aesthetic characteristics of singing voice for music contents processing and analysis. In [7], acoustic features are derived based on the instantaneous amplitude and frequency of the partials associated to vocal vibrato, to identify the singer. Mellody et al. [8] stated that temporal patterns in the vocal passage such as vibrato are likely cues to vocal quality. Besides vibrato, harmonic spectrum is also a useful feature for both SingVD and SingerID. In the study of SingVD, Goto [9] explains that, in popular songs, the harmonic structure of singing voice is often overlapped by the harmonic structure of keyboard or string instruments. Hence, one can see richer harmonic in vocals than in nonvocals. For SingerID, harmonics of soprano singer's voice are widely spaced in the spectrum in contrast to that of bass singer's voice [10]. Hence, harmonic spectrum is useful to differentiate between low and high pitch singers.

One of the basic elements of music is timbre or color. Timbre is the quality of sound which allows the ear to distinguish among different types of sounds [11]. Hence, timbre-motivated effect is universal and useful for tasks of music or audio identification systems. Several studies propose different methods to formulate timbre based features. Poli [12] measured the timbre quality from spectral envelope of MFCC features for SingerID. In [13], timber is characterized by the harmonic lines of the harmonic sound. In [11], timbre is mainly determined by the harmonic content of a sound and the dynamic characteristics of the sound such as vibrato and attack-decay envelope of the sound. In this paper, we analyse the above timbre features formulation methods [11, 12, 13] and study different combinations of these individual features in formulating timbre-motivated effects for both of SingVD and SingerID systems.

Our approach consists of three steps. First, the test song is segmented into vocal and nonvocal segments. Then, we select the vocal segments with high confidence measure using statistical method. Finally, we categorize the selected vocal segments according to their singer identity.

The rest of the paper is organized as follows. In section 2, we discuss in details the procedures to extract features from music signal. In section 3, we present the statistical methods for selecting vocal segments with high confidence measure. In section 4, we present the experiment set-up and results. Finally, we conclude our study in section 5.

2. ACOUSTIC FEATURES

As mentioned in the introduction, features based on vibrato, harmonic and timbre have characteristics that are useful to distinguish different music signals. Firstly, we study the characteristics of vibrato and harmonic in singing and instrumental signals. Then, several formulation processes for timbre based feature are discussed in the following sections.

2.1. Vibrato

Vibrato is studied in [7] for SingerID. It is modulation effect on pitch and amplitude of a musical tone [14] as illustrated in Figure 1. The two parameters: the extent and the rate are used to characterize the vibrato.

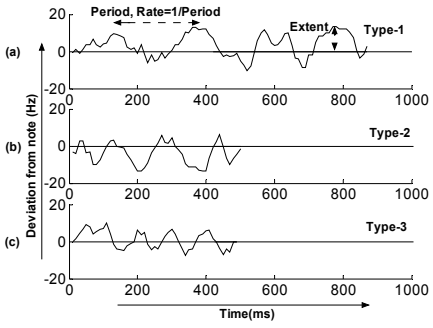

Figure 1. Three types of vibrato waveforms observed at the note of D6, 1174.6Hz being normalized to 0 at Y-axis [7]

Not all instruments can produce vibrato. For example, drum sound has no vibrato. However, vibrato presents in the instruments such as violin [15]. Figure 1 shows 3 types of vibrato. Type-1 vibrato has excursions which are balanced to the left and right of the note. And, it has wider pitch fluctuation and slower rate of vibrato which is referred to as *wobble* [16]. Type-2 and Type-3 have fluctuations which are not balanced to the left and right of the note. Type-3 vibrato has narrower pitch fluctuation and faster rate and is referred to as *bleat* [17]. Singers or performers have their personalized style of vibrato [7].

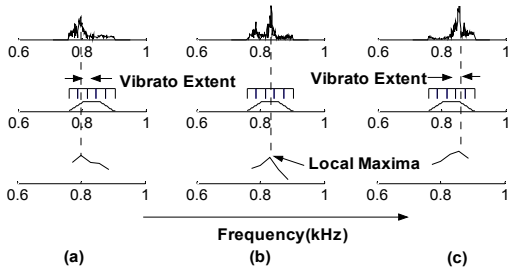

Figure 2. Vibrato fluctuations and cascaded bandpass filtering observed at the note G#5, 830.6Hz. (a) vibrato fluctuates left (b) no fluctuation (c) vibrato fluctuates right. The upper panel shows the spectrum partial. The middle panel presents the frequency response of the vibrato filter. The lower panel demonstrates the output amplitudes of the vibrato filter. [7].

