Comparing audio descriptors for singing voice detection in music audio files

Martín Rocamora[1]*, **Perfecto Herrera**[2]

[1]Instituto de Ingeniería Eléctrica – Facultad de Ingeniería de la Universidad de la República
Julio Herrera y Reissig 565 – (598) (2) 711 09 74, Montevideo, Uruguay

[2]Music Technology Group – Universitat Pompeu Fabra
Pg. Circumval·lació 8 – 08003 Barcelona, Spain

rocamora@fing.edu.uy, pherrera@iua.upf.edu

Abstract. *Given the relevance of the singing voice in popular western music, a system able to reliable identify those portions of a music audio file containing vocals would be very useful. In this work, we explore already used descriptors to perform this task and compare the performance of a statistical classifier using each kind of them, concluding that MFCC are the most appropriate. As an outcome of our study, an effective statistical classification system with a reduced set of descriptors for singing voice detection in music audio files is presented. The performance of the system is validated using independent datasets of popular music for training, validation and testing, reaching a classification performance of 78.5% on the testing set.*

1. Introduction

One of the most memorable and representative features of a song in popular western music is the singing voice melody. For this reason, a lot of research being carried out on Music Information Retrieval deals with it (singer identification, singing voice separation, singing voice melody transcription, query by humming, lyrics transcription, etc). This kind of research would benefit from a reliable segmentation of a song into singing voice fragments. Furthermore, singing voice detection has many other applications in audio processing, such as to generate "lyrics-centred" summaries of songs, or to apply a certain singing voice filter (e.g. de-esser, de-breather) only to the automatically identified vocal fragments of a music audio file.

In this paper we address the problem of processing a music audio file and segmenting it into fragments containing singing (with or without musical accompaniment) and purely instrumental (non-singing) content, referred as vocal and non-vocal respectively. Contextual information must be considered when partitioning to ignore pauses (short non-vocal fragments) during a singing melody or to correct classification errors. One of the most troublesome characteristics of the problem is the variability of music, both with regards to the singing performance and to the accompaniment. In order to limit the scope of the problem, in this work we focus on popular music (in particular music genres such as rock, pop, folk, funk and jazz).

To address the singing voice detection problem we can take advantage of the knowledge on research fields such as musical instruments classification [Martin, 1999] [Herrera et al., 2006] and speech processing [Rabiner and Schafer, 1978] [Chilton, 1999]. The former studies our ability to distinguish different musical instruments. Singing voice detection can be considered a particular case of musical instruments classification in complex mixtures, so many features used in this field may be useful for characterizing vocal and non-vocal portions of a song. Given the similarities between speech and singing voice it is reasonable to apply techniques and descriptors used to segment and recognize speech to singing voice problems. Speech/music discrimination, singing voice separation and singer identification are closely related problems, as many systems developed to perform these tasks try to identify those fragments of the audio file containing vocals.

Although singing voices resembles speech to a certain extent there are significant differences between them that need to be taken into account. To sing a melody line with lyrics it is usually necessary to stretch the voiced sounds and shrink the unvoiced sounds to match notes durations.[1] For this reason, singing voice is more than 90% voiced, where as speech is only aproximately 60% voiced [Cook, 1990]. The majority of the singing voice energy falls between 200 Hz and 2000 Hz, but in speech the unvoiced sounds

*Supported by Comisión Sectorial de Investigación Científica, UdelaR.

[1] Vocal sounds are usually divided by speech researchers in voiced and unvoiced. The vibration of the vocal folds produces quasi-periodic sound referred to as voiced. Those vocal sounds generated by the turbulence of air against the lips or tongue (such as the consonants "s" or "f") are known as unvoiced and its waveform appears random though with some limited spectral shaping [Chilton, 1999].

are more common and tend to raise this energy limit up to 4000 Hz. Speech has a characteristic energy modulation peak around the 4 Hz syllabic rate usually considered as an evidence of it presence in automatic speech processing [Scheirer and Slaney, 1997]. With regards to pitch, the pitch contour of the singing voice tends to be piece-wise constant with abrupt pitch changes in between, while in natural speech pitch slowly drifts down with smooth pitch changes. Besides this, speech pitch is normally between 80 to 400 Hz whereas singing has a wider pitch range that can reach 1400 Hz in a soprano singer [Li and Wang, 2005] [Gerhard, 2002]. Morover, singing voice is highly harmonic, that means that the partials of the sound are located at multiples of the fundamental frequency. Several musical instruments are also harmonic, so some partials of the singing voice are overlapped with those from the accompainment. Additionally, a known feature of operatic singing is the precense of an additional formant (resonance of the vocal tract), called the singing formant, in the frequency range of 2000 to 3000 Hz, that enables the voice to stand out from the accompainment [Sundberg, 1987]. However, the singing formant does not exist in many other types of singing such as the ones in pop or rock.

