Proc of Adaptive Multimedia Retrieval 2011, Barcelona, Spain

Context-aware features for singing voice detection in polyphonic music

Vishweshwara Rao, Chitralekha Gupta, and Preeti Rao

Department of Electrical Engineering
IIT Bombay, Mumbai - 76,India
{vishu,chitralekha,prao}@ee.iitb.ac.in

Abstract. The effectiveness of audio content analysis for music retrieval may be enhanced by the use of available metadata. In the present work, observed differences in singing style and instrumentation across genres are used to adapt acoustic features for the singing voice detection task. Timbral descriptors traditionally used to discriminate singing voice from accompanying instruments are complemented by new features representing the temporal dynamics of source pitch and timbre. A method to isolate the dominant source spectrum serves to increase the robustness of the extracted features in the context of polyphonic audio. While demonstrating the effectiveness of combining static and dynamic features, experiments on a culturally diverse music database clearly indicate the value of adapting feature sets to genre-specific acoustic characteristics. Thus commonly available metadata, such as genre, can be useful in the front-end of an MIR system.

1 Introduction

The automatic identification of audio segments within a song, that contain the singing voice (vocal part) is important in several Music Information Retrieval (MIR) applications such as artist identification [1], voice separation [2] and lyrics alignment [3]. SVD is typically viewed as an audio classification problem where features that distinguish vocal segments from purely instrumental segments in music are fed to a machine-learning algorithm previously trained on manually labeled data. Until recently, singing voice detection algorithms employed solely static features, typically comprising frame-level spectral measurements such as combinations of mel-frequency cepstral coefficients (MFCCs) [1], [2], [4]-[6], warped or perceptually derived linear predictive coefficients (LPCs) [1], [7]-[9], log frequency power coefficients (LFPC) [10], harmonicity related features [11]-[13] and other spectral features such as flux, centroid and roll-off [14], [15].

A consideration of acoustic attributes necessary for the detection of vocal segments in music fragments by humans is interesting for its potential in guiding the search for suitable features for the task. An experiment in which subjects listened to short excerpts (less than 500 ms long) of music from across diverse genres showed that human listeners can reliably detect the presence of vocal

content in such brief excerpts [16]. The presence of note transitions in the excerpt was found to be especially useful, indicating that both static features and dynamic features (changes) provide important perceptual cues to the presence of vocalization.

Dynamic features for singing voice detection have largely been confined to feature derivatives, representing very short-time dynamics. A few studies have explored the observed pitch instability of the voice relative to most accompanying instruments in the form of features representing longer-term dynamics of pitch and/or the harmonics such as arising from vibrato and tremolo in singing [10], [17]. Clearly, there is scope for improvement in terms of capturing the essential differences in the dynamics of the singing voice and musical instruments in a compact and effective way.

While most research results in MIR are reported on collections drawn from one or another culture (mostly Western), we are especially interested in features that work cross-culturally. It would be expected that certain feature attributes are more discriminative on particular music collections than on others, depending on the musical content due to the inherent diversity of both singing styles and instrumentation textures across cultures [18]. The work reported in this paper is an attempt to study this. A similar approach underlies the work on collection-specific features for image retrieval, also appearing in this volume [19].

We choose to focus on 'vocal' music i.e. where the singing voice is the predominant melodic entity whenever present. Further, as far as possible, the instrumental sections of the song include a predominant melodic instrument. One of the categories that was badly classified, and also negatively influenced the training set effectiveness, in the study of [20] corresponded indeed to songs with predominant melodic instruments or with singing co-occurring with such instruments. Paralleling this observation, are studies on predominant musical instrument identification in polyphony which state that pitched instruments are particularly difficult to classify due to their sparse spectra [21]. Thus our choice of evaluation datasets is guided by the known difficulty of the musical context as well as the wide availability of such a category of music cross-culturally. We consider the effective extraction and evaluation of static and dynamic features on a dataset of vocal music drawn from Western popular, Greek and three distinct Indian genres: North Indian (Hindustani) classical, South Indian (Carnatic) and popular (Bollywood or film music).

2 Dominant Source Spectrum Isolation

Previous studies on SVD have extracted features by processing the short-time Fourier transform (STFT) without attempting to isolate the spectra of individual sound sources. Recently, a spectral processing technique called accompaniment sound reduction, which uses the predominant-F0, for robust singing voice modeling in polyphony was presented by Fujihara et.al.[22]. In this section we describe a method for dominant source spectrum isolation along the lines of their method but with some important enhancements.

Harmonic sinusoidal modeling is applied to detect and track F0-specific harmonic components from the spectrum. We use a predominant F0 extraction system designed for robustness in the difficult context of concurrent pitched accompaniment [23]. In order to study the comparative performance of features unobscured by possible pitch detection errors, we carry out feature extraction in both automatic and semi-automatic modes of pitch detection for the dominant source spectrum isolation [24]. In the latter mode, pitch analysis parameters are selected considering a priori information on the pitch range of the voice in the given piece of music.

