Separation and Classification of Harmonic Sounds for Singing Voice Detection

Martín Rocamora[1] and Alvaro Pardo[2]

[1] Institute of Electrical Engineering - School of Engineering
Universidad de la República, Uruguay
[2] Department of Electrical Engineering - School of Engineering and Technologies
Universidad Católica del Uruguay, Uruguay

Abstract. This paper presents a novel method for the automatic detection of singing voice in polyphonic music recordings, that involves the extraction of harmonic sounds from the audio mixture and their classification. After being separated, sounds can be better characterized by computing features that are otherwise obscured in the mixture. A set of descriptors of typical pitch fluctuations of the singing voice is proposed, that is combined with classical spectral timbre features. The evaluation conducted shows the usefulness of the proposed pitch features and indicates that the approach is a promising alternative for tackling the problem, in particular for not much dense polyphonies where singing voice can be correctly tracked. As an outcome of this work an automatic singing voice separation system is obtained with encouraging results.

1 Introduction

Much research in audio signal processing over the last years has been devoted to music information retrieval, i.e. the extraction of musically meaningful content information from the automatic analysis of an audio recording. This involves diverse music related problems and applications, from computer aided musicology to automatic music transcription and recommendation. Not surprisingly, several research works deal with the singing voice, such as singing voice separation and melody transcription. This kind of research would benefit from a reliable segmentation of a song into singing voice fragments. Furthermore, singing voice segments of a piece are valuable information for music structure analysis.

The goal of this work is to build a computer system for the automatic detection of singing voice in polyphonic music recordings; i.e. classifying each time interval into vocal or non-vocal. The most common approach found in the literature is to extract features from audio signal frames, and then classify them using a statistical classifier. For instance, Mel Frequency Cepstral Coefficientes (MFCC)[3] and Gaussian Mixture Models are applied in [1], which can be considered an example of the standard solution. Several other classifiers have been explored, such as Hidden Markov Models, Artificial Neural Networks and Support Vector Machines. With regards to descriptors, most common features used

[3] Section 3.2 describes these classical audio features traditionally used in speech.

in previous work are different ways of characterizing spectral energy distribution. In [2] a study is conducted which concludes that MFCC are the most appropriate among the features reported to be used for the task. Most recent work began to explore other types of information, such as frequency and amplitude modulation features of the sining voice [3], with a limited success.

In our research on the problem a significant effort has been put into the improvement of the standard solution by considering different acoustic features and machine learning techniques. Results indicate that it seems rather difficult to surpass certain performance bound by variations on this approach. For this reason, in this work we propose a different strategy which involves the extraction of harmonic sound sources from the mixture and their individual classification. This pursues a better characterization of the different musical instruments present in the piece, in a way which is not feasible when dealing with the audio mixture.

2 Harmonic sounds separation

An existing harmonic sound sources extraction front-end is applied, which is very briefly summarized in what follows. It involves a time-frequency analysis, followed by polyphonic pitch tracking and sound sources separation.

The time-frequency analysis is based on [4], in which the application of the Fan Chirp Transform (FChT) to polyphonic music is introduced. The FChT offers optimal resolution for the components of an harmonic linear chirp, i.e. harmonically related sinusoids with linear frequency modulation. This is well suited for music analysis since many sounds have an harmonic structure and their frequency modulation can be approximated as linear within short time intervals. The FChT can be formulated as [4],

$$X(f,\alpha) = \int_{-\infty}^{\infty} x(t)\ \phi'_\alpha(t)\ e^{-j2\pi f \phi_\alpha(t)} dt, \tag{1}$$

where $\phi_\alpha(t) = (1 + \frac{1}{2}\,\alpha\, t)\, t$, is a time warping function. The parameter α is the variation rate of the instantaneous frequency of the analysis chirp.

In addition, based on the FChT analysis, a pitch salience representation called F0gram is proposed in [4], which reveals the evolution of pitch contours in the signal, as depicted in Figures 1 and 3. Given the FChT of a frame $X(f,\alpha)$, salience (or prominence) of fundamental frequency f_0 is obtained by summing the log-spectrum at the positions of the corresponding harmonics,

$$\rho(f_0,\alpha) = \frac{1}{n_H} \sum_{i=1}^{n_H} \log|X(if_0,\alpha)|, \tag{2}$$

where n_H is the number of harmonics considered. Polyphonic pitch tracking is carried out by means of the technique described in [5], which is based on unsupervised clustering of F0gram peaks. Finally, each of the identified pitch contours are separated from the sound mixture. To do this, the FChT spectrum is band-pass filtered at the location of the harmonics of the f_0 value, and the inverse FChT is performed to obtain the waveform of the separated sound.

