New speech/music discrimination approach based on fundamental frequency estimation

Article *in* Multimedia Tools and Applications · October 2009

DOI: 10.1007/s11042-008-0228-x · Source: DBLP

Nicolas Ruiz Reyes
Universidad de Jaén
131 PUBLICATIONS **1,180** CITATIONS
SEE PROFILE

Pedro Vera-Candeas
Universidad de Jaén
101 PUBLICATIONS **834** CITATIONS
SEE PROFILE

Sebastian Garcia Galan
Universidad de Jaén
83 PUBLICATIONS **656** CITATIONS
SEE PROFILE

Francisco Jesus Canadas Quesada
Universidad de Jaén
53 PUBLICATIONS **496** CITATIONS
SEE PROFILE

Multimed Tools Appl (2009) 41:253–286
DOI 10.1007/s11042-008-0228-x

New speech/music discrimination approach based on fundamental frequency estimation

N. Ruiz-Reyes · P. Vera-Candeas · J. E. Muñoz · S. García-Galán · F. J. Cañadas

Published online: 9 October 2008

Abstract Automatic discrimination of speech and music is an important tool in many multimedia applications. The paper presents a robust and effective approach for speech/music discrimination, which relies on a set of features derived from fundamental frequency (F0) estimation. Comparison between the proposed set of features and some commonly used timbral features is performed, aiming to assess the good discriminatory power of the proposed F0-based feature set. The classification scheme is composed of a classical Statistical Pattern Recognition classifier followed by a Fuzzy Rules Based System. Comparison with other well-proven classification schemes is also performed. Experimental results reveal that our speech/music discriminator is robust enough, making it suitable for a wide variety of multimedia applications.

Keywords Speech/music discrimination · Signal features · Pitch · Fundamental frequency estimation · Multimedia · Pattern recognition · Fuzzy systems

This work was supported in part by the Spanish Ministry of Education and Science under Project TEC2006-13883-C04-03.

N. Ruiz-Reyes (✉) · P. Vera-Candeas · J. E. Muñoz · S. García-Galán · F. J. Cañadas
Department of Telecommunication Engineering, University of Jaén Polytechnic School,
Linares, Jaén, Spain
e-mail: nicolas@ujaen.es

P. Vera-Candeas
e-mail: pvera@ujaen.es

J. E. Muñoz
e-mail: jemunoz@ujaen.es

S. García-Galán
e-mail: sgalan@ujaen.es

F. J. Cañadas
e-mail: fcanadas@ujaen.es

1 Introduction

Automatic discrimination between speech and music has become a research topic of interest in the last few years. Several approaches have been described in the recent literature for different applications [8, 16, 20, 36, 38, 39, 44, 45]. Each of them uses different features and pattern classification techniques and describes results on different material.

Saunders [38] proposed a real-time speech/music discriminator, which was used to automatically monitor the audio content of FM audio channels. Four statistical features on the zero-crossing rate and one energy-related feature were extracted, a multivariate-Gaussian classifier was applied, which resulted in high classification accuracy. Richard et al. [36] have recently developed a combined supervised and unsupervised approach, which includes feature selection, for automatic segmentation of radiophonic audio streams.

In Automatic Speech Recognition (ASR) of broadcast news, it's desirable to disable the input to the speech recognizer during the non-speech portion of the audio stream. Scheirer and Slaney [39] developed a Speech/Music Discrimination (SMD) system for ASR of audio sound tracks. Thirteen features to characterize distinct properties of speech and music, and three classification schemes (MAP Gaussian, GMM and k-NN classifiers) were exploited, resulting in an accuracy of over 90%. Matsunaga et al. proposed in [26] an audio source segmentation method for automatic indexing of broadcast news, which relies on spectral correlation features.

Another application that can benefit from distinguishing between speech and music is low bit-rate audio coding. Designing an universal coder to reproduce well both speech and music is the best approach. However, it is not a trivial problem. An alternative approach is to design a multi-mode coder that can accommodate different signals. The appropriate module is selected using the output of a speech-music classifier [10, 34, 41].

Automatic discrimination of speech and music is an important tool in many multimedia applications. Khaled El-Maleh et al. [8] combined the line spectral frequencies and zero-crossings-based features for frame-level narrow band SMD. The classification system operates using only a frame delay of 20 ms, making it suitable for real-time multimedia applications. An emerging multimedia application is content-based indexing and retrieval of audiovisual data. Audio content analysis is an important task for such application [24, 47]. Minami et al. [27] proposed an audio-based approach to video indexing, where a speech/music detector is used to help users to browse a video database. In [15], a SMD approach for content-based indexing and retrieval of cognitive multimedia is proposed. It is based on the application of the gray correlation analysis method, which makes the algorithm feasible for real-time multimedia applications.

Comparative view of the value of different types of features in speech music discrimination is provided in [3], where four types of features (amplitudes, cepstra, pitch and zero-crossings) are compared for discriminating speech and music signals. Experimental results showed cepstra and delta cepstra bring the best performance. Mel Frequencies Spectral or Cepstral Coefficients (MFSC or MFCC) are very often used features for audio classification tasks, providing quite good results. In [16], MFSC's first order statistics are combined with neural networks to form a speech music classifier that is able to generalize from a little amount of learning data. MFCC are one of the most used features in speech recognition and have recently proposed in musical genre classification of audio signals [2, 11, 42]. Comparison of statistical and information theory measures for automatic musical genre classification is reported in [11].

Unlike the previous works, SMD approaches based on only one type of features are presented in [21, 44], which result in fast and robust classification systems. The approach in [21] takes psychoacoustic knowledge into account by using the low frequency modulation amplitudes over 20 critical bands to form a good discriminator for the task. The approach in [44] exploits a new energy-related feature, called modified low energy ratio, that improves the results obtained with the classical low energy ratio.

In the problem of classification of audio signals, the requirements of low-complexity, high-accuracy and short delay are crucial for some practical scenarios. In [45], the authors propose a method of real-time speech/music classification with a hierarchical oblique decision tree, which comes with promising short delay of 10 ms, high accuracy of 98% and low complexity. Only two signal characteristics (the amplitude, measured by the Root Mean Square, and the mean frequency, measured by the average density of zero crossings) are used in [31], giving also rise to a short delay (20 ms), high accuracy (95%) and low complexity SMD approach.

Most previous works have focused on timbral features, with the idea of discriminating between music and speech from an accurate modeling of the speech spectrum. However, the vocal instrument also appears on music tracks and, sometimes, it is the predominant instrument. In that cases, speech/music discriminators based on timbral features are not well-fitted to the problem to be solved. Our approach is based on the following principle: intonation is the main difference between speech and music. Western music has common patterns in frequency and rhythm. With respect to frequency, tonal instruments are played in order to generate notes in frequency. In our approach, intonation is pursued from the estimation of the fundamental frequency to implement a robust SMD system.

In this paper, we present our contribution to the design of a robust SMD system. The main contribution of the paper is the definition and assessment of a set of seven features derived from F0 estimation. The proposed features outperform classical timbral features, mainly when discrimination between speech and vocal instrument-based music (i.e. rap music) is addressed. The signal intonation pursued from F0 estimation is responsible of the improvement in the classification accuracy rate. Besides, the F0-derived features are easy to compute, once the fundamental frequency has been estimated. On the other hand, the proposed SMD approach achieves a further improvement in the classification accuracy rate by incorporating a Fuzzy Rules Based System (FRBS) to the classification stage, which consists of a classical statistical pattern recognition (SPR) classifier followed by the FRBS (two-stage classification scheme). It results in a robust and effective approach for SMD. The behavior of both the proposed F0-derived features and the two-stage classification scheme are assessed by comparison.

This paper is structured as follows. Comprehensive review of the main existing approaches for SMD and the areas of application are discussed in Section 1. Section 2 reviews the principles and main approaches of fundamental frequency estimation. It also briefly describes the signal processing operations involved in the YIN algorithm (algorithm used in this work to estimate the fundamental frequency). Section 3 is devoted to the proposed SMD approach. It consists of three parts: 1) Brief description of classical features for SMD, 2) New features based on F0 estimation, 3) The proposed two-stage decision-taking scheme. Experimental results are shown in Section 4, which allow to assess the performance of the proposed SMD approach. Finally, Section 5 outlines some meaningful conclusions regarding the performance of the proposed SMD approach.

2 Fundamental frequency estimation

The fundamental frequency (F0) of a periodic signal is the inverse of its period. This definition applies strictly only to a *perfectly* periodic signal. However, interesting signals, such as speech or music, depart from periodicity in several ways, and the art of fundamental frequency estimation is to deal with them in a useful and consistent way. The subjective pitch of a sound usually depends on its fundamental frequency, but there are exceptions. However, over a wide range of sounds pitch and period are in a one-to-one relation, so that the word "pitch" is often used in the place of F0, and F0 estimation methods are often referred to as "pitch detection algorithms". Modern pitch perception models assume that pitch is derived either from the periodicity of neural patterns in the time domain or else from the harmonic pattern of partials resolved by the cochlea in the frequency domain [4]. Both processes yield the fundamental frequency or its inverse, the period. Some applications give for F0 a different definition, closer to their purposes. For voiced speech, F0 is usually defined as the rate of vibration of the vocal folds. However, there are several factors which conspire to make the task of obtaining a useful estimate of speech F0 rather difficult.