The audio signal is processed at a frame rate of 100 frames/sec with the STFT computed from 40 ms long Hamming windowed analysis frames. Rather than identify spectral local maxima in the vicinity of expected harmonic locations as detected harmonics, as done in [22], we first extract reliable sinusoidal partials using a main-lobe matching technique, which we have previously found to be particularly robust to polyphony and signal non-stationarity [25]. We then search the sinusoidal space in a 50-cent neighborhood (subject to a 50 Hz absolute limit) of expected harmonic locations, based on known predominant F0, to identify local harmonic components. This results in a harmonic line spectrum for each analysis frame.

Next, partial tracking is applied to the harmonic line spectra over time for better source spectral isolation during transients and note changes. Similar to Serra [26], partial tracking is improved by biasing trajectory formation towards expected harmonic locations based on the detected pitch. Tracks are formed and ordered by their respective harmonic index. We apply a one semitone threshold on track continuation i.e. a track will 'die' if there does not exist any sinusoid within 1 semitone of the last tracked frequency. In addition to frequency proximity, we incorporate sinusoid amplitudes in the cost function, as given below, to account for high amplitude partials that might lose out to spurious nearby peaks.

$$J = \left|(\omega_n^k - \omega_{n-1}^m) \times \log(A_n^k / A_{n-1}^m)\right| \tag{1}$$

where ω_{n-1}^m and A_{n-1}^m are the frequency and amplitude resp. of the m^{th} harmonic track in frame $n-1$ and ω_n^k and ω_{n-1}^m are the frequency and amplitude resp. of the k^{th} local sinusoid in the frame n that is competing for joining the track.

Rather than extract features from the discrete frequency spectra obtained after harmonic sinusoidal modeling, we use smooth spectra derived by the log-linear interpolation of the harmonic spectral amplitudes as follows. Given a set of estimated amplitudes S(ω_1), S(ω_2), , S(ω_L) at L uniformly spaced harmonic frequencies ω_1, ω_2,, ω_L we generate spectral amplitudes Q(Θ) at fixed DFT bin spacing as shown below:

$$Q(\Theta_j) = 10^{\log|S(\omega_k)| + (\frac{\Theta_j - \omega_k}{\omega_{k+1} - \omega_k})(\log|S(\omega_{k+1})| - \log|S(\omega_k)|)} \tag{2}$$

where $\omega_k < \Theta_j < \omega_{k+1}$. The interpolation serves to make any spectral envelope features extracted subsequently less dependent on the F0 and thus more representative of source timbre.

3 Feature Extraction

In this section we describe the different static and dynamic features investigated, motivated by the signal characteristics. Static features are computed locally over short sliding analysis frames while dynamic features are computed over larger non-overlapping time windows called texture windows. To avoid confusion, we will use the term *frame* to refer to the analysis frame and *window* for the texture windows. Silence frames, i.e. frames with energy lower than the song-level global maximum energy by more than 30 dB, are not processed. The features for such frames are interpolated from valid feature values in adjacent frames.

3.1 Static Timbral Features

In previous work on SVD the most commonly used features are static timbral descriptors, which attempt to capture the timbral aspects of musical sounds and do not utilize any specific long-term traits of the singing voice or instruments. Rocomora and Herrera compared the performance of a comprehensive list of static features in an SVD task and found that MFCCs resulted in the best classification accuracies [27]. The tested features were however all extracted from the overall polyphonic spectrum.

Recently Fujihara and Goto proposed that the feature set of F0 and harmonic powers were found to be highly discriminatory for SVD when source spectral isolation was possible [3]. The harmonic powers were normalized for each song. The normalized power of the h^{th} component at time t is given by

$$p_h^{'t} = \log p_h^t - \frac{\sum_t \sum_h \log p_h^t}{T \times H} \tag{3}$$

where $p_h^{'t}$ represents the original power, T is the total number of frames in the song, and H is the number of harmonic components considered per frame.

The interpolated spectrum of Eq. (2) provides a spectral envelope that is potentially indicative of the underlying instrument's resonances or formants. Two more features: the sub-band spectral centroid (SC) and sub-band energy (SE) [28] as given by

$$SC = \frac{\sum_{k=k_{low}}^{k_{high}} f(k)|X(k)|}{\sum_{k=k_{low}}^{k_{high}} |X(k)|} \qquad SE = \sum_{k=k_{low}}^{k_{high}} |X(k)|^2 \tag{4}$$

where $f(k)$ and $|X(k)|$ are frequency and magnitude spectral value of the k^{th} frequency bin, and k_{low} and k_{high} are the nearest frequency bins to the lower and upper frequency limits on the sub-band respectively. SE is normalized by its maximum value for a song. The sub-band for SC ranges from 1.2-4.5 kHz and that for SE is from 300 to 900 Hz.

