Singing Voice Detection in Polyphonic Music using Predominant Pitch

Vishweshwara Rao, S. Ramakrishnan, Preeti Rao

Electrical Engineering Department, Indian Institute of Technology Bombay, Mumbai, India

{vishu, ramakrishnan, prao}@ee.iitb.ac.in

Abstract

This paper demonstrates the superiority of energy-based features derived from the knowledge of predominant-pitch, for singing voice detection in polyphonic music over commonly used spectral features. However, such energy-based features tend to misclassify loud, pitched instruments. To provide robustness to such accompaniment we exploit the relative instability of the pitch contour of the singing voice by attenuating harmonic spectral content belonging to stable-pitch instruments, using sinusoidal modeling. The obtained feature shows high classification accuracy when applied to north Indian classical music data and is also found suitable for automatic detection of vocal-instrumental boundaries required for smoothing the frame-level classifier decisions.

Index Terms: audio segmentation, voice detection

1. Introduction

The automatic identification of audio segments that contain the singing voice is required for several Music Information Retrieval (MIR) applications such as melody extraction, artist identification and voice separation. The last decade has witnessed a significant increase in research interest in the Singing Voice Detection (SVD) problem. SVD is usually viewed as an audio classification problem where features that distinguish vocal regions from purely instrumental regions in music are fed to a classifier previously trained on manually labeled data.

Various classifiers, such as Gaussian mixture models (GMM) [1], support vector machines (SVM) [2] and multi-layer perceptrons (MLP) [3]) have been used in previous studies related to SVD but, as Berenzweig, Ellis and Lawrence note, "the methods of statistical pattern recognition can only realize their full power when real-world data are first distilled to their most relevant and essential form" [3]. This emphasizes the importance of the design and selection of features that demonstrate the ability to discriminate between singing voice and accompanying instruments. Commonly used features in previous SVD related studies, such as MFCCs [3]-[5], attempt to capture the timbral aspects of musical sounds. However, it is well known that the singing voice occupies a large and diverse timbre-space (due to the continuous variation of vocal tract characteristics with articulation of different phones and also due to variations in vocal tract dimensions across different singers), which may not be completely captured by such features. In this study, rather than timbre, we focus on the energy of the voice, based on the premise that in vocal performances, the voice, when present, is the dominant sound source. In order to extract the energy of the dominant sound source we utilize the pitch estimate as provided by a predominant pitch detection algorithm.

Singing also differs from several musical instruments in its expressivity, which is partially 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 [6]. Within non-western forms of music, specifically Indian classical music, voice pitch inflections and ornamentation are extensively used as they serve important aesthetic and musicological functions. Even when pitch modulations are not intended during singing, involuntary pitch instability in the human voice is always present (called jitter in speech literature and flutter in singing) [7]. 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[1].

There has been limited previous work on applying pitch instability to SVD. Shenoy, Wu and Wang [8] 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. The residual signal is then expected to be the singing voice since its harmonics will only be partially attenuated due to its pitch instability. An adaptive threshold, applied to the energy of the residual signal, is used to make a frame-level vocal/non-vocal decision. The choice of delays used in the design of the inverse comb filters is driven by a musical key detection stage prior to filtering. This stage extracts the key (used interchangeably with tonic), of the song based on a rule-based analysis of detected chords (triads) against the chords present in the major and minor keys. Nwe and Li [9] made use of a bank of band-pass filters to explicitly capture the extent of vibrato within individual harmonics upto 16 kHz. The centre frequencies of these band-pass filters were fixed at the known note frequencies, assuming an equally tempered scale with a tuning frequency of 440 Hz. Both of the above approaches have been designed specifically for western music and are not applicable to non-western forms of music in which pitch simultaneity (chords) is rarely present and the tonic (tuning frequency) is not standardized.

We consider the case of SVD in north Indian classical vocal music (NICM). The typical accompaniment consists of tonal percussion, a drone and, in some cases, a secondary melodic instrument (SMI). The SMI is usually a keyed instrument called the *harmonium* (akin to the accordion) and displays significant spectral harmonic content upto around 5 kHz, similar to the voice. None of the accompanying instruments exhibits continuous pitch modulation.