The basic practical problem of monophonic fundamental frequency estimation is the following: estimate the time varying fundamental frequency from a given signal. To address this problem, the following constraints are often imposed:

- *Validity*. It is assumed that the problem is solvable, that is, the signal is quasi harmonic such that a pitch will be perceived.
- *F0 range*. For each practical problem, there is a more or less clearly defined range of possible F0 values.
- *F0 variability*. Besides the range of possible F0 values, the expected rate of change will also affect the performance of the algorithm. It is usually assumed that the rate of change is sufficiently slow.

F0 estimation is a topic that continues to attract much research effort, despite the many methods that have been proposed. Most of the methods for F0 estimation are based on one of the following approaches:

- Direct evaluation of periodicity. The simplest approach to periodicity evaluation is based on the investigation of the time domain waveform. Each F0 hypothesis can be assessed by testing how well the signal will resemble a delayed version of itself. Evaluation criteria can be either the correlation or the difference between the signal and its delayed version, both criteria being closely related.
- Evaluation of periodicity in the frequency domain. Periodicity hypothesis can also be evaluated in the spectral domain either by harmonic matching or spectral period evaluation.

- Time-frequency approach, which can incorporate either psychoacoustic information or an auditory processing model.

The most comprehensive review for F0 estimation is that of Hess [17, 18]. The most simple and commonly used method to estimate the fundamental frequency of a signal is the autocorrelation method. In response to a periodic signal, the AutoCorrelation Function shows peaks at multiples of the period. The "autocorrelation method" chooses the highest non-zero-lag peak by exhaustive search within a range of lags. It compares the signal to its shifted self. In that sense, it is related to the Average Magnitude Difference Function method, that performs its comparison using differences rather than products, and more generally to time-domain methods that measure intervals between events in time. Despite its appeal and many efforts to improve its performance, the autocorrelation method (and other methods for that matter) makes too many errors for many applications.

In this paper, F0 estimation is performed by the so-called YIN method of Alain de Cheveigne [4], which produces fewer errors than other well-known methods, such as the autocorrelation method. The name YIN (from "yin" and "yang" of oriental philosophy) alludes to the interplay between autocorrelation and cancelation that it involves. Some details about the YIN method are next reported.

Given that y_n is a discrete time-domain signal with sampling rate f_s (Hz), κ is a constant absolute threshold value, and W is the summing interval, the YIN algorithm produces, at time t, a fundamental frequency (or pitch) estimate, $F0_t$, and an aperiodicity (or voicing) value, Ap_t, as follows:

1. *Difference function*. Calculate the squared difference function $d_t(\tau)$, where τ is the lag.

$$d_t(\tau) = \sum_{j=t}^{t+W} \left(y_j - y_{j+\tau}\right)^2 \tag{1}$$

   When the signal input is periodic, the difference function shows zeros at multiples of the period. In this work, W=23.22 ms (1024 samples at a samples rate of f_s= 44100 Hz) is considered.
2. *Cumulative mean normalized difference function*. Evaluate the cumulative mean normalized difference function $d'_t(\tau)$ from the squared difference function $d_t(\tau)$.

$$d'_t(\tau) = \begin{cases} 1, & \tau = 0 \\ \frac{d_t(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} d_t(j)}, & Otherwise \end{cases} \tag{2}$$

3. *Minimum lag value using a constant absolute threshold*. Find the smallest value of τ for which a local minimum of $d'_t(\tau)$ is smaller than a given absolute threshold value κ. If such value is not found, find the global minimum of $d'_t(\tau)$ instead. Denote this lag value as τ'.
4. *Parabolic interpolation*. Interpolate the $d'_t(\tau)$ function values at abscissas $\{\tau' - 1, \tau', \tau' + 1\}$ with a second order polynomial.
5. *Best local estimate*. Search the minimum of the polynomial in the range $(\tau' - 1, \tau' + 1)$ and denote the corresponding lag value with $\hat{\tau}$.

It is worth noting how the steps build upon one another. Replacing the difference function (step 1) by the cumulative mean normalization operation (step 2) paves the way for the threshold scheme (step 3), upon which are based the parabolic interpolation (step 4) and the best local estimate (step 5). The error rates decrease as the successive steps of the YIN algorithm are completed.

At time *t*, the YIN method provides three values:

- The pitch estimate, $F0_t$. It is obtained as follows:

$$F0_t = \frac{f_s}{\hat{\tau}}(Hz) \tag{3}$$

  The values of the pitch estimate, $F0_t$, can be expressed in the logarithm scale, according to the following relation:

$$O_t = log_2(F0_t) - log_2(440) \tag{4}$$

  The logarithmic scale in Eq. 4 is selected in order to obtain O_t=0 when the pitch estimate is $F0_t$= 440 Hz (note A in 4th scale). By using the base-2 logarithm, parameter O_t increases in 1 when the frequency estimate is multiplied by 2. Since the YIN algorithm provides "good" pitch estimates from 27.5 to 7040 Hz, the frequency in the logarithmic scale of Eq. 4 ranges from −4 to 4.
  In a musical transcription system, the pitch estimate could be expressed in MIDI notes as follows:

$$MIDI_t = 69 + \lfloor 12 \cdot log_2\Big(\frac{f_s/\hat{\tau}}{440}\Big) + 0.5\rfloor \tag{5}$$

- Aperiodicity, denoted as Ap_t. The aperiodicity is the value of the cumulative-mean-normalized difference function $d'_t(\tau)$ at the estimated period $\hat{\tau}$. Therefore, the aperiodicity values range from 0 to 1, the normalized feature being obtained by Eq. 6:

$$Ap = d'_t(\hat{\tau}) \tag{6}$$

  A value of parameter Ap_t below the absolute threshold ($Ap_t < \kappa$) denotes a *voiced pitch estimate*, meaning that small aperiodicity values express a high degree of periodicity. When the aperiodicity is relatively high (above the absolute threshold κ), it results that the signal is not periodic. Here, it is supposed that F0 estimation is not valid when Ap_t is above the absolute threshold κ, and the analyzed signal is considered to be unvoiced. In such way, non-periodic signals can be distinguished. In this work, the absolute threshold value $\kappa = 0.2$ is considered.
- Power of the windowed discrete-time signal, denoted as P. When this parameter is below a threshold (2^{-15} for normalized signals), the signal is considered to be a silence.

Due to the fact that it requires not more than two times the period of the minimum fundamental to be detected, the YIN algorithm is especially suitable for highly non-stationary speech signals. For musical instruments, the problem of reverberation due to slow decay of previous notes renders the analysis more or less polyphonic. Due to the possibilities to restrict the analysis to limited frequency ranges and to threshold spectral amplitudes, it appears that for musical instruments spectral domain preprocessing is helpful such that spectral domain evaluation of the autocorrelation function is more appropriate. For the harmonic instruments, the error is below 1% if less than a quarter note of deviation is excepted. For speech signals, evaluation usually allows a larger tolerance of about 20% of the pitch. With this performance criterion, the YIN algorithm operates with less then 1% error on speech signals as well.

When dealing with complex music, the presence of polyphony makes F0 estimation far more difficult. In such situations, the F0 estimated by the YIN algorithm is not reliable. If there is a dominant F0 (i.e. small reverberation in an instrument), the estimated value appears to be a good approximation. However, when the signal energy is produced by more than one fundamental frequencies, the signal is not really periodic. As a result, the estimated aperiodicity is relatively high and the estimated F0 tends to be low. Therefore, the YIN algorithm usually labels polyphonic segments as unvoiced. This behavior is later shown in the examples for music signals.

Supposing that it can be reliably estimated, F0 is useful for a wide range of applications. Speech F0 variations contribute to prosody, and in tonal languages they help distinguish lexical categories. Attempts to use F0 in speech recognition systems [13, 19] have met with mitigated success, in part because of the limited reliability of estimation algorithms. Several musical applications (automatic score transcription, real-time interactive systems, etc.) need F0 estimation [12, 32], but here again the imperfect reliability of available methods is an obstacle. F0 is a useful ingredient for a variety of signal processing applications, for example, F0-dependent spectral envelope estimation [22]. Finally, a fairly recent application of F0 is as metadata for multimedia content indexing [25, 33].

In this paper, F0 estimation is required to compute a set of musically-inspired features, which constitute the basis of the analysis stage in our SMD approach. F0 estimation has also been used for audio discrimination in [14, 28]. In [14], the strategy is based on the concept of multiple fundamental frequencies estimation, which provides the elements for the extraction of three features from the signal. The discrimination between speech and music is obtained by properly combining such features. Molla et al. propose in [28] a new technique for voiced/unvoiced discrimination based on the extraction of pitch period. Empirical Mode Decomposition (EMD) is employed for multi-band representation of speech signal in the time domain. The fundamental oscillation in a speech segment is determined in the autocorrelation function of the EMD space. A damped cosine model is fitted using least squared method to extract the frequency of the fundamental oscillation.

3 The proposed SMD approach

Section 3 is organized in three parts. First, Section 3.1 reviews the most commonly used features in audio classification tasks, which represent timbral texture and are based on the Short Time Fourier Transform (STFT). Then, the new F0-derived features are defined in Section 3.2. It also shows the normalized histograms for all defined features. Finally, the proposed two-stage decision-taking scheme is described in Subsection 3.3, which also contains the motivation of using the proposed classification scheme.