The approach proposed in this work is to build a statistical classifier trained on descriptors of accompanied singing voices and accompaniments alone. Much effort is put in the study of descriptors taking into account the great majority of those reported in previous work and comparing them in equivalent conditions. The rest of this paper is organized as follows. In the next section the different approaches proposed to address the singing voice detection problem are summarized. Section 3 describes the method we applied for the selection of descriptors and the classification approach. Experimental results are presented in section 4. The paper ends with a discussion on the present work and the main conclusions.

2. Previous work

The common procedure followed for partitioning a song into vocal and non-vocal portions is to extract feature paremeters from audio signal frames (near stationary block of samples) at each few tens milliseconds, and then to classify them to one of each class using a threshold method or a statistical classifier.

Threshold methods tend to be simple but need descriptors that clearly discriminate between classes in order to be successful. The methods proposed to classify audio frames into vocal and non-vocal apply a threshold on only one descriptor [Maddage et al., 2004] [Shenoy et al., 2005] or compute different descriptors and apply a set of thresholds on them [Zhang, 2002]. On the other hand, statistical classifiers are trained using accompanied singing voices and pure instrumentals and can learn complex boundaries between classes combining several descriptors. This has the drawbacks of a certain amount of time spent on training and the difficulty of obtaining ground truth annotations. In the singing voice detection problem, finding the exact boundaries over the entire song can be time-consuming, and sometimes could be difficult in case of slow decays or masking. Moreover, special attention has to be paid to ensure generalization beyond the training set avoiding overfitting. Several statistical classifiers have been explored to address the problem of singing voice detection, such as Hidden Markov Models (HMM) [Berenzweig and Ellis, 2001] [New et al., 2004], Gaussian Mixture Models (GMM) [Tsai and Wang, 2006] [Li and Wang, 2007], Artificial Neural Networks (ANN) [Berenzweig et al., 2002] [Tzanetakis, 2004] and Support Vector Machines (SVM) [Maddage et al., 2003] [Maddage et al., 2004].

The short-term classification of each signal frame considers only local information so it is prone to errors, and the classification obtained is typically noisy changing from one class to the other. For this reason, usually long-term information is introduced, by smoothing the classification [Tsai and Wang, 2006] or by partitioning the song into segments (much longer than frames) and assigning the same class to the whole segment. This partitioning of the song is performed based on tempo [New et al., 2004] [Maddage et al., 2003], chord change [Maddage et al., 2004] or spectral (timbre) change [Li and Wang, 2007]. More global temporal aspects of a song are also taken into account by modeling the song structure (intro, verse, chorus, etc) by means of a HMM [New et al., 2004].

With regards to descriptors, research on musical instruments classification has demonstrated the importance of temporal and spectral features [Martin, 1999] [Herrera et al., 2006], and the speech processing field has contributed on well known techniques (such as Linear Prediction) to compute voice signal attributes [Rabiner and Schafer, 1978] [Chilton, 1999]. A broad group of descriptors has been used for the purpose of singing voice detection. Singing voice carries the main melody and the lyrics of a popular song, so it is is usually one of the most salient instruments of the mixture. Therefore, vocal frames can be identified by an energy increase of the signal, and energy or power descriptors are often used [Zhang, 2002] [Tzanetakis, 2004] [New et al., 2004] [Shenoy et al., 2005] [Maddage et al., 2003]. The timbre of different instruments is partially dictated by its spectral characteristics and when new sounds enters a mixture they usually introduce significant spectral changes. Among the descriptors computed to represent this are Mel Frequency Cepstral Coefficients (MFCC) [Tsai and Wang, 2006] [Li and Wang, 2007]

[Maddage et al., 2003], Linear Prediction Coefficients (LPC) (warped LPC or perceptually derived LPC) [Berenzweig and Ellis, 2001] [Berenzweig et al., 2002] [Kim and Whitman, 2002] [Maddage et al., 2003], Log Frequency Power Coefficients (LFPC) [New et al., 2004], and spectral Flux, Centroid and Roll-Off [Zhang, 2002] [Tzanetakis, 2004]. The delta coefficients of the previous features or variances of them are also used to capture temporal information [Berenzweig et al., 2002]. As stated previously singing voice is highly harmonic, thus the harmonicity of the signal, usually computed as an Harmonic Coefficient (HC), is used as a clue for singing voice detection [Chou and Gu, 2001] [Zhang, 2002] [Kim and Whitman, 2002].

Regarding harmonicity a pre-processing technique was proposed that consist in filtering the signal by an inverse comb filter to attenuate the harmonic sounds in the mixture [New et al., 2004] [Shenoy et al., 2005]. Harmonic accompaniment instruments have a very regular harmonic structure and can be partially removed by this filtering. Although being harmonic, singing voice has some features (vibrato, intonation) that deviate the frequency of partials from perfectly harmonic and cause it to remain after filtering.

The following is a summary of the most relevant research work on singing voice detection in chronological order. To our knowledge, the first work that focused and described the problem of locating the singing voice segments in music signals was [Berenzweig and Ellis, 2001]. Posterior probability features and their statistics are derived from an ANN trained on a phone classes database to work with speech. A HMM with two states is used to discriminate singing from accompaniment. Close in time, an approach for singing detection in speech/music discrimination is described in [Chou and Gu, 2001]. It employs a set of features that comprises MFCC as well as HC, 4Hz modulation and energy based features to train a GMM.