3 Audio features

3.1 Pitch related features

In a musical piece, pitch variations are used by a singer to convey different expressive intentions and to stand out from the accompaniment. Most typical expressive features are *vibrato*, a periodic pitch modulation, and *glissando*, a slide between two pitches [6]. Although this is by no means an exclusive feature of the singing voice, in a music performance where singing voice takes part as a leading instrument, continuous modulations of its fundamental frequency are of common use. In addition, the accompaniment frequently comprises fixed-pitch musical instruments, such as piano or fretted strings. Thus, low frequency modulations of a pitch contour are considered as an indication of singing voice. Nevertheless, since other musical instruments can produce such modulations, this feature shall be combined with other sources of information.

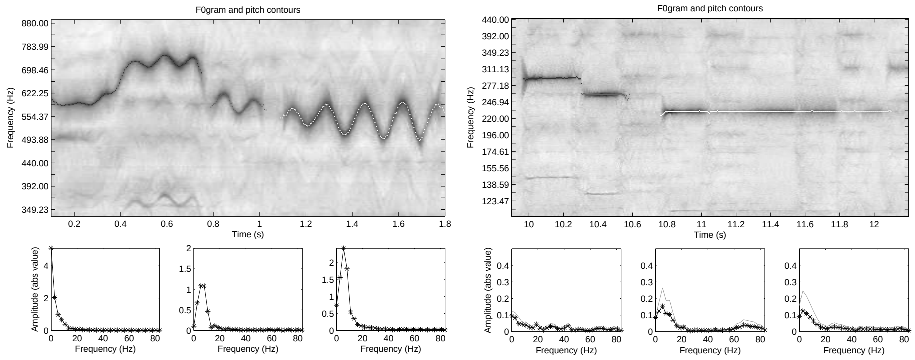

Fig. 1. Vocal notes with vibrato and low frequency modulation (*left*) and saxophone notes without pitch fluctuations (*right*) for two audio files from the MIREX melody extraction test set. Summary spectrum $\tilde{c[k]}$ is depicted at the bottom for each contour.

In order to describe the pitch variations, the contour is regarded as a time dependent signal and the following procedure is applied. The pitch values $f_0[n]$ are represented in a logarithmic scale using a 16th semitone grid. After removing the mean value, the contour is processed by a sliding window and a spectral analysis is applied to each signal frame i using the Discrete Cosine Transform,[4]

$$c[k]^i = \sum_{n=1}^{N} f_0[n]^i \cos\frac{\pi(2n-1)(k-1)}{2N}, \quad k = 1 \ldots N.$$

Frame length is set to $\sim$ 180 ms, to have two frequency components in the range of a typical vibrato. Frames are highly overlapped, using a hop size of $\sim$ 20 ms.

After analysing all the frames, the $c[k]^i$ coefficients are summed up in a single spectrum $\tilde{c}[k]$ as follows. Since we are interested in high values in low frequency,

[4] Normalized by $1/\sqrt{N}$ for $k = 1$ and by $\sqrt{2/N}$ for $k = 2, \ldots, N$.

the maximum absolute value for each frequency bin is taken, namely $\hat{c}[k]$. However, it was observed that this over estimates frames with high low energy values that occasionally arise in noisy contours due to tracking errors. Thus, the median of the absolute value of each frequency bin $\bar{c}[k]$ is also considered and both spectrums are combined as,

$$\tilde{c}[k] = \frac{\hat{c}[k] + \bar{c}[k]}{2}, \quad \text{where} \begin{cases} \hat{c}[k] = \max\limits_{i}\{|c[k]^i|\} \\ \bar{c}[k] = \operatorname*{median}\limits_{i}\{|c[k]^i|\}. \end{cases}$$