3.1 Classical features for SMD

Most of the works concerning audio classification rely on three types of features: timbral texture features, rhythmic content features and pitch content features [2, 42]. However, timbral texture features are the most commonly used in audio classification when it is reduced to SMD. The features used to represent timbral texture are based on standard features proposed for music-speech discrimination [39] and speech recognition [6]. These standard features are based on the STFT and are calculated for every short-time frame of sound. The following specific features representing timbral texture are used in our approach for comparison purposes:

1. *Spectral Centroid (SC)*. The SC is defined as the center of gravity of the magnitude spectrum of the STFT:

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]} \tag{7}$$

   where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency bin n. The centroid is a measure of spectral shape and higher centroid values correspond to "brighter" textures with more high frequencies.
2. *Spectral Rolloff (SR)*. The SR is defined as the frequency R_t below which 85% of the magnitude distribution is concentrated:

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \tag{8}$$

   The rolloff is another measure of spectral shape.
3. *Spectral Flux (SF)*. The SF is defined as the squared difference between the normalized magnitudes of successive spectral distributions:

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \tag{9}$$

   where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current frame t, and the previous frame $t-1$, respectively. The SF is a measure of the amount of local spectral change.
4. *Time Domain Zero Crossings (ZC)*. This timbral texture feature is defined as:

$$Z_t = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \tag{10}$$

   where the *sign* function is 1 for positive arguments and 0 for negative arguments and $x[n]$ is the time domain signal for frame t. ZC provide a measure of the noisiness of the signal.
5. *MFCC*. MFCC are perceptually motivated features that are also based on the STFT. After taking the log-amplitude of the magnitude spectrum, the FFT bins are grouped and smoothed according to the perceptually motivated Mel-frequency scaling. Finally, in order to decorrelate the resulting feature vector

a Discrete Cosine Transform (DCT) is performed. Although typically 13 coefficients are used for speech representation, it has been found that the first five coefficients provide the best performance for musical genre classification of audio signals [42]. Therefore, we have also used the first five coefficients in our SMD approach, without adding the derivatives. The use of MFCCs to separate music and speech has been explored in [23].

3.2 New features based on fundamental frequency estimation

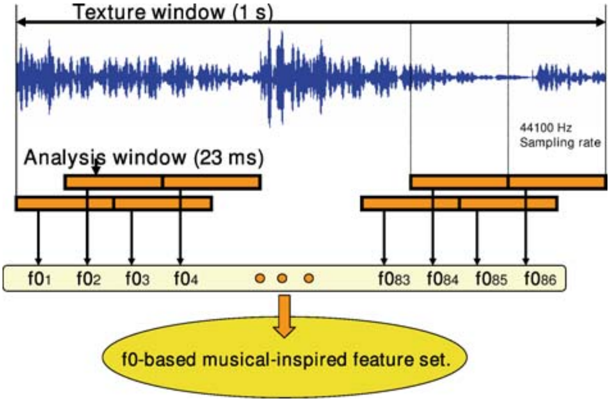
In this paper, a set of features derived from F0 estimation are defined for SMD. In particular, the set is composed of seven features, most of them having musical meaning. Other types of F0-derived features have been used for different applications [9, 35, 37]. A speech-to-singing synthesis system that can synthesize a singing voice is proposed in [37]. The F0 contour of the singing voice is generated from four types of F0-derived features: overshoot, vibrato, preparation, and fine fluctuation. In [35], a voice conversion method based on transformation of spectral and intonation features is proposed. The F0 contour is used for modeling the pitch and intonation patterns of speech.

In this work, an *analysis window* of 23.22 ms (1024 samples at 44100 Hz sampling rate) and a *texture window* of approximately 1 s (43 analysis windows) are defined. Overlapping with a hop size of 512 samples is performed. F0 estimation is performed at each analysis window, which results in a 86-length vector for each texture window. From the vector containing the F0 estimated at each texture window, a set of features is computed. The texture window is shifted by 250 ms, which entails updating the feature set every 250 ms. At a first approximation, statistical features, such as standard deviation and skewness, could be computed. However, since F0 is related to musical concepts, we have defined a set of features based on musical principles. As just stated, these features are computed every second from the F0 estimated at each texture window. Figure 1 shows the windows scheme to compute the F0-based feature set.

When dealing with speech signals, the estimated F0 fits to a characteristic pattern for most of the analyzed signals. Speech signals contain voiced frames (near-periodic) and unvoiced frames (aperiodic), which are alternated in short time intervals. In most of languages, words are composed of voiced and unvoiced phonemes, which results in several voiced-unvoiced boundaries within a word. Good estimates of F0 are accomplished for voiced frames, while it does not make sense to estimate F0 for unvoiced frames. Moreover, voiced speech frames have a time-varying F0, because the pitch changes when voiced phonemes are pronounced. An example of F0 estimation for a representative speech signal (Male speech, French) is showed in Fig. 2. As can be seen, the F0 estimated in voiced frames (aperiodicity below 0.2) slowly varies in terms of the speaker's intonation, while the F0 estimated in unvoiced frames (aperiodicity above 0.2) has a quite different behavior (sudden changes of F0 can happen).

Instead, music signals show a quite changing behavior. There is no generic pattern for such signals. It depends on several factors, such as the music genre, polyphony, instruments involved, etc. However, two specific patterns can be identified in music signals: 1) F0 does not change when only one note is played at any time (steady state within the same note); 2) Step-wise changes often happen when passing from a

Fig. 1 Example illustrating how to compute the F0-based feature set


Texture window (1 s)
44100 Hz Sampling rate
Analysis window (23 ms)
f01 f02 f03 f04 f083 f084 f085 f086
f0-based musical-inspired feature set.

musical note to another. Two examples of music signals are shown in Figs. 3 and 4. Figure 3 depicts the F0 estimated for a music signal composed of a single instrument (violin, melodious phrase) of 1 second. As can be seen, most of the time the signal

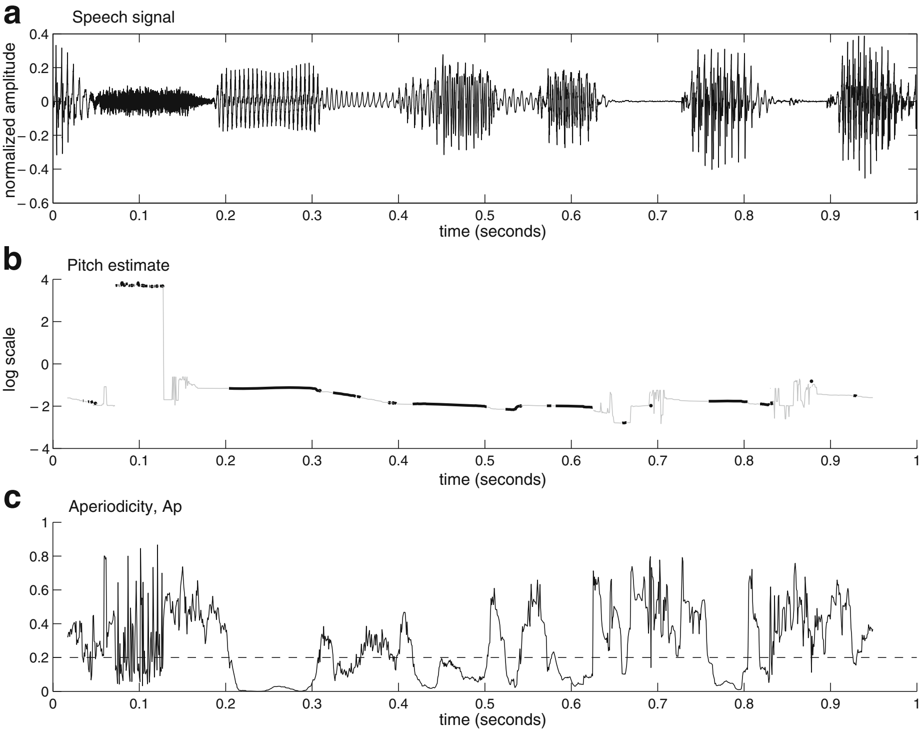


Fig. 2 F0 estimate for a representative speech signal of 1 s. **a** Normalized waveform; **b** Estimated F0. The *thick line* corresponds to the segments that are "classified" as voiced by applying the 0.2 threshold to the aperiodicity (Ap0); **c** Aperiodicity. The *dashed line* represents the boundary to "classify" the signal frames as voiced or unvoiced