Following this early work, new proposals for singing detection were suggested. Artist classification is improved in [Berenzweig et al., 2002] by using only voice segments detected with a multi-layer perceptron that is fed with Perceptual LPC (PLPC), plus deltas and double deltas. An harmonic sound detector is presented in [Kim and Whitman, 2002] to identify vocal regions of an audio signal for singer identification. It works under the hypothesis that most harmonic sounds correspond to regions of singing. By means of an inverse comb filter bank, the signal is attenuated and the Harmonicity is computed as the ratio between the total energy in a frame over the energy of the most attenuated signal. The automatic singer identification system described in [Zhang, 2002] identifies the starting point of the singing voice in a song using energy features, zero-crossing rate (ZCR), HC and Spectral Flux and classifying with a set of thresholds. In [Maddage et al., 2003] a study which aims to establish if there are significant statistical differences between vocal music, instrumental music and mixed vocal and instruments is described. Descriptors set contains LPC, LPC derived Cepstrum, MFCC, Spectral Power, Short Time Energy and ZCR. Classification performance of a SVM is shown to be superior to an ANN and a GMM.

Subsequent research explored some other descriptors and classification approaches. A technique for singing voice detection is proposed in [Maddage et al., 2004]. Musical signals are segmented into beat-length frames using a rhythm extraction algorithm, the Fast Fourier Transform (FFT) of each frame is calculated, and after filtering to a narrow bandwith containing mostly voice, another FFT is applied (Twice Iterated Composite Fourier Transform, TICFT). Based on a threshold on the energy of the TICFT, singing voice frames are separated from instrumental frames. Performance is improved by some frame-correction rules based on chord pattern changes. In [Tzanetakis, 2004] singing voice detection is performed by a bootstrapping process that consists in manually annotating a few random fragments of the song being processed to train a classifier. The feature set includes Mean Relative Energy, and Mean and Standard Deviation of Spectral Centroid, Roll-off, Flux and Pitch. Different classifiers are tested, being Logistic Regression and ANN those which performed best. In the work described in [New et al., 2004] , based on the observation that a more regular harmonic structure is present in non-vocal sections as compared to vocal sections, a key estimation is performed to attenuate the harmonics of the spectrum and LFPC are computed. The energy distribution of the LFPC shows that vocal segments have relatively higher energy values in the higher frequency bands. Classification is done with a multi-model HMM that models the different sections of a typical song structure (intro, verse, chorus, bridge, outro). A classification refinement is provided by a verification step based on classification confidence and an automatic bootstrapping process (similar to that proposed in [Tzanetakis, 2004]). The work in [Shenoy et al., 2005] also addresses the singing voice segmentation problem by estimating the key of the song in order to perform an inverse comb filtering to attenuate the harmonic sounds. Singing voice is retained after filtering due to vibrato and intonation. The energy in different frequency sub-bands is computed and the highest energy frames are classified as vocal.

Most recent works use MFCC as feature vectors and GMM as classifiers. A vocal/non-vocal detection algorithm is proposed in [Tsai and Wang, 2006] applied to singer recognition. The class of each frame is hypothesized according to log-likelihoods and the final decision is made at homogeneous segments. In [Li and Wang, 2007], the problem of singing voice separation from music accompaniment is addressed and a singing voice detection procedure is used. The audio is partitioned by detecting large spectral changes, and segments are classified according to the log-likelihoods of all the frames of a portion.

3. Method

3.1. Databases

Independent datasets of popular music recordings were used for training, validation and testing. Audio of all datasets is stereo at a sampling rate (f_s) of 44.1 kHz. The training database was build extracting short audio excerpts from music recordings and manually classifying them into vocal and non-vocal. Three different excerpt-length sets were constructed of 0.5, 1 and 3 seconds length. Each set contains 500 instances of each class. Music utilized belongs to a music genres database comprising alternative, blues, classical, country, electronic, folk, funk, heavy-metal, hip-hop, jazz, pop, religious, rock and soul. In addition, approximately 25% of pure instrumental and a capella music was also added. The validation database consists of 63 fragments of 10 seconds that were manually annotated. Music was extracted from Magnatune[2] recordings belonging to similar genres as the ones used for training. Once the features and the classifier were set and its parameters finely tuned, an independent evaluation was conducted on a testing database of 46 manually annotated songs, for a total duration of 3 hours. Music in this set comprises 7 different singers performing in genres that were used for training and validation.

3.2. Descriptors implemented

Based on the existing literature the following timbre descriptors were implemented: Mel Frequency Cepstral Coefficients (MFCC), Perceptually derived LPC (PLPC), Log Frequency Power Coefficients (LFPC) and Harmonic Coefficient (HC). In addition, a general purpose musical instruments classification feature set was built, including Spectral Centroid, Roll-off, Flux, Skewness, Kurtosis and Flatness. Pitch was also included, being the only non-spectral feature reported that was considered relevant (4Hz modulation is appropriate for speech but not for singing [Chou and Gu, 2001], ZCR strongly correlates with the Spectral Centroid [Herrera et al., 2006] and other power or energy features can be regarded as variants of spectral descriptors such as LFPC).