Vibrato information of music signal is extracted using cascaded subband filter (referred to as vibrato filter) [7] which is shown in middle panels of Figure 2. The vibrato filter has two cascaded layers of subbands. The first layer has 96 overlapped trapezoidal filters which span up to 16kHz (8 octaves). The center frequencies are located at each of the musical notes. Bandwidths of the filters are ± 1.5 semitone from each note since vibrato extent can increase more than ± 1 semitone [18]. The list of the frequencies of the musical notes can be found in [19]. The second layer has 5 non-overlapped rectangular filters of equal bandwidths for each trapezoidal subband of first layer. Trapezoidal filters are tapered between ± 0.5 semitone to ± 1.5 semitone. The vibrato fluctuations are observed by tracking the sinusoids which is the local maxima in the instantaneous amplitude output of the subbands in the second layer as shown in the lower panel of Figure 2. We refer the sinusoids with slowly time-varying amplitude and frequency as partials. Local maxima indicate the position of the partial. The distance between the center frequency of the corresponding filter and the position of the partial 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 vibrato filters are able to capture 3 different vibrato types.

2.2. Harmonic

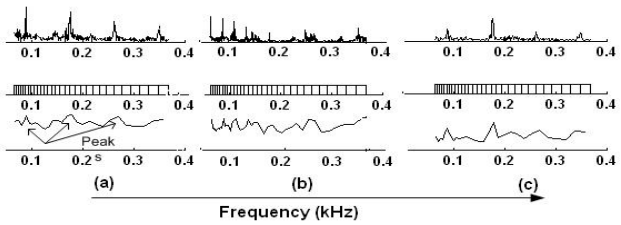

Figure 3. Harmonics and harmonic filtering (a) Singing voice of soprano singer (b) Singing voice of bass singer (c) nonvocal signal

The harmonic structures of singing and instruments are often overlapped [9]. Hence, harmonic intensity of singing voice segments are higher than that of instrumental segments [3] as shown in upper panels of Figure 3. Sopranos have higher fundamental frequency than bass singers. Hence, harmonics of soprano's voice is widely spaced in contrast to that of bass singing as shown in upper panels of Figures 3(a) and 3(b). These harmonic structures are captured using harmonic filter which is shown in middle panels of Figures 3. The centre frequencies of the harmonic filter are located at each musical notes listed in [19]. The bandwidths are ± 1.5 semitone from each note and the filters spans up to 16kHz. Outputs of harmonic filters are given in lower panels of Figure 3. For soprano, lower panel of Figure 3(a) shows widely spaced peaks. However, the peaks are narrowly spaced in lower panel of Figure 3(b) for bass singers. Output subband energies of instrumental signal is lower than that of singing voice in lower panels of Figure 3.

2.3. Acoustic feature formulation

A music signal is divided into frames of 15ms with 10ms overlapping. Each frame is multiplied by a Hamming window to minimize signal discontinuities at the end of each frame. Then, the audio frame is passed through a bank of vibrato filters. Then, log energy of each band in the second layer is calculated. Finally, a total of 13 Octave Frequency Cepstral Coefficients ($OFCC_{VIB}$) are

computed from log energies using Discrete Cosine Transform for each audio frame. We then replace the vibrato filters with harmonic filters to compute the $OFCC_{HAR}$.

We compute Mel Frequency Cepstral Coefficients (MFCC) using Mel-frequency scale and Log Frequency Cepstral Coefficients (LFCC) using logarithmic frequency scale [20]. LFCC was shown to be useful for SingVD in [20]. For all features, we augment the feature coefficients with time derivatives from two neighboring frames to capture the temporal information to take care of attack-decay envelope in timbre feature [11].

2.4. Formulation of timbre effect

Sounds may be generally characterized by pitch, loudness and quality. Sound quality or timbre describes those characteristics of sound which allow human ears to distinguish sounds which have the same pitch and loudness. Timbre is a general term for the distinguishable characteristics of a tone. Several timbre formulation methods using 1) MFCC [12], 2) harmonic lines of the harmonic sound [13] and 3) combination of harmonic, vibrato and attack-decay envelope of the sound [11] are proposed in literature. All these components seem to be important in timbre formulation and we formulate timbre effect through fusion of these methods. We build 4 systems using $OFCC_{VIB}$, $OFCC_{HAR}$, MFCC and LFCC features for each of SingVD and SingerID. We consider each system as an 'expert'. We combine these experts to achieve timbre-motivated effect. In this fusion, we weight better experts more than others based on their prior performance. To recognize the timbre of a tone, it takes duration of about 60ms. If a tone is shorter than 4ms, it is perceived as an atonal click [11].