In the next section two harmonic energy based features, one of which attempts to exploit temporal pitch instability to discriminate the singing voice from accompanying instruments, are described. Section 3 comparatively evaluates the frame-level classification performance of the proposed features versus other feature sets (MFCCs, Spectral features [10]) using a Gaussian Mixture Model (GMM) classifier. Frame-level decisions of a statistical classifier for SVD are known to be noisy due to the local variability in the underlying signal [4], [5]. Post-processing by smoothing between known or detected boundaries typically improves the frame-level

[1] While certain wind instruments, such as the piccolo and oboe, can produce vibrato and fretless stringed instruments, such as the violin, can produce similar pitch modulations as the human voice, these are not considered in this study.

accuracy significantly. In Section 4 we investigate the suitability of the proposed pitch-based features for automatic boundary detection. Classification accuracies after post-processing using detected boundaries are then computed. The last section presents conclusions and future study directions.

2. Pitch-based Feature Extraction

The computation of the two pitch-based features, described in this section, requires the predominant pitch and local spectral information. The predominant pitch (voice pitch during sung phrases) is estimated by a harmonic-matching melody extraction algorithm known to have high accuracy for vocals in polyphony including the typical NICM setting [11]. During instrumental segments the melody extractor will by-and-large output the pitch of the secondary melodic instrument (SMI), if present, or the drone, which is continuously present. The spectral information refers to the local sinusoid frequencies and amplitudes. These are computed by applying a main-lobe shape matching criterion to the magnitude spectrum, computed using an 8192 point FFT (sampling frequency of 22.05 kHz) from a Hamming windowed signal of 40 ms duration, and further refined using parabolic interpolation. Both the predominant pitch estimate and the spectral information are computed at frame intervals of 10 ms.

2.1. Normalized harmonic energy (NHE)

NHE is based on detecting the singing voice by the energy of the predominant pitch source. First, the harmonic energy (HE), defined as the sum of the energies of individual harmonics (multiples of the pitch) in the frequency region up to 5 kHz, is computed as

$$HE = \sum_{i=1}^{N} \left| X\left[f_i \right] \right|^2 \quad (1)$$

where $|X[f_i]|$ is the magnitude of the closest detected sinusoid, with frequency f_i, within a 50 cent neighborhood of the expected location of the i^{th} harmonic. N is the total number of expected harmonics below 5 kHz for the estimated pitch. A value of 5 kHz is used since significant voice harmonics are rarely found above this limit. The HE is normalized by its maximum attained value over a single musical performance.

2.2. Sinusoidal track harmonic energy (STHE)

The NHE feature may fail if a pitched instrument between vocal segments has comparable loudness to the voice. Hence an additional attribute, namely the temporal instability of voice harmonics as compared to the harmonics of a keyed instrument, is considered. This is clearly visible in Figure 1(a), which displays the spectrogram of a mixture of *harmonium* (present throughout) and an Indian classical vocal phrase (starting at 1.2 sec). In order to capture this difference we perform the frame-level energy computation after applying a modified partial tracking algorithm, originally used for sinusoidal modeling [12], and a novel track pruning criterion.

2.2.1. Predominant pitch based partial tracking

We adopt an approach similar to Serra's, in which partial tracking is improved by biasing trajectory formation towards expected harmonic locations based on a pitch detection stage [13]. Specifically, tracks are now indexed by harmonic number and only sinusoids in the 50-cent vicinity of local harmonic frequency estimates (computed from local predominant pitch) can be assigned to the corresponding track. A two semitone threshold is applied on track continuation i.e. a track will 'die' if there does not exist any sinusoid within 2 semitones of the last tracked frequency. In order to resolve competition between multiple sinusoids for being linked to a given track, a novel cost function that takes into account the frequency and amplitude proximity, as opposed to the originally proposed cost function that only uses frequency proximity [12], is then used. The amplitude is also given consideration in the tracking decision as, without this, the high amplitude peaks, which correspond to genuine partials, may be completely missed in favor of other relatively low amplitude peaks. Specifically the k^{th} sinusoid in frame n with frequency ω_n^k and amplitude A_n^k will be linked to a track, whose last tracked sinusoidal frequency and amplitude are ω_{n-1}^m and A_{n-1}^m resp., if it minimizes a cost (J) given by

$$J = \left| (\omega_n^k - \omega_{n-1}^m) * \log(A_n^k / A_{n-1}^m) \right| \quad (2)$$