Audio signal is processed in frames of 25 ms using a Hamming window and a hop size of 10 ms. Considering that the majority of energy in the singing voice falls between 200 Hz and 2000 Hz [Kim and Whitman, 2002], the frequency range is established in 200 Hz to 16 kHz for all spectral descriptors.

Implementation of MFCC derives 13 coefficients from 40 mel scale frequency bands for each signal frame. An FFT is applied to each signal frame and the magnitude spectrum is obtained by taking the absolute value. After that, the magnitude spectrum is processed by a filter bank whose center frequencies are spaced according to the mel scale. Then, energy on each band is computed and the logarithm is taken. The elements of these vectors are highly correlated so a Discrete Cosine Transform (DCT) is applied to finally obtain the MFCC.

Some psychoacoustic concepts are introduced in the PLP analysis technique that make it more consistent with human hearing in comparison with conventional LP analysis. PLP coefficients are obtained by a critical band integration of the signal, followed by equal loudness weighting and intensity to loudness conversion. Finally, a Linear Prediction analysis of order 12 is applied. Both MFCC and PLPC are implemented using [Ellis, 2005][3].

To derive LFPC the signal frames are passed through a bank of 12 band-pass filters spaced logarithmically, and the coefficient for each band is obtained by computing the power of the band divided over the band bandwith and expressed in decibels [New et al., 2004], as follows,

$$\mathrm{LFPC}_t(m) = 10\ \log_{10}\left|\frac{S_t(m)}{N_m}\right|, \quad S_t(m) = \sum_{k=f_{m-1}}^{f_m} X_t(k)^2, \quad m = 1, 2, \ldots, 12$$

where t is the frame number, m indicates the band number, $S_t(m)$ is the power of the band, N_m is the number of spectral components in the band, $X_t(k)$ is the k^{th} spectral component of the signal frame, and f_m are the indexes of the band boundaries corresponding to frequencies spaced logarithmically from 200Hz to 16kHz as represented in figure 1.

Implementation of HC follows the procedure described in [Chou and Gu, 2001], where temporal and spectral autocorrelation of the signal frame are computed (TA and SA respectively), and the HC is obtained as the maximum of the sum of the autocorrelation functions, $\mathrm{HC}_t = \max_\tau\ [\ \mathrm{TA}_t(\tau) + \mathrm{SA}_t(\tau)\ ]$, where t is the frame number, and τ is the temporal delay. Temporal and spectral autocorrelation functions

[2] http://magnatune.com/

[3] http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

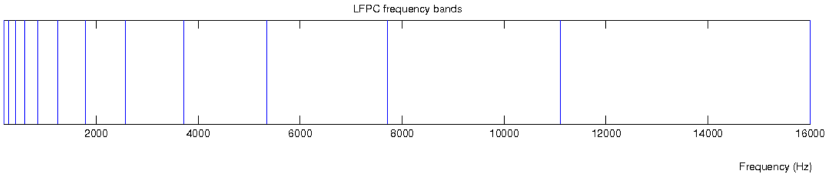

Figure 1: Frequency bands used for LFPC computation.

are calculated as,

$$\mathrm{TA}(\tau) = \frac{\sum_{n=1}^{N-\tau} \bar{x}_t(n)\,\bar{x}_t(n+\tau)}{\sqrt{\sum_{n=1}^{N-\tau} \bar{x}_t^2(n) \sum_{n=1}^{N-\tau} \bar{x}_t^2(n+\tau)}}, \quad \mathrm{SA}(\tau) = \frac{\sum_{k=1}^{\frac{M}{2}-k_\tau} \bar{X}_t(k)\,\bar{X}_t(k+k_\tau)}{\sqrt{\sum_{k=1}^{\frac{M}{2}-k_\tau} \bar{X}_t^2(k) \sum_{k=1}^{\frac{M}{2}-k_\tau} \bar{X}_t^2(k+k_\tau)}}$$

where $x_t(n)$ is a signal frame of N samples, $X_t(k)$ is the magnitude spectrum of the signal frame computed by an M point FFT, $\bar{x}_t(n)$ and $\bar{X}_t(k)$ are the zero mean versions of $x_t(n)$ and $X_t(k)$, and $k_\tau = \frac{M}{\tau f_s}$ is the frequency bean index that corresponds to the time delay τ.