3.2 Dynamic Timbral Features (C2)

The explicit modeling of temporal dynamics, as an important component of perceived timbre, has found a place in recent research on musical instrument

recognition [29], [30]. Based on similar considerations, features linked to the temporal evolution of the spectral envelope can be designed to capture specific attributes of the instrument sound. For example, in view of the well-known source-filter model of sound production applicable to human speech and to musical instruments, variations in spectral envelope with time represent variations of the filter component independent of the source component (i.e. F0). Such a dynamic feature could potentially discriminate between singing voice and (even similarly pitch-modulated expressive) musical instruments due to the absence of formant articulation dynamics in the latter. Another distinguishing aspect could be the attack-decay envelope peculiar to a particular instrument as reflected in the variation of spectral amplitudes over note durations.

One of the problems with the effective use of timbral dynamics for instrument classification in polyphony was found to be the lack of ability to pay selective attention to isolated instrumental spectra [31]. Here we describe the extraction of the dynamics of the two static timbral features using our isolated dominant source spectral representation. The first of these is the sub-band spectral centroid (SC) as computed in the previous section. The specific sub-band chosen [1.2-4.5 kHz] is expected to enhance the variations in the 2nd, 3rd and 4th formants across phone transitions in the singing voice. This feature is expected to remain relatively invariant over note transitions in the instrument. Although the band-limits are restrictive, even very high pitched instruments will have at least 3 harmonics present within this band so that their spectral envelope can be relatively well represented. The second feature for dynamic analysis is the sub-band energy (SE) as computed in the previous section. The band limits for SE ([300-900 Hz]) are expected to enhance the fluctuations between voiced and unvoiced utterances while remaining relatively invariant to small instrumental note transitions. Fluctuations in this feature should be evident even if the signal representation captures some pitched accompanying instrument information during unvoiced sung sounds.

To capture meaningful temporal variations in the dynamics of the above timbral features, it is necessary to choose the duration of the observation interval appropriately [29]. We choose three different time scales (texture windows) for our feature set: 0.5 sec (long note duration), 1 and 2 sec intervals (to capture note articulation changes in both fast and slow singing). We represent the dynamics via the standard deviation (std. dev.) and specific modulation energies over the different observation intervals. These modulation energies are represented by a modulation energy ratio (MER). The MER is extracted by computing the DFT of the feature trajectory over a texture window and then computing the ratio of the energy in the 1-6 Hz region in this modulation spectrum to that in the 1-20 Hz region as shown below:

$$MER = \frac{\sum_{k=k_{1Hz}}^{k_{6Hz}} |Z(k)|^2}{\sum_{k=k_{1Hz}}^{k_{20Hz}} |Z(k)|^2} \tag{5}$$

where $Z(k)$ is the DFT of the mean-subtracted feature trajectory $z(n)$ and k_{fHz} is the frequency bin closest to fHz. We assume that the fastest syllabic rate

possible, if we link each uttered phone to a different note in normal singing, should not exceed 6 Hz. Steady note durations are not expected to cross 2 seconds. The std. dev. and MER of the above features are expected to be higher for singing than instrumentation.

3.3 Dynamic F0-Harmonic Features (C3)

Singing differs from several musical instruments in its expressivity, which is physically manifested as the instability of its pitch contour. In western singing, especially operatic singing, voice pitch instability is marked by the widespread use of vibrato, a periodic, sinusoidal modulation of phonation frequency during sustained notes [32]. Within non-western forms of music, such as Greek Rembetiko and Indian classical music, voice pitch inflections and ornamentation are extensively used as they serve important aesthetic and musicological functions. On the other hand, the pitch contours of several accompanying musical instruments, especially keyed instruments, are usually very stable and incapable of producing pitch modulation.

There has been limited previous work on utilizing voice-pitch instability to SVD. Shenoy, Wu and Wang [33] exploit pitch instability in an indirect way by applying a bank of inverse comb filters to suppress the spectral content (harmonics) of stable-pitch instruments. Nwe and Li [10] made use of a bank of band-pass filters to explicitly capture the extent of vibrato within individual harmonics upto 16 kHz. Regnier and Peeters attempted a more direct use of frequency and amplitude instability of voice harmonics (manifested as vibrato and tremolo respectively in western music) [17]. Their method is based on the observation that the extents of vibrato and tremolo in singing are different than those for most instruments. The authors have also previously used pitch stability cues to prune flat-pitched instruments tracks in a harmonic sinusoidal model [34].