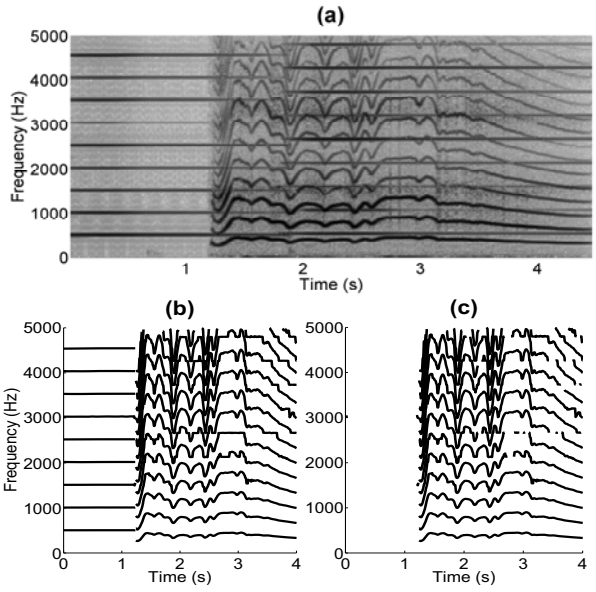

Figure 1: (a) *Spectrogram of harmonium-voice mixture (b) Sinusoidal tracks before and (c) after SD pruning with a 2 Hz threshold.*

Figure 1(b) displays the result of partial tracking for the *harmonium*-voice mixture, whose spectrogram is shown in Figure 1(a). Spurious tracks i.e. shorter than 60 ms, are first deleted. Both the (clearly stable) *harmonium* and voice tracks are well formed. However, some tracks formed at collisions between the voice and *harmonium* harmonics can be observed to approximately follow the latter.

2.2.2. Standard deviation based track pruning

In order to attenuate stable instrument tracks, we use a novel track pruning criteria based on standard deviation (SD). We would like to prune tracks whose SDs are below some threshold (indicating that they belong to a stable-pitch instrument). Computing the SD for individual tracks, however, assumes that the entire track belongs to a single sound source. The intersection of harmonics of different sound sources may result in different segments of the same sinusoidal track belonging to different sources. In such cases, regions of tracks that actually belong to stable-pitch instruments may escape pruning since the SD of the entire track may be greater than the threshold. To avoid such an occurrence, the SD is computed over short non-overlapping track segments of length 200 ms and only track segments whose SD is below a particular threshold (here 2 Hz) are pruned. Figure 1(c) shows

the result of this pruning. The *harmonium* tracks have been erased but the majority of the voice tracks survive.

From the pruned sinusoidal model of the audio signal, the proposed feature, called sinusoidal track harmonic energy (STHE), is computed as the frame-level energy of all surviving sinusoids. As with NHE, the STHE is normalized by its maximum attained value over a single musical performance.

3. Singing voice detection experiment

In this section both the NHE and STHE features are evaluated against previously used features for SVD within the same classification framework on the same set of NICM data.

3.1. Description of training and testing data

A typical NICM performance gradually progresses from slow to fast tempo resulting in the variation of signal characteristics over the performance. With this in mind, the training data contains excerpts from the start, middle and end of polyphonic recordings of seven different north Indian classical vocal performances spanning 23 minutes in total. Three of the performers were female and four were male. Accompanying instruments consist of the drone, percussion and an SMI.

Although the accompanying instruments in NICM are usually soft relative to the voice, occasionally the SMI levels are found to be comparable to the voice. In order to evaluate the features for these two scenarios we use two different datasets. The first set (testing dataset 1) comprises of two songs, sung by a male and female, which are similar to the training dataset. The second dataset (testing dataset 2) contains multi-track data, obtained from live Indian classical vocal performances, down-mixed such that the signal-to-accompaniment ratio of the voice relative to the SMI (*harmonium*) is low (5 dB). Ground truth vocal and instrumental frames are manually marked in the training and testing data. Statistics of the two test datasets are given in Table 1. It can be observed that the vocal segments comprise about 75-80 % of the total duration. The frame-rate over the durations shown in Table 1 is 100/sec.

3.2. Feature sets

Four different feature sets were extracted, for comparative evaluation, from the previously mentioned audio data. The first of these (FS1) comprised the mel-frequency cepstral coefficients (MFCCs). The second feature set (FS2) consisted of spectral features. The optimal number of MFCC coefficients and the best combination of spectral features were determined using a 10-fold cross validation (CV) experiment on the training data. From this experiment it was found that the first 13 MFCC coefficients and a combination of 7 spectral features (spectral flatness, spectral roll-off, spectral centroid, sub-band flux, spectral spread, sub-band energy and sub-band energy ratio) each exhibited the best performance in their respective categories. The third and fourth feature sets (FS3 and FS4) contain the NHE and STHE features respectively.

3.3. Results

All feature sets extracted from the training data were used to train individual GMMs consisting of 4 mixtures per class. This number was again arrived at using a 10-fold CV experiment. Classification performance is quantified in terms of the vocal and instrumental accuracies (recalls) individually. Recall is defined as the ratio of the number of correctly detected frames to the total number of labeled frames for a particular class.