Examples of the behaviour of $\tilde{c}[k]$ are given in Figure 1. Then, two features are derived from this spectrum. The low frequency power (LFP) is computed as the sum of absolute values up to 20 Hz ($k = k_L$). Since well-behaved pitch contours do not exhibit prominent components in the high frequency range, a low to high frequency power ratio is considered (PR), which tries to exploit this property,

$$\text{LFP} = \sum_{k=1}^{k_L} \tilde{c}[k], \quad \text{PR} = \frac{\text{LFP}}{\sum_{k_L+1}^{N} \tilde{c}[k]}. \tag{3}$$

Besides, two additional pitch related features are computed. One of them is simply the extent of pitch variation,

$$\Delta f_0 = \max_n\{f_0[n]\} - \min_n\{f_0[n]\}. \tag{4}$$

The other is the mean value of pitch salience in the contour,

$$\Gamma_{f_0} = \operatorname*{mean}_n\{\rho(f_0[n])\}. \tag{5}$$

This gives an indication of the prominence of the sound source, but it also includes some additional information. As noted in [4], pitch salience computation favours harmonic sounds with high number of harmonics, such as the singing voice. Besides, as done in [4], a *pitch preference* weighting function is introduced that highlights most probable values for a singing voice in the f_0 selected range.

3.2 Mel-frequency Cepstral Coefficients

Mel-frequency Cepstral Coefficients (MFCC) are one of the most common features used in speech and music modeling for describing the spectral timbre of audio signals. The implementation of MFCC is based on [7]. Frame length is set to 25 ms, using a Hamming window and a hop size of 10 ms. The signal frame is processed by a filter bank of 40 bands, whose center frequencies are equally-spaced according to the mel scale (approximately linear in log frequency). An FFT is applied and log-power on each band is computed. The elements of these vectors are highly correlated so a DCT is applied and only the 13 lowest order coefficients are retained. Temporal integration is done by computing median and standard deviation of the frame-based coefficients within the whole pitch contour. In order to capture temporal information first order derivatives of the coefficients are also included, for a total of 50 audio features.

4 Classification methods

4.1 Training database

An advantage of the proposed method over the classical polyphonic audio approach is that monophonic audio clips can be used for training, that is music in which only a single musical instrument takes part. There is a lot of monophonic music available and collecting such a database requires much less effort than manually labeling songs. A training database was built based on more than 2000 audio files, comprising singing voice on one hand and typical musical instruments found in popular music on the other.

The procedure for building the database involves the FChT analysis followed by pitch tracking and sound source extraction. Finally the audio features are computed for each extracted sound. In this way, a database of 13598 audio sounds was obtained, where vocal/non-vocal classes are exactly balanced.

4.2 Classifiers and training

First of all, to assess the discrimination ability of the proposed pitch related features, histograms and box-plots are presented in Figure 2 for the training patterns. Although these features should be combined with other sources of information, it seems they are informative about the class of the sound. In addition, using only the proposed pitch related features, a Polynomial Kernel SVM operating as a linear discriminant reaches 93.7% of correctly classified instances when trained and tested by 10-fold cross validation (CV) on the training data.[5]

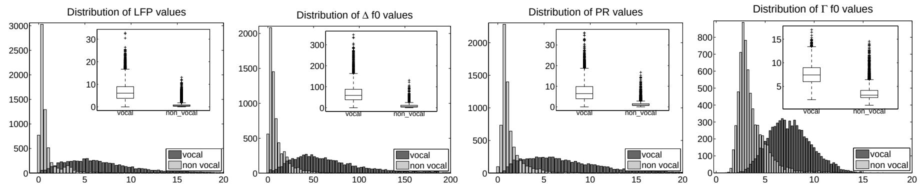

Fig. 2. Analysis of the pitch related features on the training database.

A feature selection experiment on the pitch related features using the Correlation Based selection method and the Best First searching algorithm [8] indicates all features provide some relevant information. An SVM classifier with a Gaussian RBF Kernel was selected for further classification experiments. Optimal values for the γ kernel parameter and the penalty factor C were selected by grid-search. Performance estimated by 10-fold CV on the training set is presented in Table 1 for the MFCC set and by adding the pitch features. Some remarks can be made on these results. Firstly, the performance is encouraging, though it is based on training data. Then, pitch related features seem to contribute to the discrimination between classes (considering that an increase of 1% is more relevant in a high performance level). Finally, the confusion matrix is well balanced, so there seems not to be a significant classification bias.

[5] All classification and selection experiments are performed using Weka software [8].