In this study we do not restrict ourselves to targeting particular types of pitch modulation such as vibrato but extract some statistical descriptors (mean, median, std. dev.) of general pitch instability-based features over texture windows of expected minimum note duration (here 200 ms). These features are the first-order differences of the predominant-F0 contour and the subsequently formed harmonic frequency tracks. The track frequencies are first normalized by harmonic index and then converted to the logarithmic cents scale so as to maintain the same range of variation across harmonics and singers' pitch ranges. For the latter we group the tracks by harmonic index (harmonics 1-5, harmonics 6-10, harmonics 1-10) and also by low and high frequency bands ([0-2 kHz] and [2-5 kHz]). This separation of lower and higher harmonics/frequency bands is due to the observation that when the voice pitch is quite stable, the lower harmonics do not display much instability but this is clearly visible in the higher harmonics. However when the voice pitch exhibits large modulations the instability in the lower harmonic tracks is much more clearly observed but often the higher harmonic tracks are distorted and broken because of the inability of the sinusoidal model to reliably track their proportionately larger fluctuations. We

Table 1. List of features in each category. Bold indicate finally selected feature

C1 **Static timbral**	**C2** **Dynamic timbral**	**C3** **Dynamic F0-Harmonic**
F0	Δ 10 Harmonic powers	Mean and **median** of **ΔF0**
10 Harmonic powers	ΔSC and ΔSE	Mean, median and Std.Dev. of ΔHarmonic ϵ[0 2 kHz]
Spectral centroid (SE)	**Std. Dev.** of **SC** for 0.5, **1** and 2 sec	Mean, median and Std.Dev. of ΔHarmonic ϵ[2 5 kHz]
Sub-band energy (SE)	**MER** of **SC** for 0.5, 1 and **2** sec	Mean, median and Std.Dev. of ΔHarmonics 1 to 5
	Std. Dev. of **SE** for **0.5**, 1 and 2 sec	Mean, **median** and Std.Dev. of **ΔHarmonics 6 to 10**
	MER of SE for 0.5, 1 and 2 sec	Mean, **median** and Std.Dev. of **ΔHarmonics 1 to10**
		Ratio of mean, median and Std.dev. of ΔHarmonics 1 to 5 : ΔHarmonics 6 to 10

also compute the ratio of the statistics of the lower harmonic tracks to those of the higher harmonic tracks since we expect these to be much less than 1 for the voice but nearly equal to 1 for flat-note instruments.

A complete list of features appears in Table 1. All features are brought to the time-scale of 200 ms long decision windows. The frame-level static timbral features, generated every 10 ms, are averaged over this time-scale and the timbral dynamic features,generated over larger windows: 0.5, 1 and 2 sec, are repeated within 200 ms intervals. The F0-harmonic dynamic features were generated at 200 ms non-overlapping windows in the first place and do not need to be adjusted. Next, feature subset selection is applied to identify a small number of highly predictive features and remove as much redundant information as possible. Reducing the dimensionality of the data allows machine learning algorithms to operate more effectively from available training data. Feature selection is achieved by measuring the information gain ratio of the feature with respect to a class [35]. Each feature is assigned a score based on the information gain ratio to obtain a ranked feature list.

4 Classification

Statistical classification methods are very effective in exploiting the overall information provided about the underlying class by the set of diverse features

if suitable data is provided for the training of the statistical models. Previous studies on SVD have employed a variety of classifiers. We use a standard GMM classifier [36] for the evaluation of our features. Using 4 mixtures for each of the two models (singing voice and instrumentation) with full covariance matrices was found suitable in preliminary testing. The dimensionality of the GMM is determined by that of the feature vector. We would like to study the effectiveness of individual feature sets as well as that of the combination of individual classifiers trained with individual feature sets, which can improve the effectiveness of the system while offsetting difficulties arising from high dimensionality [37]. Combining the likelihood scores of classifiers is particularly beneficial if the corresponding individual feature sets represent complementary information about the underlying signal. Weighted linear combination of likelihoods provides a flexible method of combining multiple classifiers with the provision of varying the weights to optimize performance [37].

5 Experiments

The singing voice detection task is carried out on a database comprising excerpts from 5 distinct music genres. A feature set of the first 13 MFCCs extracted from the frame-level magnitude spectrum applied to a GMM classifier with 4 mixtures per class is considered as a baseline system. As mentioned earlier, Rocamora and Herrera had found the performance of the MFCC features among the best performing for in the SVD task [27]. The newly proposed features of the present work are applied to the same classifier framework in order to evaluate the performance improvement with respect to the baseline feature set and to derive a system based on possible feature combinations that performs best for a specific genre and across genres.

5.1 Database Description

All the audio excerpts in our database contain polyphonic music with lead vocals and dominant pitched melodic accompaniment, and are in 22 kHz 16-bit mono format. Vocal and purely instrumental sections of songs categorized as above from five different genres of music have been selected - Western popular, Greek Rembetiko, Indian popular (Bollywood), North Indian classical (Hindustani) and South Indian classical music (Carnatic). The Western and Greek clips have been selected from the datasets used in [9] and [38] respectively. The total size of the database is about 65 min. with roughly 13 min. per genre. Information about the number of songs, and vocal and instrumental durations for each genre appears in Table 2. In a given genre a particular artist is represented by only one song.