Spectral Flux, Roll-off, Centroid, Skewness, Kurtosis and Flatness are implemented based on [Herrera et al., 2006]. The Spectral Flux (SFX) is a measure of local spectral change, and it is computed as the spectral difference of two consecutive frames as,

$$\mathrm{SFX}_t = \sum_{k=1}^{\frac{M}{2}} \left(\hat{X}_t(k) - \hat{X}_{t-1}(k) \right)^2$$

where $\hat{X}_t(k)$ is the energy normalized magnitude spectrum of the signal frame. Spectral Roll-off is computed as the frequency index R below which the majority of the spectral energy is concentrated ($\gamma = 0.85$ is used),

$$\sum_{k=1}^{R} X_t(k)^2 \leq \gamma \sum_{k=1}^{\frac{M}{2}} X_t(k)^2.$$

The rest of the spectral descriptors consider the spectrum as a distribution, which values are the frequencies and the probabilities are the normalized spectral amplitude, and compute measures of the distribution shape. The Spectral Centroid is the barycenter or center of gravity of the spectrum and is computed as,

$$\mathrm{SC} = \frac{\sum_{k=1}^{\frac{M}{2}} k\; X_t(k)}{\sum_{k=0}^{\frac{M}{2}-1} X_t(k)}.$$

The Skewness is a measure of the asymmetry of a distribution around its mean, while the Kurtosis is a measure of the flatness of a distribution around its mean. They are computed as the normalized moments of order 3 and 4 respectively by,

$$\mathrm{SS} = \frac{\sqrt{\frac{M}{2}} \sum_{k=1}^{\frac{M}{2}} (X_t(k) - \tilde{X}_t)^3}{\left(\sum_{k=1}^{\frac{M}{2}} (X_t(k) - \tilde{X}_t)^2 \right)^{\frac{3}{2}}}, \quad \mathrm{SK} = \frac{\frac{M}{2} \sum_{k=1}^{\frac{M}{2}} (X_t(k) - \tilde{X}_t)^4}{\left(\sum_{k=1}^{\frac{M}{2}} (X_t(k) - \tilde{X}_t)^2 \right)^2} - 3.$$

where $\tilde{X}_t$ is the mean value of the magnitude spectrum $X_t(k)$. The Spectral Flatness is a measure of how flat or similar to white noise the spectrum is. It is computed as the ratio of the geometric mean to the arithmetic mean of the spectrum,

$$\mathrm{SF} = \frac{\sqrt[\frac{M}{2}]{\prod_{k=1}^{\frac{M}{2}} X_t(k)}}{\frac{1}{\frac{M}{2}} \sum_{k=1}^{\frac{M}{2}} X_t(k)}.$$

A low value indicates a tonal signal, while for a noisy signal the value is close to 1. It is typically computed for several frequency bands. We used the following four overlapped frequency bands: 200 to 1000 Hz, 800 to 2500 Hz, 2000 to 3500 Hz and 2500 to 5000 Hz.

Reliable pitch estimation in polyphonic music signals remains up to the moment a very complex open problem, so for the sake of simplicity pitch estimation is performed by applying a monophonic fundamental frequency estimation algorithm based on [A. de Cheveignè, H. Kawahara, 2002]. This has the drawback of an unreliable estimation, noisy and prone to octave errors due to the many sounds present at the same time. The implemented algorithm uses the Difference Function (DF), a variation of the autocorrelation function that calculates the difference between the signal frame and a delayed version of it,

$$DF_t(\tau) = \sum_{n=1}^{N-\tau} (x_t(n) - x_t(n+\tau))^2 .$$

For a periodic signal this function is zero at values of τ multiples of the signal period, while in case of quasiperiodic signal it has minimuns close to zero. The pitch of the signal is estimated as the inverse of the delay value of the first minimun.

As stated previously, the audio signal is processed in overlapped frames, so the audio features computed describe the signal at each 10 ms. To take into account descriptors information over several consecutive frames, an audio fragment is considered, of 0.5, 1 or 3 seconds. Mean, median, standard deviation, skewness and kurtosis are extracted for this fragments. Additionally, deltas and double deltas are computed for each descriptor coefficient and the same statistical measures are calculated.

3.3. Validation procedures

To compare descriptors, in order to select them, classification performance on the training dataset is considered. The training database is composed of audio excerpts, each of them labelled as belonging to either one class. Performance is computed as the percentage of correctly classified instances using 10-fold cross-validation (CV). Different classifiers provided by [Witten and Frank, 2005][4] were considered and, after detailed comparisons, SVM was selected (see table 3).

The validation database is used to adjust system parameters such as the fragment length to be used to process the music audio file, or the classifier options. It is also used for evaluating the inclusion of different post-processing strategies to the system. On the other hand, testing database provides an independent dataset to test the final system performance. In both cases, music audio files manually labelled must be processed. The adopted procedure divides the file into 50% overlapped fragments, descriptors for each fragment are computed at each 10 ms and statistical measures over the whole fragment are calculated. After that, each fragment is classified as vocal or non-vocal. Finally, vocal and non-vocal regions of the audio file are build considering that the class of each fragment extends from its center to half hop-size on either side (due to overlapping). Performance is computed as the percentage of time that the classification and the manual annotation coincides. It is important to note that, due to the rough discretization of the audio in fragments, even in a correctly classified case there is always a small divergence between the annotated and the computed boundaries (see figure 2).