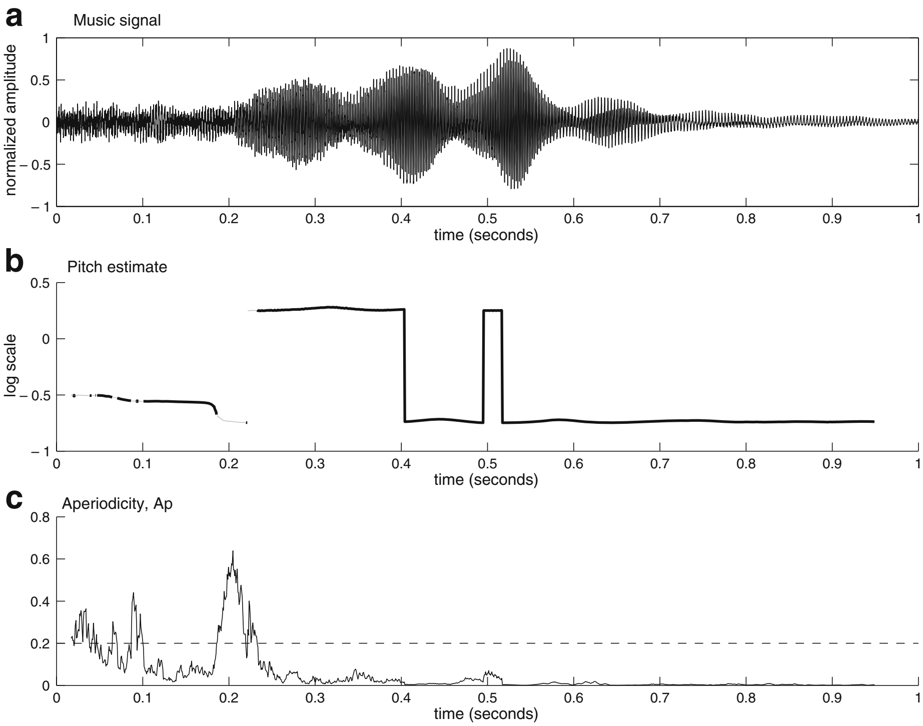


Fig. 3 F0 estimate for a single instrument music signal of 1 s. **a** Normalized waveform; **b** Estimated F0. The *thick line* corresponds to the segments that are "classified" as voiced by applying the 0.2 threshold to the aperiodicity (Ap0); **c** Aperiodicity. The *dashed line* represents the boundary to "classify" the signal frames as voiced or unvoiced

is monophonic, which involves that most of the signal frames are considered to be voiced. Further, when there is a dominant fundamental frequency, the estimated F0 is nearly-steady during the time interval of the played note. Figure 4 shows the F0 estimated for a vocal music signal generated by a quartet. As can be seen, most of the time the signal is polyphonic, which involves that most of the signal frames are considered to be unvoiced.

The signals analyzed in Figs. 2, 3 and 4 have been taken from the European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) Compact Disc, which contains a set of audio programme signals that are recommended by the EBU for subjective test purposes. The EBU SQAM CD is available online at http://www.ebu.ch/en/technical/publications/tech3000_series/tech3253/index.php. The information concerning the signals in Figs. 2, 3 and 4 is shown in Table 1.

The audio test database has been carefully prepared. It has been used for both obtaining the normalized histograms of the F0-derived features, and for assessing the proposed SMD approach. The database consists of a continuous 2-h audio signal representative of the two corresponding audio classes (speech and music). The speech data has been taken from news programs, dialogs and announcements of radio and TV stations, and the languages involve English, Spanish, French, German, Catalonian and Portuguese with different levels of noise, especially in news

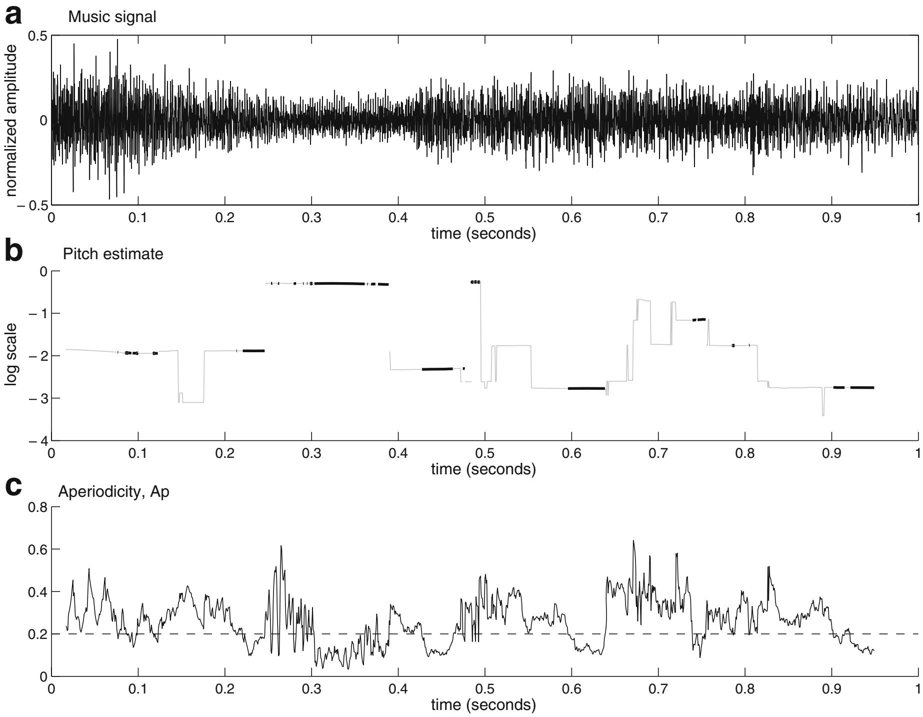


Fig. 4 F0 estimate for a vocal (quartet) music signal of 1 s. **a** Normalized waveform; **b** Estimated F0. The *thick line* corresponds to the segments that are "classified" as voiced by applying the 0.2 threshold to the aperiodicity (Ap0); **c** Aperiodicity. The *dashed line* represents the boundary to "classify" the signal frames as voiced or unvoiced

programs. The speakers involve male and female with different ages. The music data has been taken from musical programs of radio and TV stations too, and consist of songs and instrumental music. The songs cover several musical genres, such as rock, pop, folk, rap and funky, and they are sung by male and female in English and Spanish. The instrumental music covers different instruments (piano, violin, cello, pipe, clarinet) and styles (symphonic music, chamber music, jazz, electronic music). Some soundtracks which consists of several instruments are also included. We have attempted that the data sets are representative of the two classes to be classified (speech and music) so that the results are indicative of the discrimination performance with real-world unknown signals.

Table 1 Information concerning the signals in Figs. 2, 3 and 4

Figure	Track	Section	Duration	Interval	Signal	File
2	52	01	0:24	7:8	Male speech, French	52bwf.wav
3	08	02	0:42	7:8	Violin, melodious phrase	08mbwf.wav
4	48	01	0:28	7:8	Quartet (vocal)	48bwf.wav

Table 2 Detailed description about the audio database

Speech	Percentage (%)	Music	Percentage (%)
Male, English	7.5	Flamenco	2.5
Male, German	2.5	Pop	5
Male, Catalonian	2.5	Rock	5
Male, Spanish	15	Hip-hop	2.5
Male, Portuguese	2.5	Symphonic	5
Male, French	2.5	Chamber	5
Female, English	5	Folk	5
Female, German	2.5	Metal	5
Female, Catalonian	2.5	Jazz	5
Female, Spanish	5	Funky	2.5
Female, Portuguese	0	Rap	2.5
Female, French	2.5	Electronic	5
	50		50

For experimental fairness, a detailed description about the audio database used in the experiments is provided in Table 2.

Next, the set of features derived from F0 estimation, and computed each texture window, is defined:

1. *Dynamic range of aperiodicity* (D_{Ap}). It is defined as the difference between the maximum and minimum values of the normalized aperiodicity (Ap) within a texture window. Feature D_{Ap} is expressed as follows:

$$D_{Ap} = max(\mathbf{Ap}) - min(\mathbf{Ap}) \tag{11}$$

$\mathbf{Ap} = [Ap_1, Ap_2, ..., Ap_t, ..., Ap_T]$ being the vector containing the values of the normalized aperiodicity computed for a given texture window, and T the number of analysis windows in the computation interval (texture window).
This feature is intended to discriminate between speech and music when the music signal is either noisy (unvoiced) or voiced during the whole texture window. Speech signals typically alternate voiced frames (low aperiodicity) and unvoiced frames (high aperiodicity) during the texture window (1 second). Typically, speech signals show high dynamic range of aperiodicity in the computation interval, while music signals tend to provide lower values. Figure 5 shows the normalized histograms of feature D_{Ap} for both speech and music. As shown in Fig. 5, feature D_{Ap} entails a good discriminatory capability.

2. *Average of the estimated F0* ($F0_{av}$). It is defined as the mean value of the F0 estimated at the current texture window.
Before computing this feature, the pitch estimate (in Hz) is converted to logarithmic values, according to Eq. 4. In this way, the logarithmic behavior of the ear is taken into account. Therefore, computation of feature $F0_{av}$ is performed as follows:

$$F0_{av} = \frac{\sum_{t=1}^{T} O_t}{T} \tag{12}$$

where $T = 85$ is the number of analysis windows at each texture window.
The normalized histograms of feature $F0_{av}$ for both speech and music are shown in Fig. 6. As can be seen, speech signals have a typical pitch range, which goes

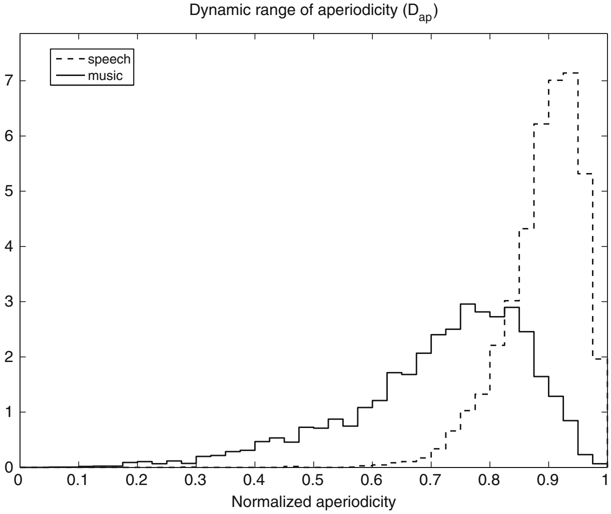


Fig. 5 Normalized histograms of feature D_{Ap} for both speech and music

from −2.5 to −1. However, the pitch range of music signals goes from −3.5 to −2. This result is justified from the fundamental frequency estimation that the YIN algorithm detects for polyphonic music. This situation (presence of polyphonic musical passages in music frames) is very common in our database.