The selected genres are marked by distinct singing styles and instrumentation. A noticeable difference between the singing styles of the Western and non-Western genres is the extensive use of pitch-modulation (other than vibrato) in the latter. Pitch modulations further show large variations across non-Western

Table 2. Duration Information of Audio Database

Genre	Number of songs	Vocal duration	Instrumental duration	Overall duration
I. Western	11	7m 19s	7m 02s	14m 21s
II. Greek	10	6m 30s	6m 29s	12m 59s
III. Bollywood	13	6m 10s	6m 26s	12m 36s
IV. Hindustani	8	7m 10s	5m 24s	12m 54s
V. Carnatic	12	6m 15s	5m 58s	12m 13s
Total	**45**	**33m 44s**	**31m 19s**	**65m 03s**

genres in the nature, shape, extents, rates and frequency of use of specific pitch ornaments. Further, whereas Western, Greek and Bollywood songs use syllabic singing with meaningful lyrics, the Hindustani and Carnatic music data is dominated by melismatic singing (several notes on a single syllable in the form of continuous pitch variation). The instruments in Indian popular and Carnatic genres are typically pitch-continuous such as the violin, saxophone, flute, *shehnai*, and *been*, whose expressiveness resembles that of the singing voice in terms of similar large and continuous pitch movements. Although there are instances of pitch-continuous instruments such as electric guitar and violin in the Western and Greek genres as well, these, and the Hindustani genre, are largely dominated by discrete-pitch instruments such as the piano and guitar, accordion and the harmonium. A summary of genre-specific singing voice and instrumental characteristics appears in Table 3.

5.2 Selected Features

Each of the feature sets (C1, C2 and C3) is fed to the feature selection system to generate a ranked list for each genre. A feature vector comprising the top-N features common across genres was tested for SVD in a cross-validation classification experiment to select N best features. For C1 it was observed that using all the features in this category consistently maximized the classification accuracies across genres and so we did not discard any of these features. For C2 and C3 we observed that the top six selected features for each of the genres consistently maximized their respective classification accuracies. The finally selected features in each of the categories appear in bold in Table 1.

In the dynamic timbral feature set, the Δ values of the static features are ignored by the feature selection algorithm in favour of the std. dev. and MER values of the SC and SE. The feature selection algorithm took into account the expected high degree of correlation between the same dynamic features at different time-scales and only selected at most one time-scale for each dynamic feature. For the F0-harmonic dynamic feature set, the final selected features (C3) are the medians of ΔF0 and ΔHarmonic-tracks rather than their means or std. dev. The choice of medians was seen to be driven by the common occurrence of intra-window flat-pitched instruments note-transitions where the F0/Harmonic

Table 3. Genre-specific singing and instrumental characteristics

Genre	Singing	Dominant Instrument
I. Western	Syllabic. No large pitch modulations. Voice often softer than instrument.	Mainly flat-note (piano, guitar). Pitch range overlapping with voice.
II. Greek	Syllabic. Replete with fast, pitch modulations.	Equal occurrence of flat-note plucked-string /accordion and of pitch-modulated violin.
III. Bollywood	Syllabic. More pitch modulations than Western but less than other Indian.	Mainly pitch-modulated wood-wind & bowed instruments. Pitches often much higher than voice.
IV. Hindustani	Syllabic and melismatic. Varies from long, pitch-flat, vowel-only notes to large & rapid pitch modulations.	Mainly flat-note harmonium (woodwind). Pitch range overlapping with voice.
V. Carnatic	Syllabic and melismatic. Replete with rapid pitch modulations.	Mainly pitch-modulated violin. F0 range generally higher than voice but has some overlap in pitch range.

tracks make a discontinuous jump. In such cases, the means and standard deviations of the Δs exhibit large values as opposed to the relatively unaffected median values which remain low.

5.3 Evaluation

An N-fold cross-validation classification experiment is carried out within each genre. Since the durations of different songs within a particular genre are unequal we consider each song to be a fold so as to avoid the presence of tokens of the same song in the training and testing data to achieve a 'Leave 1 Song out' cross-validation. This is done separately for the semi- and fully-automatic F0 extraction based source spectrum isolation methods. In each case, we first evaluate the performance of the baseline features (MFCCs), before and after applying dominant source spectrum isolation. We next evaluate the performance of the different categories of feature sets individually (C1, C2 and C3). Further, we evaluate the performance of different feature set combinations: C1+C2, C1+C3 and C1+C2+C3, using a linear combination (equal weights) of the log-likelihood outputs per class of separate classifiers for each feature set. Vocal/non-vocal decision labels are generated for every 200 ms texture window. While the ground-truth labels for the Western and Greek genres were available with the datasets, the