3. VOCAL DETECTION

Vocal detection errors can affect SingerID performance. We formulate the vocal detection using the following hypothesis test [21]. The following likelihood ratio and a threshold δ are used for vocal detection decision verification.

$$\frac{p(O|\lambda^v)}{p(O|\lambda^n)} \underset{\text{Reject}}{\overset{\text{Accept}}{\gtrless}} \delta \qquad (1)$$

where λ^v and λ^n denotes vocal and nonvocal models respectively. And, O is the sequence of input feature vectors representing a song segment. False alarms are removed by using higher threshold δ.

4. EXPERIMENTS

We compile 6 different data sets as listed in Table 1. The databases in SingerID include songs from 12 solo singers [7].

Table 1. Number of songs in the 6 data sets

Systems	TrainDB	DevelopmentDB	TestDB
SingVD	35	25	45
SingerID	48	25	36

The 6 song databases do not overlap. Each song is annotated manually to obtain the vocal and nonvocal segments to provide the ground truth labels. These labels are used for performance benchmarking in SingVD.

Several experiments are conducted to observe the different types of timbre-motivated effect. In theory, timbre is a universal feature and can be used to identify different audio types. Hence, all formulations of timbre-motivated effects are used in both SingVD and SingerID systems. We use the continuous density HMM with four states and two Gaussian mixtures per state for all HMM models in our experiments. As mentioned in Section 3, duration of about 60ms is necessary to recognize a timbre of a tone. Hence, each of 4 features listed in 1st column of Table 2 is derived at 6 different window sizes: 15ms, 25ms, 35ms, 45ms, 55ms and 65ms. Frame rate of 10 ms is used for all window sizes. Four individual feature systems and four fused feature systems are built for each of the 6 window sizes. For brevity, we give each system a reference name as in Table 2.

Table 2. Choice of features and their systems

Individual Feature	System	Timbre-motivated Feature Fusion	System
MFCC	F1	F3 + F4	T1
LFCC	F2	F1 + F3 + F4	T2
$OFCC_{VIB}$	F3	F2 + F3 + F4	T3
$OFCC_{HAR}$	F4	F1 + F2 + F3 + F4	T4

When fusing systems, the individual system is weighted based on their prior performance. The weights are obtained through the Development data set.

For SingVD system, we train three models, $\Lambda_{SingVD} = \{\lambda_v^M, \lambda_v^F, \lambda_n^I\}$, each for male vocal, female vocal and nonvocal sound using TrainDB. The average error rates for 4 individual systems (F1, F2, F3 and F4) and 4 timbre-motivated fusion systems (T1, T2, T3 and T4) are presented in Figure 4 where SingVD error rates (ER) are computed in 20ms frames.

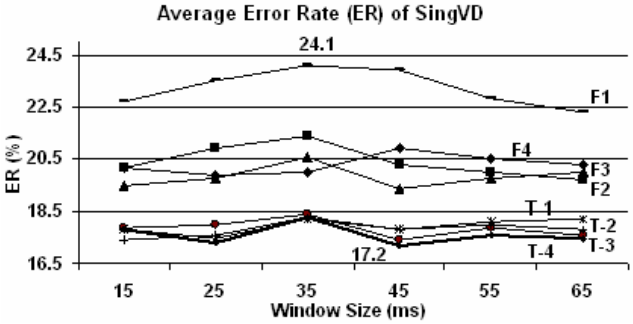

Figure 4. Average error rates on TestDB of SingVD system

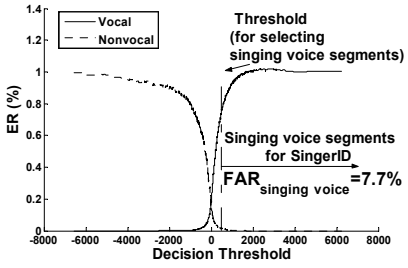

Figure 5. Receiver Operating Curve (ROC) to select high confidence vocal segments

In SingerID experiments, we extract high quality vocal segments by using those segments of high confidence. Using equation (1), we draw ROC curves (Figure. 5) of vocal and nonvocal classes using DevelopmentDB. Then, we select samples

with high level of confidence by discarding some unsure samples failing below the threshold as shown in Figure 5. The threshold to select the high confidence vocal segments is 500 which give FAR of 7.7%. We use the selected samples to conduct SingerID experiments. We train an HMM model for each singer using TrainDB. The average error rates for 4 individual systems (F1, F2, F3 and F4) and 4 timbre-motivated fusion systems (T1, T2, T3 and T4) for 6 window sizes for SingerID system are presented in Figure 6. The weights for timbre-motivated system fusions are determined using DevelopmentDB of SingerID. During identification, a SingerID decision is made on every 5 to 10 seconds test segment. SingerID error rates (ER) are computed on 1 second segments.