	MFCC		MFCC + Pitch	
	classified as		classified as	
class	vocal	non-vocal	vocal	non-vocal
vocal	6671	128	6740	59
non-vocal	123	6676	38	6761
performance	98.2%		99.3%	

Table 1. Percentage of correctly classified instances and confusion matrix for each set of features, obtained by 10-fold CV on the training dataset using an SVM classifier.

4.3 Automatic labeling of polyphonic music

In order to deal with polyphonic music the following procedure is applied. The audio is processed with the sound source extraction front-end. Each extracted sound is classified based on MFCC and pitch features. A time interval of the polyphonic audio file is labeled as vocal if any of the identified pitch contours it contains is classified as vocal. This is shown in the examples of Figure 3.

When manual labeling a musical piece, very short pure instrumental regions are usually ignored. For this reason, the automatic labels are further processed and two vocal regions are merged if they are separated by less than 500 ms. It is important to notice that manually generated labels include unvoiced sounds (such as fricative consonants). Although this type of sounds are shrunk when singing, this constitute a systematic source of errors of the proposed approach.

The analysis of pitch fluctuations described in section 3.1 imposes a certain minimum contour length, so a threshold of approximately 200 ms is applied and pitch fluctuations analysis is avoided for shorter segments.

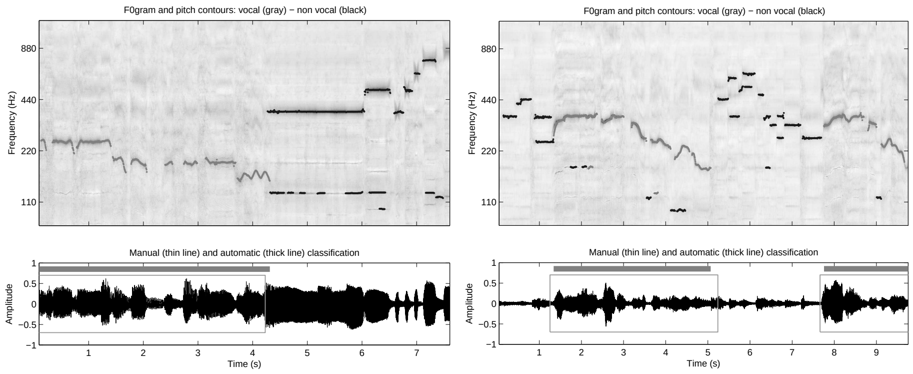

Fig. 3. Examples of automatic vocal labeling. *Left:* Fragment of the song *For no one* by The Beatles. A singing voice in the beginning is followed by a French horn solo. There is a soft accompaniment of bass and tambourine. *Right:* Blues song excerpt from the testing dataset. It comprises singing voice, piano, bass and drums. Singing voice notes are correctly distinguished from piano and bass notes that are also detected.

5 Evaluation and results

An evaluation was conducted to estimate the performance of the singing voice detection approach applied on polyphonic music audio files, and to assess the usefulness of the proposed pitch related features. For this purpose a testing database of 30 manually labeled audio fragments of 10 seconds length was utilized. Music was extracted from Magnatune[6] recordings labeled as: blues, country, funk, pop, rock and soul. A few musical genres were avoided, such as heavy-metal or electronica, because of the high density of sources and the ubiquity of prominent noise-like sounds, what makes the pitch tracking rather troublesome. Classification results are presented in Figure 4. Performance is measured as the percentage of time in which the manual and automatic labeling match. The addition of pitch information produces a noticeable performance increase in the overall results, as well as for almost every file of the database. The performance of the standard approach, MFCC of the audio mixture and an SVM classifier [2], is also reported.

	matching
MFCC	71.3%
MFCC + Pitch	78.0%
MFCC [2]	73.7%

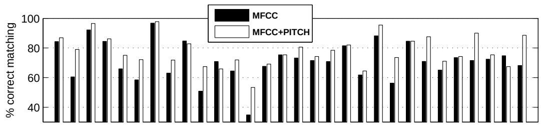

Fig. 4. Classification performance as percentage of time in which the manual and automatic vocal labels match, for both set of features and for the standard approach.