Table 4. %Correct classification for different genres using semi-automatic predominant F0 extraction;Bold indicates best achieved in each genre

Genre	I	II	III	IV	V	Total
Baseline	77.2	66.0	65.6	82.6	83.2	74.9
MFCCs(after source isolation)	78.9	77.8	78.0	85.9	85.9	81.2
C1	79.6	77.4	79.3	82.3	87.1	81.0
C2	72.0	77.9	80.0	70.1	65.3	73.2
C3	64.3	77.0	68.3	83.7	70.2	72.6
C1+C2	**82.3**	83.6	85.4	83.3	86.8	84.2
C1+C3	80.2	83.4	81.7	**89.7**	**88.2**	84.5
C1+C2+C3	81.1	**86.9**	**86.4**	88.5	87.3	**85.9**

remaining datasets were manually labeled. In all cases classification performance is given by the percentage of decision windows that are correctly classified [2].

5.4 Results and Discussion

The 'Leave 1 Song out' cross-genre classification results for semi- and fully-automatic predominant F0 extraction based source spectrum isolation methods are given in Tables 4 and 5 respectively. From Table 4, we see that the best overall performance is achieved for the combination of all three feature sets and is significantly (10-12%) higher than the baseline performance. For the static feature comparison it can be seen that the feature sets C1 and MFCCs after source isolation show similar performance and are, in general, superior to the baseline features(non-source-isolated MFCC). The clear superiority of the C1+C2+C3 feature combination over the static feature set C1 and over the MFCC feature set after source isolation can also be observed by the across-genre average vocal precision v/s recall curves in Fig. 1. A detailed analysis of the genre-specific performance of each feature set follows. The feature set C2 shows relatively high performance for the Western, Greek and Bollywood genres as compared to the Hindustani and Carnatic genres. This can be attributed to the presence of normal syllabic singing in the former and long duration vowel and melismatic singing in the latter. The relatively high performance of this feature set in the Bollywood genres where the instruments are mainly pitch-continuous corroborates with the static timbral characteristics of these instruments despite their continuously changing pitch. Table 5 shows reduced results for different feature sets and their combinations as compared to Table 4. This is due to the pitch detection errors inherent in the fully-automatic F0 extraction system. However the general trends across different feature sets hold for this case as well, with the results of Table 4 providing an upper limit on performance achievable with a better pitch detector.

The suitability of C2 and C3 for specific signal conditions can be understood from Fig. 2a and 2b, which show spectrograms of 30-sec excerpts from the Bollywood and Hindustani genres respectively. For both plots the left half contains

Table 5. %Correct classification for different genres using fully automatic predominant F0 extraction; Bold indicates best achieved in each genre

Genre	I	II	III	IV	V	Total
Baseline	77.2	66.0	65.6	82.6	83.2	74.9
MFCCs(after source isolation)	76.9	72.3	70.0	78.9	83.0	76.2
C1	81.1	67.8	74.8	78.9	84.5	77.4
C2	72.0	75.9	77.1	67.5	65.8	71.7
C3	65.6	69.1	62.1	77.1	66.0	68.0
C1+C2	**82.9**	78.5	82.8	79.5	**85.0**	81.7
C1+C3	81.5	72.9	77.6	**83.9**	84.9	80.2
C1+C2+C3	82.1	**81.1**	**83.5**	83.0	84.7	**82.8**

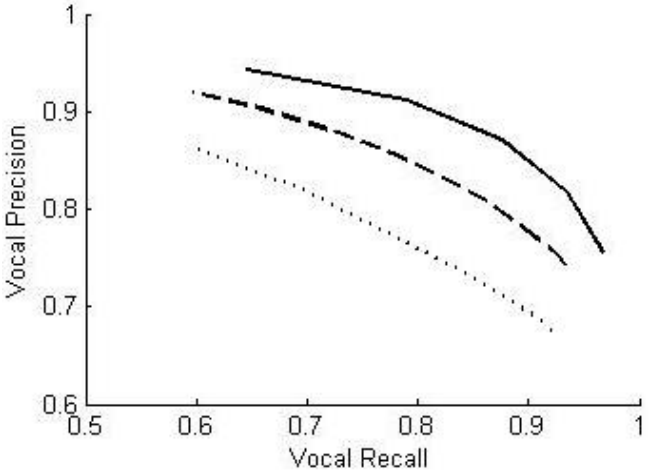

Fig. 1. Avg. Vocal Recall v/s Precision curves for different feature sets - baseline, C1 and C1+C2+C3 (using semi-automatic F0 extraction) across genres in the Leave 1 song out classification experiment

a dominant melodic instrument and right half contains vocals. In the Bollywood case the instrument is replete with large pitch modulations but the vocal part has mainly flatter note-pitches. However the instrumental timbre is largely invariant while the vocal part contains several phonemic transitions. In this case the timbral dynamic feature set (C2) is able to discriminate between the voice and instrument but the F0-harmonic dynamics feature set fails. The situation is reversed for the Hindustani excerpt since, although the instrumental part still displays timbral invariance, this is also exhibited by the vocal part, which consists of a long single utterance i.e. rapid pitch modulations on the held-out vowel /a/. C2 is ineffective in this case due to the absence of phonetic transitions. However the relative flatness of the instrument harmonics as compared with the vocal harmonics leads to good performance for C3.