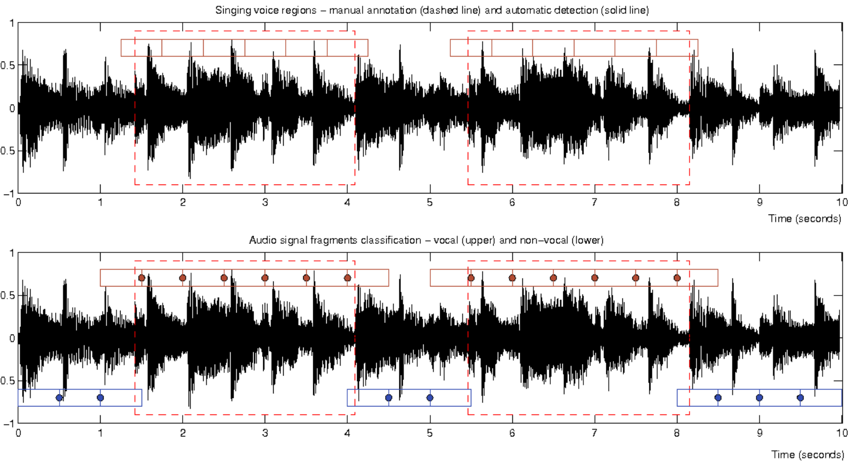

Figure 2: Singing voice detection example. Manual annotation (dashed line) versus automatic detection (solid line) is shown at the top. Audio signal is divided into 50% overlapped fragments of 1 second length. The classification of the fragments into vocal (upper) and non-vocal (lower) is depicted at the bottom. This rough discretization of the audio in fragments produces a small divergence between the annotated and the computed boundaries of the vocal regions. Although this example could be considered a correctly classified case, performance computed as the percentage of time that the classification and the manual annotation coincides is 93.3%.

[4] http://www.cs.waikato.ac.nz/ml/weka/

3.4. Selection of descriptors

As stated previously, for each descriptor coefficient (as well as deltas and double deltas), mean, median, standard deviation, skewness and kurtosis are computed. Adding irrelevant attributes to a dataset is not practical and often confuses a machine learning system [Witten and Frank, 2005]. Thus, feature selection is performed to discard less significant descriptors among each category (e.g. MFCC). For this task a correlation-based feature subset selection method provided by [Witten and Frank, 2005] is applied. However, automatic selection is not reliable when the number of features is not sufficiently small compared to the available observations, as spurious correlations between features are more likely. In those cases, a selection procedure is applied that is motivated by practical considerations concerning the inclusion of the selected features on the final system. It would be desirable to relieve the system of computing whole groups of attributes (double deltas for instance) if their contribution is not so relevant. However this is not generally ensured when appling automatic selection or dimensionality reduction techniques. For this reason, the selection firstly determines the individual classification performance of a SVM classifier for each group of statistical attributes (e.g MFCC medians). Then a procedure similar to backward elimination is applied, that is, starting with the full set and deleting a group of attributes one at a time, trying to leave out as much as possible without reducing classification performance significantly. Individual performance of each group is used as a clue for selecting those to eliminate. As a result of this feature selection, each attribute category is finally represented by a reduced set of descriptors.

3.5. Classification strategy

In order to select a statistical classifier for the system, different methods provided by [Witten and Frank, 2005] were applied to the training set, and 10 times 10-fold CV classification performance was compared. The best performing classification model is finally incorporated into our singing detection system. Frame length used by the system is selected by comparing the performance of the selected model trained with the different training sets (of 0.5, 1 and 3 seconds excerpt length) and applied to the validation dataset.

Some post-processing strategies were considered to improve classification performance of the system based on classification confidence and contextual information. The first one is motivated by the fact that in some cases a fragment lies over a transition between vocal and non-vocal. For this reason, if a frame has a low classification confidence (based on probability estimates for each class) it is subdivided into two new fragments and each of them is re-classified. In case of a transition each new fragment could be classified to a different class (see figure 3). Additionally, two simple context rules are proposed. If a low probability fragment is surrounded by elements of the same class re-classification is avoided. On the other hand, if one of the half size fragments produced by re-classification is surrounded by elements of the other class, it is deleted.

4. Results

4.1. Selection of descriptors

Descriptors selection results are summarized in table 1. The selected features and the total number of coefficients for each category are presented. MFCC and PLPC set includes delta coefficients, while LFPC set includes delta and double delta coefficients. The only discarded type of descriptor in the Spectral set was Flux. The complete Spectral set includes: Centroid mean and median, Roll-off mean, median and standard deviation, Skewness mean and median, Kurtosis mean and median and Flatness mean, median and standard deviation.

Category	Performance	# Coefficients	Features selected
MFCC+D	84.5%	52	median, stdev
LFPC+D+DD	78.7%	72	median, stdev
Spectral	76.0%	21	whole set without Flux
PLPC+D	70.8%	78	median, stdev, skewness
HC	63.6%	2	mean, median
Pitch	59.1%	3	mean, stdev, kurtosis

Table 1: Feature selection results and 10 times 10-fold CV classification performance of a SVM on the training dataset of 1 second length audio excerpts. Performance is measured as percentage of correct decisions.