Table 1: *Statistics of vocal (V) and instrumental (I) segments for testing datasets 1 and 2*

Statistic	Testing dataset 1		Testing dataset 2	
	V	I	V	I
Number	75	80	39	43
Avg. duration (sec)	4.4	1.3	5.6	1.2
Total duration (sec)	326.0	106.5	219.9	50.4

Table 2: *Frame-level vocal (V) and instrumental (I) classification results for testing datasets 1 and 2 for all feature sets (FS1-FS4)*

Feature set	Testing dataset 1		Testing dataset 2	
	V (%)	I (%)	V (%)	I (%)
FS1	92.17	56.14	91.61	40.91
FS2	92.38	66.29	87.53	57.40
FS3	89.21	92.09	86.60	45.22
FS4	83.45	90.24	85.62	86.34

Table 3: V*ocal (V) and instrumental (I) classification results after post-processing for testing datasets 1 and 2 for all feature sets (FS1-FS4)*

Feature set	Testing dataset 1		Testing dataset 2	
	V (%)	I (%)	V (%)	I (%)
FS1	97.31	60.70	98.52	31.19
FS2	96.59	69.31	93.44	53.65
FS3	93.41	96.08	92.18	41.24
FS4	89.11	93.63	91.28	81.41

Results comparing the performance of each of the feature sets, in terms of frame-level classification accuracies for testing datasets 1 and 2 are presented in Table 2. For testing dataset 1, FS3 and FS4 show very high instrument classification accuracies and significantly outperform FS1 and FS2. FS4 shows a significantly lower vocal accuracy than FS1 – FS3. This can be attributed to the low frequency harmonics of some steady pitch vocal sounds, which contain most of the energy, but do not survive the SD-based pruning. This could lead to misclassification of these segments as instrumental. For testing dataset 2, the instrumental classification accuracy of FS4 is far superior to all the other feature sets. FS3, which showed high values for instrument accuracy for testing dataset 1, now exhibits degraded performance. This is attributed to the strong presence of the *harmonium*.

As mentioned before, the performance of both NHE and STHE requires that the voice pitch be detected by the melody extractor. However, it can be argued that the voice is not always the pre-dominant sound source. This, in fact, often occurs for testing dataset 2, when the melody extractor output the SMI pitch even though the voice is simultaneously present. In this case we computed these features using the ground truth predominant pitches obtained from the clean voice and *harmonium* signals (available in the multi-track data). Recent enhancements to the melody extractor [14] have shown significantly improved robustness to co-occurring SMI, but are yet to be incorporated in the above feature extraction system.

4. Post-processing

In this section we investigate whether smoothing the frame-level classifier decisions over homogenous segments of audio will result in an improvement in classification accuracy. The boundaries of these segments are automatically detected by the use of the STHE feature in an audio segmentation framework.

4.1. Boundary detection algorithm

We use the framework for audio novelty detection as proposed by Foote [15]. The goal of this method is to generate a novelty function, which will have strong peaks at actual boundaries. The inputs to the novelty function generator will typically be features, which show sharp, but relatively stable, changes at V↔I boundary locations. From these a similarity matrix, a 2-dimensional representation of how similar each frame is to every other frame, is computed. The novelty function is generated by convolving the similarity matrix with a 2-d Gaussian difference kernel along the diagonal. Peaks in the novelty function above a global threshold correspond to significant changes in the audio content and are picked as potential segment boundaries. We then prune detected boundaries using a minimum segment duration criterion i.e. if two boundaries are closer than the duration threshold then the one with the lower novelty score is discarded.

For the input to the boundary detector, we consider only the STHE feature since the frame-level classification results of Section 3.3 show that it demonstrates high classification accuracy even in the presence of strong SMI. The optimal values i.e. ones that give the best trade-off between true boundaries and false alarms, of the difference kernel duration, the novelty function threshold and the min. segment duration are empirically found to be 500 ms, 0.15 and 200 ms resp.

4.2. Results

The performance of the boundary detector is evaluated based on whether it helps improve classification accuracy of any of the feature sets. Grouping of frame-level classification labels can be done either by combining log likelihoods or majority voting [5]. We have found that both methods lead to very similar results and have only shown results for the majority voting process here. Table 3 shows the classification accuracies after grouping frame-level labels over automatically detected boundaries for testing datasets 1 and 2.

For testing data set 1, it can be observed that the vocal and instrumental classification accuracies of all feature sets show a significant increase after post-processing. For testing dataset 2 all feature sets show a significant increase in vocal classification accuracy after grouping but a drop in instrumental accuracy.