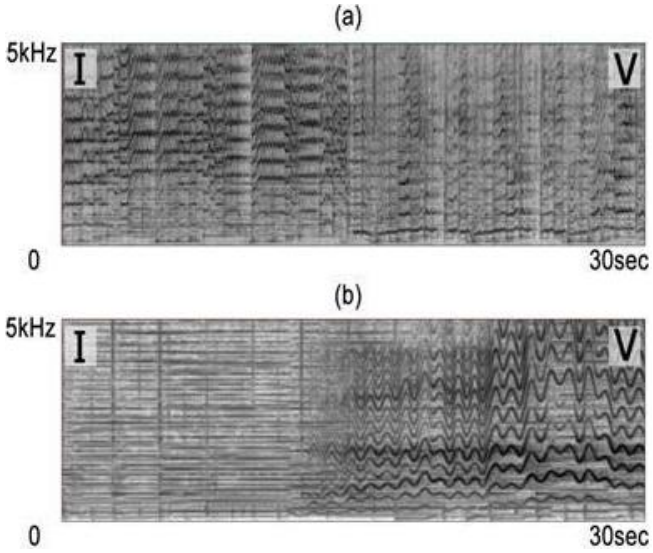

Fig. 2. Spectrograms of excerpts from (a) Bollywood and (b) Hindustani genres. For both excerpts the left section contains a dominant melodic instrument (I) and the right section contains vocals (V).

6 Conclusions

In this paper we have investigated the use of a combination of static and dynamic features for effective detection of lead vocal segments within polyphonic music in a cross-cultural context. Several of the features are novel and have been motivated by considering the distinctive characteristics of singing voice across genres. The introduction of an isolated dominant source spectral representation resulted in a significant increase in the performance of static timbral features in the polyphonic setting over a popularly used baseline feature set. The features representing timbral dynamics and F0-harmonic dynamics were found to provide complementary information for different underlying signal conditions related to singing styles and instrumentation specific to individual genres. While the overall combination of the static and dynamic features was found to result in the highest overall classification performance, individual genre accuracies clearly indicate the value of adapting feature sets to genre-specific acoustic characteristics. Thus commonly available metadata, such as genre, may be effectively utilized in the front-end of a MIR system.

References

1. Berenzweig, A., Ellis, D., Lawrence, S.: Using voice segments to improve artist classification of music. In: 22nd International Conference of Audio Engineering Society, Finland (2002)

2. Li, Y., Wang, D.: Separation of singing voice from music accompaniment for monoaural recordings. In: IEEE Trans. of Audio, Speech Lang. Proc., vol. 15(4), pp. 1475–1487 (2007)
3. Fujihara, H., Goto, M.: Three techniques for improving automatic synchronization between music and lyrics: Fricative detection, filler model and novel feature vectors for vocal activity detection. In: IEEE International Conference on Acoust., Speech, Signal Proc., Las Vegas (2008)
4. Lukashevich, H., Gruhne, M., Dittmar, C.: Effective singing voice detection in popular music using ARMA filtering. In: 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France (2007)
5. Xiao, L., Zhou, J., Zhang, T.: Using DTW based unsupervised segmentation to improve the vocal part detection in pop music. In: IEEE International Conference on Multimedia and Expo, Hannover, Germany (2008)
6. Fujihara et. al.: F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search. In: IEEE International Conference on Acoust. Speech and Sig. Processing, Toulouse, France (2006)
7. Berenzweig, A., Ellis, D.: Locating singing voice segments within music signals. In: IEEE Workshop Applications of Sig. Process. to Audio and Acoust., New York (2001)
8. Maddage, N., Xu, C., Wang, Y.: A SVM-based classification approach to musical audio. In: International Conference on Music Information Retrieval, Baltimore (2003)
9. Ramona, M., Richard, G., David, B.: Vocal detection in music with support vector machines. In: IEEE International Conference on Acoust. Speech and Sig. Process. (2008)
10. Nwe, T., Li, H.: Exploring vibrato-motivated acoustic features for singer identification. In: IEEE Trans. Audio Speech Lang. Process., vol. 15, no. 2, pp. 519–530 (2007)
11. Kim, Y., Whitman, B.: Singer identification in popular music recordings using voice coding features. Proc.5th Intl. Conf. on Music Information Retrieval, Spain (2004)
12. Nwe, T., Li, H.: On fusion of timbre-motivated features for singing voice detection and singer identification. In: IEEE International Conference Acoust., Speech, Signal Proc., Las Vegas (2008)
13. Chou, W., Gu, L.: Robust singing detection in speech/music discriminator design. In: IEEE International Conference Acoust. Speech Sig. Process. (2001)
14. Tzanetakis, G.: Song-specific bootstrapping of singing voice structure. In: IEEE International Conference Multimedia and Expo, Taipei, Taiwan. (2004)
15. Zhang, T.: System and method for automatic singer identification. In: IEEE International Conference Multimedia and Expo, Baltimore. (2003)
16. Vallet, F., McKinney, M.: Perceptual constraints for automatic vocal detection in music recordings. In: Conference Interdisciplinary Musicology (2007)
17. Regnier, L., Peeters, G.: Singing voice detection in music tracks using direct voice vibrato detection. In: IEEE International Conference Acoust. Speech Sig. Process., Taipei, Taiwan (2009)
18. Lidy, T., et. al.: On the Suitability of State-of-the-art Music Information Retrieval Methods for Analyzing, Categorizing and Accessing Non-Western and Ethnic Music Collections. In: Elsevier Signal Processing Special issue on Ethnic Music Audio Documents: from the preservation to the fruition (2009)