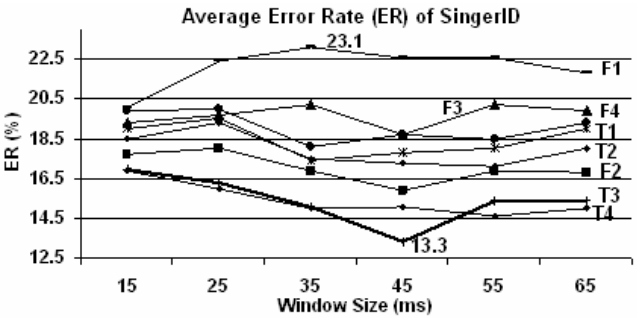

Figure 6. Average error rates on TestDB of SingerID system

The results in Figures 4 and 6 shows that timbre-motivated fusions, T3 and T4, perform the best among other features and fusion systems for both of SingVD and SingerID respectively. The average error rates (ER) are 17.2 % and 13.3% respectively. These systems outperform previously reported F2 (for SingVD) [20] and F3 (for SingerID) [7] systems. The timbre-motivated fusion systems, T1, T2, T3 and T4, in general give better results than individual systems. Among fusion systems, T3 and T4 perform better than T1 and T2 fusions. It can be conjectured from the results that timbre effect can be better formulated by fusing not only harmonic, vibrato and attack-decay envelope of the sound (T2 system [11]) but also MFCC (F1) and LFCC (F2) together. As for further analysis of the results, ERs are the lowest at 45ms window size for both of SingVD and SingerID. Hence, we believe that longer window size is suitable to extract timbre characteristics from a music signal.

T3 system (45ms window size) that use vocal segment selection method achieve 13.3% ER in SingerID (Figure 6). To show the effect of selecting high confidence vocal segments, we re-run the experiments T3 (45ms window size) without selection. It gives an ER rate of 15%. Hence, vocal segment selection method helps to reduce ER by 1.7% or 11.3% relative ER reduction.

5. CONCLUSIONS

We have presented an approach for fusion of timbre-motivated features, which is found effective in both SingVD and SingerID systems. The contributions of this work include: 1) we formulate timbre motivated effects by fusing features of vibrato, harmonic, MFCC and LFCC in several combinations. 2) we employ statistical method for selecting vocal segments with high confidence measure for SingerID. 3) we show that high quality vocal detection is desired for SingerID.

6. REFERENCES

[1] G. Tzanetakis, "Song-Specific Bootstrapping of Singing Voice Structure", *IEEE Int. Conf. Multimedia and Expo,* 2004.

[2] A.L. Berenzweig, D.P.W. Ellis, and S. Lawrence, "Using Voice Segments To Improve Artist Classification of Music," *AES 22nd Int. Conf. Espoo*, Finland, 2002.

[3] T. Zhang, "System and Method for Automatic Singer Identification", *IEEE Int. Conf. Multimedia and Expo*, Baltimore, MD, 2003.

[4] 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.

[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] T.L. Nwe and H. Li, "Exploring Vibrato-Motivated Acoustic Features for Singer Identification", *IEEE Trans. on Speech and Audio Processing*, vol. 15, no.2, pp: 519-530, 2007.

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

[9] M. Goto, "A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals", *Speech Communication,* vol. 43, no. 4, pp. 311-329, September 2004.

[10]E. Joliveau, J. Smith and J. Wolfe, "Vocal Tract Resonances in Singing: The Soprano Voice," *Journal of Acoustical Society of America*, vol. 116, pp. 2434-2439, 2004.

[11] F. Winckell, *Music, sound and sensation*. Dover, NY ,1967.

[12] G. D. Poli, and P. Prandoni, "Sonological Models for Timber Characterization," *Journal of New Music Research*, vol. 26, pp. 170-197.

[13] T. Zhang, and C.C.J. Kuo, *Content-Based Audio Classification and Retrieval for Data Parsing*, Kluwer Academic, USA, 2001.

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

[15]R.B. Macleod, "Influences of Dynamic Level and Pitch Height on The Vibrato Rates and Widths of Violin and Viola Players", PhD thesis, college of music, 2006, Florida state university.

[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," *J. Voice*, vol. 17, pp. 168-178, June 2003.

[18]J. Sundberg,. *The Science of Singing Voice*. Northern Illinois University Press , 1987, ch. 8.

[19] F.A. Everest, *The Master Handbook of Acoustics*. New York, McGraw-Hill, 2001.

[20] T.L. Nwe and Y. Wang., "Automatic Detection of Vocal Segments in Popular Songs", In Proceedings of *5th International Conference on Music Information Retrieval (ISMIR)*, pp. 138-145, October 10-14, 2004, Barcelona, Spain.

[21] C. Fredouille, J-F. Bonastre, and T. Merlin, "Bayesian Approach Based-Decision in Speaker Verification," *A Speaker Odyssey*, Crete, Greece, 2001.