Descriptor sets are ranked in table 1 according to the performance obtained with a SVM using 10 times 10-fold CV over the training set (note that random guess rate is 50% in this task). The results

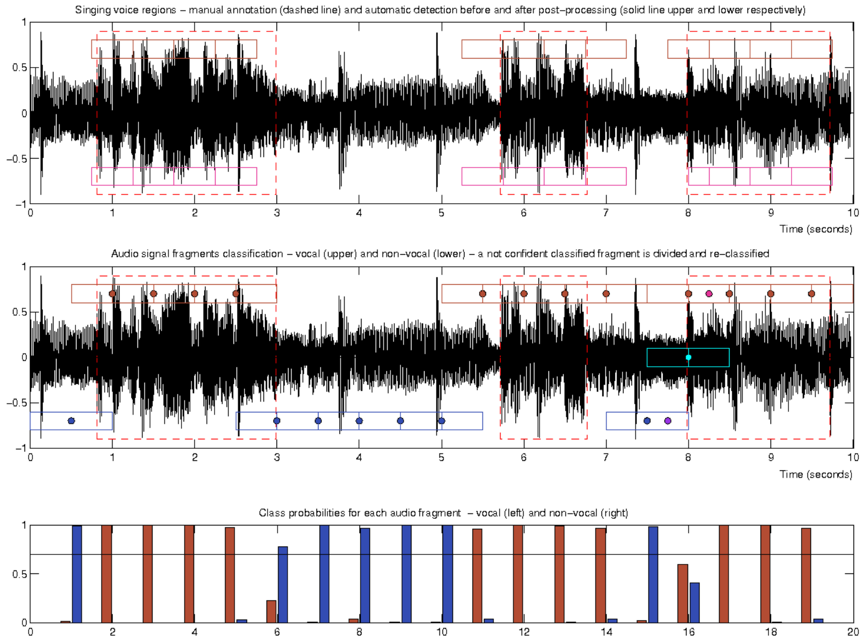

Figure 3: Post-processing based on classification confidence. An audio frame with low probability estimates for each class is subdivided into two new fragments and each of them is re-classified. In this example, fragment centered on second 8 lies over a transition between non-vocal and vocal and its probability estimates fall below a threshold of 0.7. The fragment is divided and each new portion is correctly classified. Classification is improved after post-processing (from 84.2% to 86.5%) as it is shown in the manual anotation versus automatic detection comparison at the top.

of paired t-tests (corrected resampled t-test [Witten and Frank, 2005]) comparing the performance of the various feature categories show that there are statistically significant differences between the MFCC set and each of the other sets (at a significance level of $p < 0.05$). The confussion matrix of the classification model based on MFCC features over the training set is presented in table 2.

		classified as	
		vocal	non-vocal
class	vocal	428	72
	non-vocal	77	423

Table 2: Confusion matrix of the classification on the training dataset of 1 second length audio excerpts using MFCC descriptors, obtained by 10-fold CV. The rows of the matrix correspond to the ground-truth of the audio excerpt and the columns indicate the hypothesis.

4.2. Classification strategy

Different classifiers were considered using the MFCC feature set on the training database comparing 10-fold CV classification performance. Table 3 shows the results obtained with a SVM trained using the Sequential Minimal Optimization method, a backpropagation ANN, a decision tree classifier implementing the well-known C4.5 algorithm and two different K-Nearest Neighbors (KNN).

Based on the classification performance on the validation set, 1 second turned out to be the most suitable fragment length. Results obtained were 73.3%, 74.2% and 73.0%, for 0.5, 1 and 3 seconds respectively. According to these results, our singing detection system was built with a SVM model using the MFCC feature set and a fragment length of 1 second.

The post-processing strategies proposed were added to the system one at a time in order to evaluate them on the validation database. Results obtained are reported in table 4.

Classifier	SVM	ANN	KNN-3	KNN-1	C4.5
Performance	85.1%	82.6%	76.5%	73.5%	73.1%

Table 3: Performances of different classifiers after 10-fold CV on the training database of 1 second length audio excerpts using the MFCC set.

Post-processing	none	re-classify	add rule 1	add rule 2
Performance	75.7%	76.3%	76.6%	76.8%

Table 4: Evaluation of the post-processing strategies on the validation dataset using the MFCC descriptors. Performance is computed as the percentage of time that the classification and the manual annotation coincides.

Post-processing	none	all
Performance	77.6%	78.5%

Table 5: Performance of the system on the testing dataset with no post-processing and performing all the post-processing strategies proposed, computed as the percentage of time that the classification and the manual annotation coincides.

Finally the system was used to process the testing set achieving a performance similar to that obtained in the validation set. Table 5 shows the results with no post-processing and considering all the post-processing strategies.