3. *Dynamic range of estimated F0* (D_{F0}). It is defined as the difference between the maximum and minimum values of the estimated F0 within the current texture window. Feature D_{F0} is expressed as follows:

$$D_{F0} = max(\mathbf{O}) - min(\mathbf{O}) \tag{13}$$

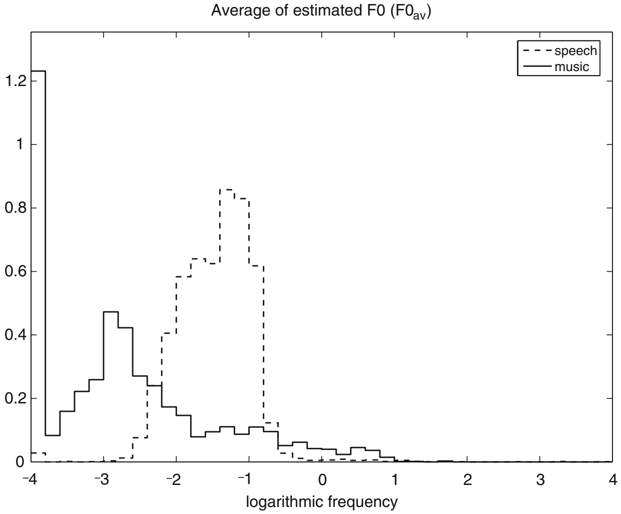


Fig. 6 Normalized histograms of feature $F0_{av}$ for both speech and music

$\mathbf{O} = [O_1, O_2, ..., O_t, ..., O_T]$ being the vector containing the values of the fundamental frequency (expressed in a logarithmic scale) computed for a given texture window.

In speech signals, speaker's intonation makes the estimated F0 varies in a typical range (≤ 1 in the logarithmic scale). Further, noisy speech frames are sometimes labeled as voiced, the estimated F0 being very high. In these cases, feature D_{F0} is very high in the current texture window. On the other hand, music signals present a completely different histogram of feature D_{F0}. These results can be assessed by the normalized histograms presented in Fig. 7.

4. *Maximum note duration* (ND_{max}). It is defined from the number of consecutive analysis windows comprising the longest musical note within the observation interval (the current texture window). Therefore, computation of the musical note corresponding to the each analysis window from the estimated F0 must be first addressed. The musical note at the t-th analysis window is here computed as follows:

$$Note_t = \lfloor 12 \cdot (O_t + 4) + 0.5 \rfloor + 1 \tag{14}$$

In this way, since parameter O_t range from −4 to 4, musical notes are ordered from 1 to 96. To understand Eq. 14, note that 12 consecutive semitones represent an octave.

Once all musical notes in the current texture window have been computed, feature ND_{max} is obtained from the longest time interval containing the same musical note.

The normalized histograms of feature ND_{max} for both speech and music are shown in Fig. 8. As can be seen, speech signals have typical values of the maximum note duration, unlike music signals.

5. *Number of notes* (N_{note}). This parameter is defined as the number of different notes contained within the observation interval (the current texture window). From the fundamental frequencies estimated in the observation interval, we

Fig. 7 Normalized histograms of feature D_{F0} for both speech and music

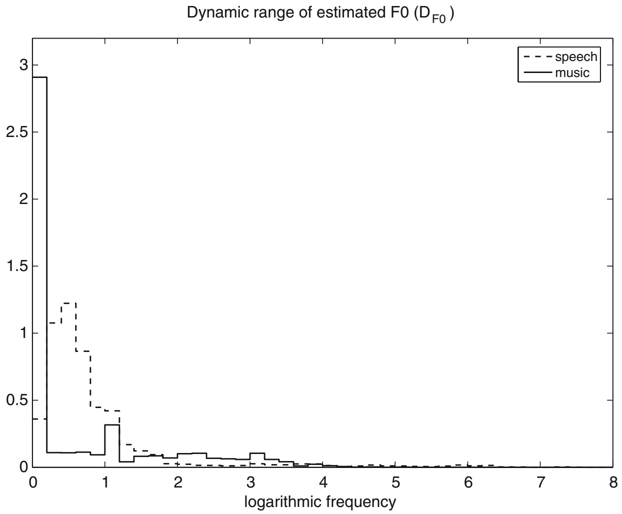
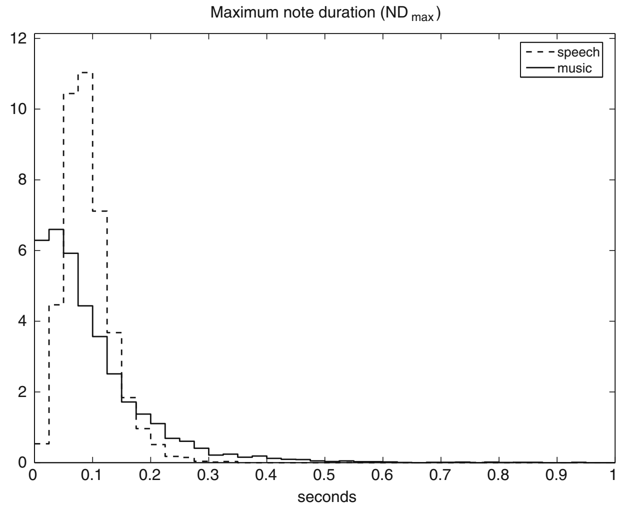


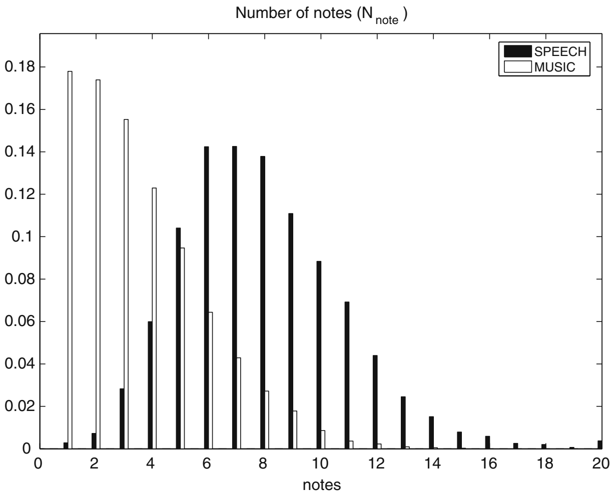
Fig. 8 Normalized histograms of feature ND_{max} for both speech and music



compute how many different notes are detected. For speech signals, it is common to obtain high values of parameter N_{note} (above 6 notes), because the estimated F0 changes with the speaker's intonation. On the other hand, lower values of parameter N_{note} are usually obtained for music signals (below 5 notes), because the estimated F0 remains steady in variable duration intervals. It is justified with the results of Fig. 9, where the normalized histograms of feature N_{note} for speech and music are depicted.

6. *Voiced ratio* (VR). It is defined as the ratio between the number of voiced frames and the total number of frames within the observation interval. This parameter

Fig. 9 Normalized histograms of feature N_{note} for both speech and music



informs us about the percentage of frames in which F0 is properly estimated at each observation interval. It can be expressed as follows:

$$VR = \frac{N_{voiced}}{T} \tag{15}$$

where N_{voiced} is the number of frames that fulfil the following condition: $Ap0_t \leq 0.2$. $Ap0_t$ is the aperiodicity at the t-th analysis window of the current texture window.

Generally, speech signals have a balanced ratio of voiced frames (parameter VR usually ranges from 0.3 to 0.6). However, the ratio of voiced frames tends to be small for music signals (parameter VR is very often below 0.2), because polyphonic frames are usually labeled as unvoiced. These results can be assessed from normalized histograms of Fig. 10.

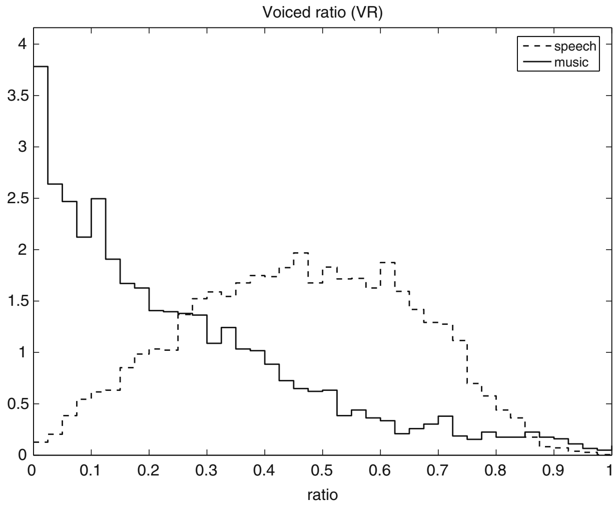
7. *Average value of the aperiodicity* (Ap_{av}). Mean value of the normalized aperiodicity at the current texture window. It is only defined for voiced frames, and can be expressed as follows:

$$Ap0_{av} = \frac{\sum_{t\in V} Ap_t}{N_{voiced}} \tag{16}$$

where set V is composed of those frames that fulfil the following condition: $Ap_t \leq 0.2$.