6 Discussion and future work

A novel approach for singing voice detection in polyphonic music was introduced that makes use of an harmonic sound sources extraction front-end [4, 5]. The extracted sounds are then classified based on the classical MFCC coefficients and on some new features devised to capture characteristic of typical singing voice pitch contours. Results obtained indicate that the proposed features provide additional information for singing voice discrimination. Besides, an advantage of the sound source separation approach is that it enables the application of other acoustic features for describing isolated sounds that are otherwise obscured in the polyphonic sound mixture (e.g. the estimation of formants will be tackled in future work). Although the sound sources extraction introduces new challenges, the proposed approach seems a feasible alternative to the standard solution.

As an interesting outcome of this research an automatic singing voice separation system is obtained which yields very promising results, mainly for not much dense polyphonies where singing voice can be correctly tracked. Examples of the automatic singing voice separation are provided in Figure 5, for the audio excerpts previously introduced.[7] The improvement of this tool and its application to different music scenarios are the most appealing ideas for future research.

[6] `http://magnatune.com/`

[7] Audio and more examples at `http://iie.fing.edu.uy/~rocamora/mscthesis/`

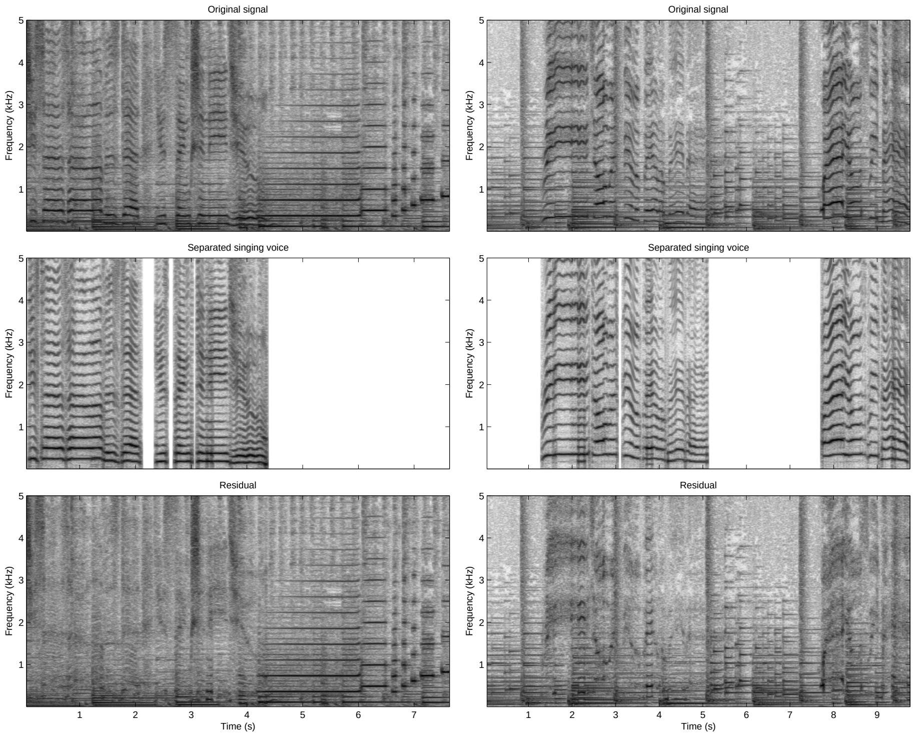

Fig. 5. Automatic singing voice separation for the audio introduced in Figure 3.

References

1. Tsai, W.H., Wang, H.M.: Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signal. IEEE Transactions on Speech and Audio Processing (2006) Vol. 14, No 1
2. Rocamora, M., Herrera, P.: Comparing audio descriptors for singing voice detection in music audio files. In: Brazilian Symposium on Computer Music, 11th. São Paulo, Brazil. (2007)
3. Regnier, L., Peeters, G.: Singing voice detection in music tracks using direct voice vibrato detection. In: ICASSP IEEE Int. Conf. (2009) 1685 –1688
4. Cancela, P., López, E., Rocamora, M.: Fan chirp transform for music representation. In: Int. Conf. on Digital Audio Effects, 13th DAFx-10. Graz, Austria. (2010)
5. Rocamora, M., Cancela, P.: Pitch tracking in polyphonic audio by clustering local fundamental frequency estimates. In: Brazilian AES Audio Engineering Congress, 9th. São Paulo, Brazil. (2011)
6. Sundberg, J.: The science of the singing voice. De Kalb, Il., Northern Illinois University Press (1987)
7. Ellis, D.P.W.: PLP and RASTA (and MFCC, and inversion) in Matlab (2005)
8. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco (2005)