5. Discussion and conclusions

Analysis of the results obtained indicate that those descriptors that model the spectral content of the audio signal were the most appropriate for the problem (MFCC, LFPC, Spectral, PLPC). It is interesting that such a simple descriptor as LFPC outperformed all other features sets but MFCC, and that the general purpose Spectral outperformed PLPC. The results confirm that HC is not able to discriminate singing voice sounds from other harmonic musical instruments. The poor performance obtained with the Pitch descriptors is due to the utilization of a monophonic fundamental frequency estimation algorithm. We plan to apply other pitch estimation methods in our future work that could deal with polyphonic audio and to develop pitch descriptors that exploit singing voice pitch contour distinctive features such as intonation. Regarding MFCC, the feature selection performed points out that considering delta coefficients can boost performance (2% on the training set). However, they are generally not used when applying MFCC to this type of problem [Li and Wang, 2007] [Tsai and Wang, 2006]. Classification performance decreased significantly in validation and testing compared to 10-fold CV on the training set, which is not surprising because of the different origins and data variances of the databases. Additionally, the developed system roughly divides the audio file in fragments, so given our validation approach, we can take these results as worst-case or lower-bound estimations of the system performance.

We have studied the singing voice detection problem in music audio files by a statistical classification approach and we have compared, under equivalent conditions, the performance of several types of acoustic descriptors reported to be used for the problem. The results obtained confirm the usefulness of MFCC for this problem. As an outcome of our study, an effective singing voice detection system to process popular music audio files with a reduced set of descriptors has been developed. It is difficult to compare the performance achieved with other research work because there is no standard dataset for evaluation, but results obtained are promising and similar to the ones reported. Although our primary intention was to compare already used descriptors for this task, we have attempted to combine different descriptors as well as different classifiers. The overall classification performance was not improved so the results are not included. Also some other descriptors were tested without success. Our future work will follow this direction as it is reasonable to expect better results by combining different sources of information.

6. Acknowledgments

This work was carried out during an internship at the Music Technology Group (MTG) financially supported by Comisión Sectorial de Investigación Científica, UdelaR. First author is very grateful to all the people at the MTG for their support and to Dr. Alvaro Pardo for his careful reading of this article and his suggestions.

References

A. de Cheveignè, H. Kawahara (2002). YIN, a fundamental frecuency estimator for speech and music. *Journal Acoustic Society of America*, 111:1917–1930.

Berenzweig, A. and Ellis, D. (2001). Locating singing voice segments within music signals. *Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio, Mohonk NY, October 2001*, page 4pp.

Berenzweig, A., Ellis, D., and Lawrence, S. (2002). Using voice segments to improve artist classification of music. *AES 22nd International Conference*.

Chilton, T. (1999). Speech analisys. School of Electronic and Physical Sciences, University of Surrey.

Chou, W. and Gu, L. (2001). Robust singing detection in speech/music discriminator design. *International Conference on Acoustics, Speech, and Signal Processing*.

Cook, P. R. (1990). *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing*. PhD thesis, Stanford Univ., Stanford, CA.

Ellis, D. P. W. (2005). PLP and RASTA (and MFCC, and inversion) in Matlab. Online web resource: `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`.

Gerhard, D. (2002). Pitch-based acoustic feature analysis for the discrimination of speech and monofonic singing. *Journal of the Canadian Acoustical Assosiation*, pages 152–153.

Herrera, P., Klapuri, A., and Davy, M. (2006). Automatic Classification of Pitched Musical Instrument Sounds. In Klapuri, A. and Davy, M., editors, *Signal Processing Methods for Music Transcription*, pages 163–200. Springer, New York.

Kim, Y. E. and Whitman, B. P. (2002). Singer identification in popular music recordings using voice coding features. *In Proceedings of Interational Conference on Music Information Retrieval*, pages 164–169. Paris, France.

Li, Y. and Wang, D. (2007). Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech and Language Processing*.

Li, Y. and Wang, D. L. (2005). Separation of singing voice from music accompaniment for monaural recordings. Technical Report OSU-CISRC-9/05-TR61, Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA.

Maddage, N. C., Wan, K., Xu, C., and Wang, Y. (June 2004). Singing voice detection using twice-iterated composite fourier transform. *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference*, pages Page(s):1347 – 1350 Vol.2.

Maddage, N. C., Xu, C., and Wang, Y. (2003). A svm-based classification approach to musical audio. *Proc. ISMIR*.

Martin, K. D. (1999). *Sound-Source Recognition: A Theory and Computational Model*. PhD thesis, MIT. Cambridge, MA.

New, T. L., Shenoy, A., and Wang, Y. (2004). Singing voice detection in popular music. Technical report, Department of Computer Science, University of Singapore, Singapore, October 2004.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice Hall, New Jersey.

Scheirer, E. D. and Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. *ICASSP, Munich, Germany*.

Shenoy, A., Wu, Y., and Wang, Y. (2005). Singing voice detection for karaoke application. *Visual Communications and Image Processing 2005, Proc. of SPIE*, page Vol. 5960.

Sundberg, J. (1987). *The science of the singing voice*. De Kalb, Il., Northern Illinois University Press.

Tsai, W. H. and Wang, H. M. (2006). Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signal. *IEEE Transactions on Speech and Audio Processing, January 2006.*, pages Vol. 14, No 1.

Tzanetakis, G. (2004). Song-specific bootstrapping of singing voice structure. Technical report, Department of computer science, University of Victoria.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco.

Zhang, T. (2002). System and method for automatic singer identification. HP Labs Technical Report.