Figure 11 shows the normalized histograms of feature Ap_{av} for both speech and music. As shown in Fig. 11, parameter Ap_{av} usually ranges from 0.08 to 0.12 for speech signals. In voiced speech frames, vocal folds are vibrating most of the time, and the aperiodicity is typically around 0.1. Regarding music signals, most of the frames are polyphonic, which implies high values (above 0.2) of aperiodicity $Ap0$, as just stated. However, polyphonic frames sometimes have aperiodicity values below (but close to) 0.2, being labeled as voiced. This fact explains that the average value of the aperiodicity for voiced music frames tends to be high (ranging from 0.15 to 0.19). Finally, as shown in Fig. 11, feature Ap_{av}

Fig. 10 Normalized histograms of feature VR for both speech and music



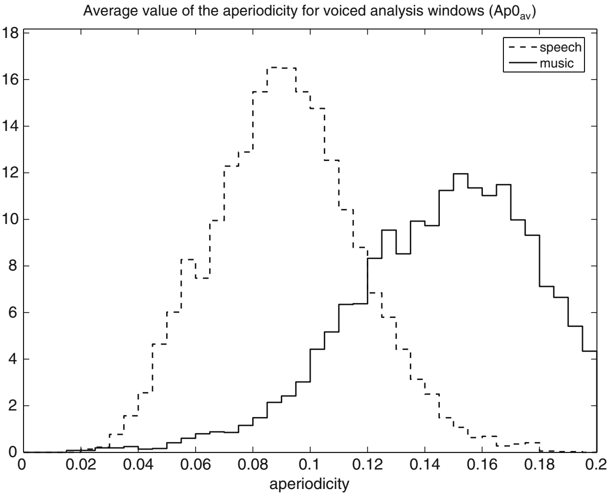


Fig. 11 Normalized histograms of feature Ap_{av} for both speech and music

has good discrimination capability due to its different behavior for speech and music.

As just stated, all histograms shown in Figures from 5 to 11 have been obtained using a continuous 2-h audio signal representative of the audio classes (speech and music) to be distinguished. The behavior of each feature is illustrated by two normalized histograms (one for each audio class to be distinguished), which evidence the discrimination capability of the proposed F0-derived features. The histograms in figures from 5 to 11 have been normalized so as to resemble probability density functions.

When all analysis windows in a texture window are labeled as either unvoiced or silence, the previously described features have no sense, and a boolean flag is activated to inform the classifier.

3.3 Two-stage classification scheme

In this work, the classification scheme consists of two constitutive elements that operate in series. First, the seven features provided by the analysis stage at each 1 second-length texture window are used as an input to a SPR classifier. As just stated, the feature set is updated each 250 ms, which involves updating the probability provided by the SPR classifier each 250 ms. Next, a FRBS processes the last four probabilities at the output of the SPR classifier, taking a decision (speech or music). With four probabilities, a trade-off solution between the classification accuracy rate and the training computational cost is found, as explained at the end of this Section. According to this scheme, the FRBS makes a decision each 250 ms, taking into account not only the probability provided by the SPR classifier at the current texture window, but also the probabilities at the three previous texture windows, the hop size being 250 ms. Therefore, the proposed decision-taking stage, composed of a SPR classifier followed by a FRBS, incorporates memory to the speech/music discriminator, which allows to increase the classification accuracy rate, as shown in Section 4.

For classification purposes, a number of SPR classifiers [7] were evaluated. The basic idea behind SPR is to estimate the probability density function (pdf) for the feature vectors of each class. In supervised learning, a labeled training set is used to estimate the pdf of each class. In the simple Gaussian (GS) classifier, each pdf is assumed to be a multidimensional Gaussian distribution whose parameters are estimated using the training set. In the Gaussian Mixture Model (GMM) classifier, each class pdf is assumed to consist of a mixture of a specific number K of multidimensional Gaussian distributions. Unlike the k-NN classifier, which needs to store all the training feature vectors in order to compute the distances to the input feature vector, the GMM classifier only needs to store the set of estimated parameters for each class. The iterative Expectation-Maximization algorithm can be used to estimate the parameters of each Gaussian component and the mixture weights. In this work, we have used two SPR classifiers (GMM and k-NN) and modern classification techniques, such as Neural Networks (NN) and Support Vector Machines (SVM). We have employed a Multi Layer Perceptron NN, which consists of three layers: the input layer with seven neurons (one for each F0-derived feature), the hidden layer with seven neurons and the output layer with only one neuron. The SVM here used is based on Radial Basis Functions, which has been properly trained and adjusted.

A FRBS is a system based on fuzzy logic, which aims to solve imprecise, uncertain or qualitative decision-making problems in similar way to humans. FRBS can be characterized in terms of their fundamental constituents: Fuzzification, Knowledge Base, Inference Engine and Defuzzification.

- *Fuzzification*. Process of mapping from the system inputs space to fuzzy sets in a defined universe, which gives rise to the membership functions associated to each one of the system inputs.
- *Knowledge Base*. It is composed of the Base of Rules and the Data Base. The Data Base contains the definition of the fuzzy linguistic terms. The Base of Rules is constituted by the collection of fuzzy rules representing the expert knowledge. A fuzzy rule is a conditional statement with the following form:

$$\text{R: IF } (X_1 \text{ is } A_1) \text{ AND } (X_2 \text{ is } A_2) \text{ AND ... AND } (X_n \text{ is } A_n) \text{ THEN } Y \text{ is } B$$

  where X_i and Y are the linguistic variables, and A_i and B are fuzzy sets.
- *Inference Engine*. It uses the Knowledge Base and the Fuzzy Inputs to make inference by means of a reasoning method. The inference engine is based on the application of the generalized modus ponens, extension of the classical logic modus ponens, in the following way:

$$\text{R: IF } (X \text{ is } A) \text{ THEN } Y \text{ is } B; \quad X \text{ is } A^* \quad \Rightarrow \quad Y \text{ is } B^*$$

  where the consequent B^* is deduced by projection on Y of the compositional rule of inference $B^* = A^* \circ R$, $\circ$ being the composition operator. The inference procedure is determined by two factors: the implication operator applied to each rule and the composition operator used to combine outputs from all rules. In this work, the minimum operator is considered for the fuzzy implication and the max-min operator for the composition.

- *Defuzzification*. It is used to reconvert the fuzzy output values, derived from the inference mechanism, into crisp values. The output of the Inference Engine is a fuzzy set, and for practical applications a crisp value is needed. The most used strategy for Defuzzification is the Center Of Area (COA) method, which gives the center of gravity of the output membership function.

In the proposed SMD scheme, a k-NN SPR classifier has been used by default, because it has been widely used in the literature for pattern recognition or classification tasks. The considered classifier does not provide good enough classification accuracy rate, because it only operates on the 1s-length current texture frame, without taking into account the probabilities at the output of the classifier for previous frames. In order to reduce the number of misclassification errors, a FRBS is cascaded with the k-NN classifier. The FRBS makes a decision about the class (speech or music) of the current audio frame from four input parameters: probabilities p_0, p_1, p_2 and p_3 obtained by the classifier for four consecutive 1s-length texture frames, the last of them being the current frame and the hop size 250 ms. The FRBS has only one output parameter, called *Decision*, which ranges from 0 to 1. If *Decision* is higher than 0.5, the fuzzy system decides in favor of *speech*. Otherwise, it decides in favor of *music*. All probabilities are [0,1] normalized.

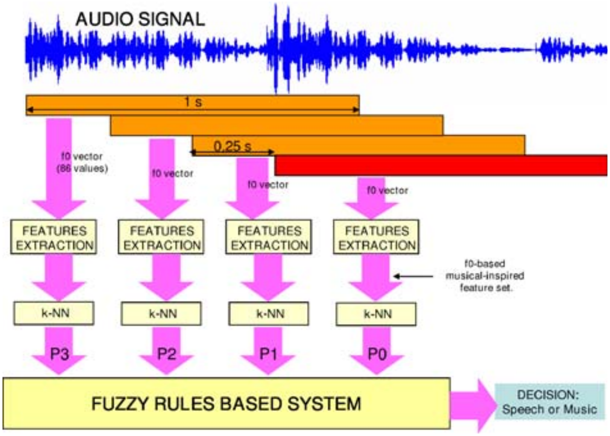
The structure of the proposed classification scheme is shown in Fig. 12.

The proposed approach for SMD is summarized in the following algorithmic representation in order to keep in mind the main features of the proposed approach.

Input and output membership functions are represented in Figs. 13 and 14, respectively. Some details about membership functions are now reported in order to explain the meaning of Figs. 13 and 14.

The membership function of a fuzzy set is a generalization of the indicator function in classical sets. In fuzzy logic, it represents the degree of truth as an extension of valuation. Membership functions were introduced by Zadeh in the first paper on fuzzy sets [46]. The membership function which represents a fuzzy set $\tilde{A}$ is usually denoted by $\mu_{\tilde{A}}$. For an element x of set X, the value $\mu_{\tilde{A}}(x)$ is called the membership degree of x in the fuzzy set $\tilde{A}$. The membership degree $\mu_{\tilde{A}}(x)$ quantifies the grade of membership of element x to the fuzzy set $\tilde{A}$. The value 0 means that x is not a

Fig. 12 General structure of the proposed classification scheme



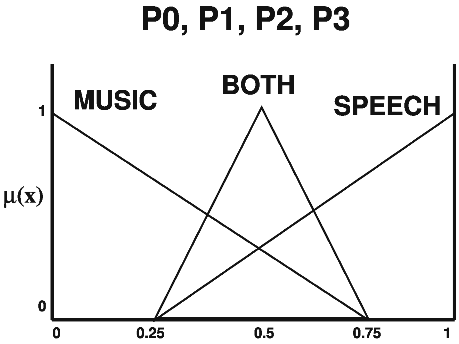


Fig. 13 Membership functions for input variables

member of the fuzzy set; the value 1 means that x is fully a member of the fuzzy set; the values between 0 and 1 characterize fuzzy members, which belong to the fuzzy set only partially.