19. Mohammed, N. and McG.Squire, D.: Effectiveness of ICF features for collection-specific CBIR. In: 9th International Workshop on Adaptive Multimedia Retrieval (AMR 2011), Barcelona, (2011).
20. Proutskova, P., Casey, M.: You call that singing? Ensemble classification for multi-cultural collections of music recordings. In: 10th International Conference on Music Information Retrieval, Kobe, Japan (2009)
21. Fuhrmann, F., Haro, M., Herrera, P.: Scalability, Generality and Temporal Aspects in Automatic Recognition of Predominant Musical Instruments in Polyphonic Music. In: 10th International Conference on Music Information Retrieval, Kobe, Japan (2009)
22. Fuhijara, H., Goto, M., Kitahara, T., Okuno, H.: A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. In: IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 3, pp. 638–648 (2010)
23. Rao, V., Rao, P.: Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. In: IEEE Trans. Audio Speech and Lang. Process., vol. 18, no. 8, pp. 2145-2154 (2010)
24. Pant, S., Rao, V., Rao, P.: A melody detection user interface for polyphonic music. In: National Conference Comm., Chennai, India (2010)
25. Rao, V., Gaddipati, P., Rao, P.: Signal-driven adaptation for singing voice processing in polyphony. In: IEEE Trans. Audio, Speech, Lang. Process.(accepted with minor mandatory revisions) (2011)
26. Serra, X.: Music sound modeling with sinusoids plus noise. In: Roads, C., Pope, S., Picialli, A., De Poli, G. (Ed), Musical Signal Processing, Swets and Zeitlinger (1997)
27. Rocamora, M., Herrera, P.: Comparing audio descriptors for singing voice detection in music audio files. In: Brazilian Symposium on Computer Music (2007)
28. Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project. In: CUIDADO I.S.T. Project Report (2004)
29. Lagrange, M., Raspaud, M., Badeau, R., Richard, G.: Explicit modeling of temporal dynamics within musical signals for acoustic unit similarity. In: Pattern Recog. Letters, vol. 31, no. 12, pp. 1498-1506 (2010)
30. Burred, J., Robel, A., Sikora, T.: Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. In: IEEE Trans. Audio Speech Lang. Process., vol. 18, no. 3, pp. 663-674 (2010)
31. Aucouturier, J-J., Patchet, F.: The influence of polyphony on the dynamic modeling of musical timbre. Pattern Recog. Letters, vol. 28, no. 5, pp. 654-661 (2007)
32. Sundberg, J.: A rhapsody on perception. In: The Science of Singing Voice, Northern Illinois University Press (1987)
33. Shenoy, A., Wu, Y., Wang, Y.: Singing voice detection for karaoke application. In: Visual Comm. and Image Proc., Beijing, China (2005)
34. Rao, V., Rao, P.: Singing voice detection using predominant pitch. In: InterSpeech, Brighton, U.K. (2009)
35. Hall, M., et. al.: The WEKA Data Mining Software: An Update. In: SIGKDD Explorations, vol. 11, issue 1 (2009)
36. Bouman, C.: Cluster: An unsupervised algorithm for modeling Gaussian mixtures. `http://www.ece.purdue.edu/~bouman`
37. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. In: IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 3 (1998)
38. Markaki, M., Holzapfel, A., Stylianou, Y.: Singing voice detection using modulation frequency features. In: Workshop on Statistical and Perceptual Audition (2008)