5. Conclusions

In this study, two predominant pitch-based features (NHE and STHE) were evaluated against commonly used features (MFCCs and spectral features) for singing voice detection in north Indian classical music (NICM). The STHE specifically attempts to use the temporal characteristic of vocal pitch instability to discriminate between the voice and keyed (stable-pitch) accompanying instruments. The NHE and STHE were found to outperform the local spectral features when evaluated on typical NICM performances. Further, the STHE was also found to be robust to the presence of a loud secondary melodic instrument with spectral content similar to the voice. The STHE feature was then used in an automatic audio novelty detection framework to generate segment boundaries for use in smoothing the frame-level classification labels to achieve high vocal detection accuracies with low false alarm rates.

It is observed that the gains in classification accuracy with the pitch-based energy features come solely from better performance on instrument recall. In fact they appear to do slightly worse on voice recall. Further it can also be argued that the voice is not always the dominant sound source, which questions the use of purely energy features. It was also observed, in separate experiments not reported here, that grouping the frame-level classification decisions, using the STHE feature, over manually marked boundaries results in near perfect vocal and instrumental classification accuracies. This indicates that the frame-level classification is indeed reliable and there is scope for improvement in the boundary detection algorithm. Consequently we intend to investigate a hybrid approach, where timbral and energy features are suitably combined, for classification and boundary detection.

6. Acknowledgements

The authors would like to thank the National Centre for the Performing Arts (NCPA), Mumbai for providing the multi-track Indian classical music data used in testing database 2.

7. References

[1] Chou, W and Gu, L., "Robust singing detection in speech/music discriminator design", in Proc. IEEE Intl. Conf. Acoust., Speech, Signal Proc., 2001.

[2] Maddage, N., Xu, C. and Wang, Y., "An SVM-based classification approach to musical audio", in Proc. 4th Intl. Conf. on Music Information Retrieval, Washington D.C., 2003.

[3] Berenzweig, A., Ellis, D. and Lawrence, S., "Using voice segments to improve artist classification of music", AES 22nd Intl. Conf., Finland, 2002.

[4] Lukashevich, H., Gruhne, M. and Dittmar, C., "Effective singing voice detection in popular music using ARMA filtering", 10th Intl. Conf. Digital Audio Effects (DAFx-07), Bordeaux, 2007.

[5] Li, Y. and Wang, D., "Separation of singing voice from music accompaniment for monoaural recordings", IEEE Trans. Audio, Speech and Lang. Proc., 15(4):1475-1487, 2007.

[6] Sundberg, J., "A rhapsody on perception", in The Science of Singing Voice, Northern Illinois University Press, 1987.

[7] Cook, P., "Pitch, periodicity and noise in the voice," in P. Cook [Ed], Music, Cognition and Computerized Sound, 195-208, MIT Press, 1999.

[8] Shenoy, A., Wu, Y. and Wang, Y. "Singing voice detection for karaoke application," Visual Comm. Image Proc., Beijing, 2005.

[9] Nwe, T. and Li, H., "On fusion of timbre-motivated features for singing voice detection and singer identification", in Proc. IEEE Intl. Conf. Acoust., Speech, Signal Proc., Las Vegas, 2008.

[10] Peeters, G., "A large set of audio features for sound description (similarity and classification) in the CUIDADO project", CUIDADO I.S.T. Project Report 2004.

[11] Rao, V. and Rao, P., "Vocal melody detection in the presence of pitched accompaniment using harmonic matching methods", Proc. of the 11th Intl. Conf. on Digital Audio Effects (DAFx-08), Espoo, Finland, 2008.

[12] Mcaulay, R. and Quatieri, T., "Speech analysis/synthesis based on sinusoid representation", IEEE Trans. Acoustics, Speech and Signal Proc., ASSP-34(4):744-754, 1986.

[13] Serra, X., "Music sound modeling with sinusoids plus noise", in C. Roads, S. Pope, A. Picialli, G. De Poli [Ed], Musical Signal Processing, Swets & Zeitlinger, 1997.

[14] Rao, V. and Rao, P., "Improving polyphonic melody extraction by dynamic programming based multiple F0 tracking", Proc. of the 12th Intl. Conf. on Digital Audio Effects (DAFx-09), Como, Italy, Sept. 2009. Accepted for publication.

[15] Foote, J., "Automatic audio segmentation using a measure of audio novelty," in Proc. IEEE Intl. Conf. Multimedia and Expo (ICME), 2000.