As can be seen in Fig. 13, there are three fuzzy sets (MUSIC, SPEECH, BOTH) for each input variable (P_0, P_1, P_2, P_3). However, there are only two fuzzy sets (MUSIC, SPEECH) for the output variable (see Fig. 14).

When the obtained knowledge for the FRBS is not considered good enough to be used, some kind of learning is needed. In such sense, automatic definition of FRBSs can be considered in many cases as an optimization or search process. Genetic Algorithms are known to be capable of finding near optimal solutions in complex search spaces. In this work, the new rules added to the knowledge base of the FRBS have been obtained using Genetic Algorithms-based evolutionary computation (genetic learning algorithms), giving rise to a Genetic Fuzzy System (GFS). This means that the FRBS is evolved by a genetic learning process. A good review of GFS is found in [5]. The main genetic learning algorithms for FRBSs are known as Michigan [1], Pittsburgh [40] and Iterative Rule Learning [43].

The genetic learning algorithm used in this work to evolve the FRBS is the Pittsburgh algorithm. Next, the genetic learning process is described. In the Pittsburgh approach, each chromosome represents an entire base of rules and evolution is accomplished by means of genetic operators applied at the level of fuzzy rule sets. The fitness function evaluates the accuracy of the entire rule base encoded in the chromosome. The genetic learning process proposed in this work for SMD using the Pittsburgh approach is illustrated in Fig. 15.

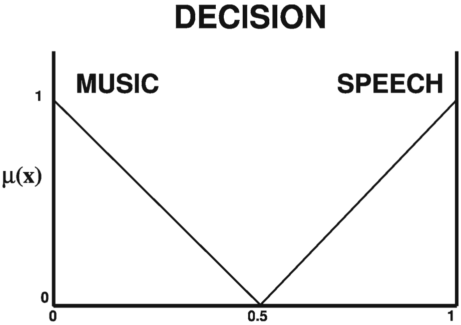


Fig. 14 Membership functions for the output variable

Algorithm 1 Proposed SMD approach

```
{DEFINITION of VARIABLES AND FUNCTIONS}
{F0 values at the current texture window}
$\mathbf{F0_0} = [F0_{01}, F0_{02}, ..., F0_{086}]$;
{F0 values at the three previous texture windows}
$\mathbf{F0_1} = [F0_{11}, F0_{12}, ..., F0_{186}]$;
$\mathbf{F0_2} = [F0_{21}, F0_{22}, ..., F0_{286}]$;
$\mathbf{F0_3} = [F0_{31}, F0_{32}, ..., F0_{386}]$;
{F0-derived features for the current texture windows}
$\mathbf{F_0} = [F_{01}, F_{02}, ..., F_{07}]$;
{F0-derived features for the three previous texture windows}
$\mathbf{F_1} = [F_{11}, F_{12}, ..., F_{17}]$;
$\mathbf{F_2} = [F_{21}, F_{22}, ..., F_{27}]$;
$\mathbf{F_3} = [F_{31}, F_{32}, ..., F_{37}]$;
{Membership functions for the input variables}
$\mu(x)_{P_0}; \mu(x)_{P_1}; \mu(x)_{P_2}; \mu(x)_{P_3}$;
{Membership function for the output variable}
$\mu(x)_{OUTPUT}$;
{Rule Base composed of n rules}
$\mathbf{RB} = \{R_1, R_2, ...R_n\}$;

{ANALYSIS STAGE}
for $i = 0$ to 3 do
  $\mathbf{F0_i} \Leftarrow ComputeF0(TextureWindow_i)$;
  $\mathbf{F_i} \Leftarrow ComputeFeatures(\mathbf{F0_i})$;
end for

{CLASSIFICATION SCHEME}
{Classification}
for $i = 0$ to 3 do
  P$i \Leftarrow Classifier(\mathbf{F_i})$;
end for

{Fuzzy Rules Based System}
$[FP_0, FP_1, FP_2, FP_3] \Leftarrow Fuzzifier(P_0, P_1, P_2, P_3)$;
for $i = 1$ to $n$ do
  $FuzzyOut = FuzzyOut \bigcup FuzzyInference(FP_0, FP_1, FP_2, FP_3, R_i)$;
end for
$Out = Defuzzifier(FuzzyOut)$;

{Decision-taking}
if $Out <= 0.5$ then
  $DECISION \Leftarrow MUSIC$;
else
  $DECISION \Leftarrow SPEECH$;
end if

return DECISION
```

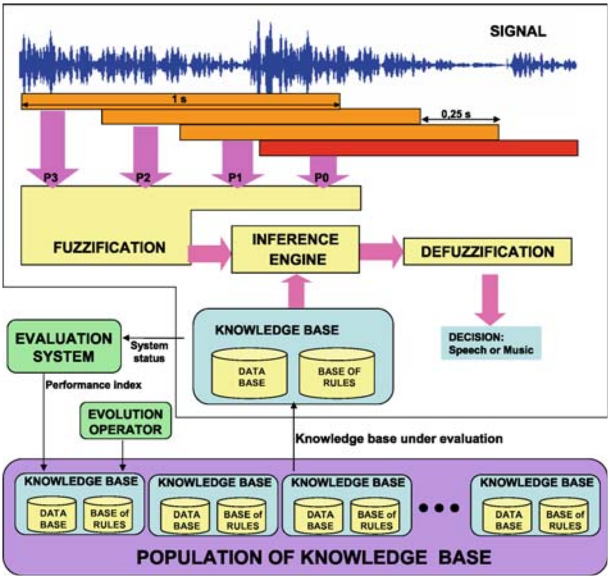


Fig. 15 General structure of the Pittsburgh-based GFS

Finally, the basic mechanisms of the GFS adopted in this work for SMD are presented:

- *Coding of the fuzzy rule base*. Each chromosome encodes an entire fuzzy rule base. The rules are coded by integer numbers that represent the index of fuzzy sets that appear in the antecedent and consequent part of the rule. Each rule contains another integer number to indicate the connector between the antecedents (“0” for the OR connector and “1” for the AND connector). It is necessary to define previously the number of fuzzy rules to be coded in the chromosomes.
- *Initial population*. The initial population is randomly generated.
- *Fitness function*. The GFS takes a decision every 250 ms. Two types of error can appear: an audio frame is labeled as speech when it is a music frame (Music as Speech Error, MSE) and the opposite (Speech as Music Error, SME). In order to design and evaluate the GFS, the fitness function considers both types of error:

$$Ev = MSE + SME \tag{17}$$

- *Genetic operators*. The genetic operators used in this work are one-point crossover, stochastic universal sampling selection and mutation. Mutation involves substituting an existing rule by another randomly generated one. All these genetic operators work according to the elitist strategy.
- *Stopping condition*. In computer simulations, the maximum number of generations has been used as the stopping criterion.

We have chosen four 1 s-length texture windows as a trade-off solution between the classification accuracy rate and the computation cost to obtain the knowledge base in the GFS. As the number of 1 s-length windows increases, the classification rate asymptotically grows until a maximum is reached, but the computational cost for obtaining the knowledge base exponentially grows. Therefore, the chosen value allows achieving high classification rates with a feasible computational cost.

4 Experiment evaluation

The classification results are calculated using a ten-fold cross-validation evaluation where the dataset to be evaluated is randomly partitioned so that 10% is used for testing and 90% is used for training. The process is iterated with different random partitions and the results are averaged. The results presented in this section are obtained with 50 iterations. This ensures that the calculated accuracy will not be biased because of a particular partitioning of the whole dataset for training and testing. If the datasets are representative of the two audio classes (speech and music), the results here presented are also indicative of the classification performance with real-world unknown signals. The ± sign in the tables of Section 4 shows the standard deviation of classification accuracy for the iterations.

First, we assess the SMD capability of the proposed F0-based feature set. To achieve such goal, comparison with the timbral features proposed in [42] is performed. The following specific features are used in [42] to represent timbral texture: SC, SR, SF, ZC and MFCC. The vector for describing all timbral texture features consists of the mean, variance and skewness computed over each 1s-length texture window.

The values considered in the proposed SMD approach for the length of the analysis and texture windows (23.22 ms and 1 s, respectively) are widely used in other related works dealing with SMD [2, 8, 42, 44]. Although other different values are possible, the chosen values provide a good trade-off between accuracy, complexity and delay. They have also been used for computing the classical features explained in Section 3.1 in order to perform a fair comparison between all features (the proposed features and the classical ones). All compared features are computed using the 1s-length texture window, and updated every 250 ms.

Table 3 shows the classification accuracy percentage when the proposed F0-based feature set is compared to the timbral features. The robustness of the F0-derived features against different widely used classifiers (k-NN, GMM, NN and SVM) is also assessed in Table 3.

From results of Table 3, a detailed comparative analysis of all considered features is accomplished, which provides reliable knowledge about the best ranked features for SMD. The results in Table 3 evidence that the proposed feature set always performs better than all tested timbral features for SMD, showing an improvement close to 2% compared to the MFCC (widely used in pattern recognition tasks). The results in Table 3 for timbral texture features are in line with the results given in other published works dealing with the same problem [2, 29, 30, 42]. As expected, the classification accuracy percentages when using the SPR classifiers are usually lower than that obtained by NN and SVM. Further, NN slightly outperforms SVM, giving rise to an accuracy rate of about 95% for the proposed features.

To corroborate the results in Table 3, four Receiver Operating Characteristics (ROC) curves have been generated, one for each classifier, which provide additional information regarding the performance of all considered features and classifiers. These curves are shown in Figs. 16, 17, 18 and 19.

Table 3 Classification accuracy percentage

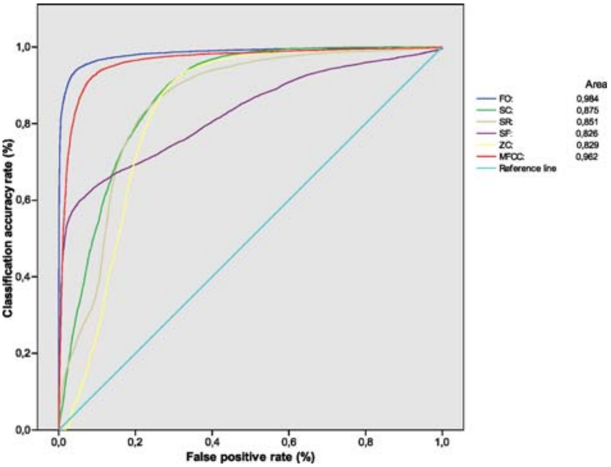
F0-based features vs. timbral features for different classifiers

Classifier	Feature	Speech (%)	Music (%)	Global (%)
k-NN	SC	77.83±3.8	85.86±4.5	81.84±3.3
	SR	72.89±4.6	81.97±6	77.43±3.5
	SF	71.26±3.6	80.39±6.8	75.83±4.3
	ZC	70.71±4.5	82.97±6.8	76.79±4.4
	MFCC	86.15±3.5	97.22±1	91.69±2
	F0	90.36±2.6	96.24±1.7	93.30±1.6
GMM	SC	85.99±4.5	86.79±7	86.39±4.6
	SR	89.64±2.6	75.07±8	82.36±4.4
	SF	71.36±6.5	75.47±10	73.42±7
	ZC	75.95±6.6	81.27±7	78.61±5.3
	MFCC	86.60±4	98.64±1	92.62±2.1
	F0	89.72±2.9	96.85±1.7	93.28±1.8
NN	SC	80.17±4	87.02±5.3	83.59±4
	SR	78.95±4.8	86.70±6	82.83±3.8
	SF	77.41±5	76.86±10	77.14±6.6
	ZC	77.77±5	83.41±7	80.59±4.8
	MFCC	87.70±3.4	98.95±0.7	93.33±1.9
	F0	92.39±2.5	97.69±1.3	95.04±1.5
SVM	SC	81.00±4	79.28±10	80.14±6
	SR	83.81±2.7	76.67±8	80.24±4.6
	SF	73.47±5	71.98±10	72.73±5.8
	ZC	75.21±5	78.58±10	76.89±6
	MFCC	86.94±3.7	98.54±0.7	92.74±1.8
	F0	92.39±2.7	96.26±2.5	94.33±2

Figures 16, 17, 18 and 19 confirms the results in Table 3, showing that the proposed F0-based features always lead to the best classification rates, regardless of the considered classifier. These figures also show the ranking between the timbral features, the MFCC being the best ranked timbral feature in all cases.

We are also interested in knowing the discrimination capability of each single F0-derived feature. Therefore, Table 4 shows the classification accuracy percentage

Fig. 16 ROC curves for all considered features when a k-NN classifier is used



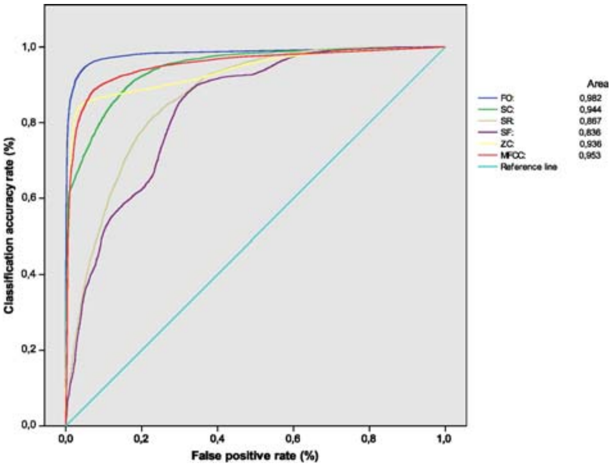


Fig. 17 ROC curves for all considered features when a GMM classifier is used

of each single F0-derived feature, as well as the classification rate of the F0-based feature set. The results in Table 4 have been obtained by applying the widely used k-NN SPR classifier.

As can be seen in Table 4, the discrimination capability of the F0-derived features is related to the histograms shown in Figs. 5–11. Given a F0-derived feature, the more separated the feature histograms are, the higher the accuracy rate is. From Table 4, it also results that most of the meaningful information is provided by only three features: the average of the estimated F0 ($F0_{av}$), the dynamic range of the estimated F0 (D_{F0}) and the number of notes (N_{note}). When the three features are combined, the classification percentage goes up to about 90%, which is close to the value obtained when all F0-derived features are considered (93.30%).

Further, we are interested in knowing the improvement (if any) due to the F0-based features when they are combined with each one of the timbral features. Table 5 shows the improvement achieved by the F0-derived features when they are used

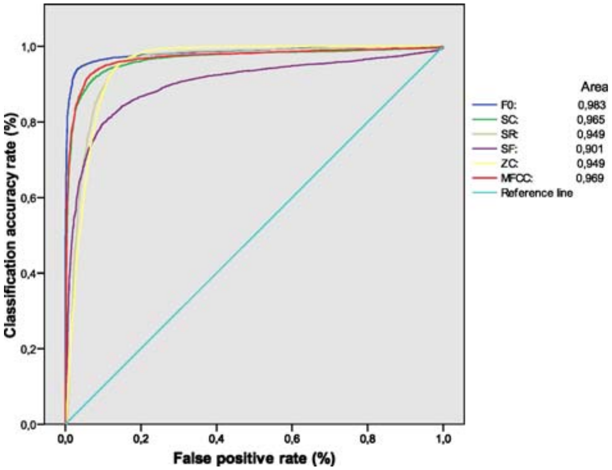


Fig. 18 ROC curves for all considered features when a NN classifier is used

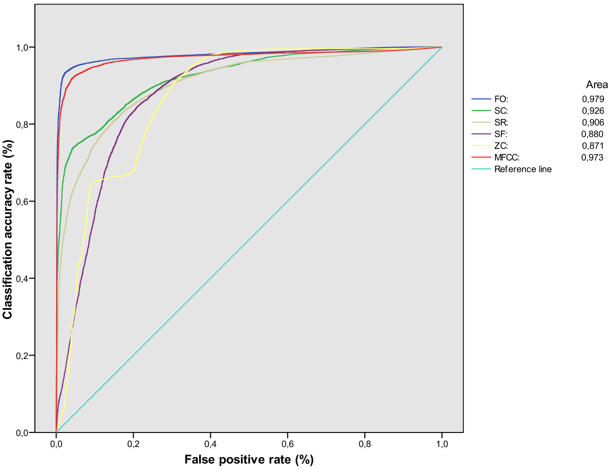


Fig. 19 ROC curves for all considered features when a SVM classifier is used

together with the timbral features. As before, the results in Table 5 have been obtained by using the k-NN SPR classifier.

As can be seen in Table 5, a great increase in the classification accuracy rate is always produced when combining each timbral feature with the F0-derived ones. Consequently, the F0-derived features (pitch features) constitute a good complement to the timbral features for SMD. Note that the results obtained when combining each timbral feature with the F0-derived ones are always above the reference (93.30%). The best results are obtained by combining MFCC and the F0-derived features, with an average classification percentage of about 97%.

Now, we investigate the influence of the GFS on the classification accuracy rate for SMD. Table 6 shows the improvement in the accuracy rate (averaged results) due to the inclusion of the GFS within the proposed two-stage classification scheme regarding the case of using only the classifier (first stage). The robustness of the GFS against different widely used classifiers (k-NN, GMM, NN and SVM) has also been assessed in Table 6. The results in Table 6 have been obtained when using the proposed F0-based feature set, the Pittsburgh learning algorithm is applied to evolve the GFS, and the same audio database has been considered for testing the different classification schemes.

Table 4 Discrimination capability of each single F0-derived feature

Feature	Speech (%)	Music (%)	Global (%)
D_{Ap}	68.53±2.6	61.21±7	64.87±4
$F0_{av}$	73.50±3	68.67±8	71.08±4.7
D_{F0}	70.15±2.9	64.11±5.5	67.13±3.6
ND_{max}	75.98±1.8	51.21±5.3	63.59±2.6
N_{note}	86.64±2.7	62.86±10	74.75±5
VR	65.00±2.8	56.46±7	60.73±3.6
Ap_{av}	59.35±2.8	64.07±6.8	61.71±3.8
$F0_{av}+D_{F0}+N_{note}$	87.61±2.6	93.21±1.4	90.46±1.5
F0 feature set	90.36±2.6	96.24±1.7	93.30±1.6

Table 5 Improvement (%) due to the F0-based features when combined with the timbral features

Feature	Speech (%)	Music (%)	Global (%)
Reference (F0)	90.36±2.6	96.24±1.7	93.30±1.6
SC + F0	94.20±2.9	97.52±2.6	95.86±2
SC	77.83±3.8	85.86±4.5	81.84±3.3
Difference	16.37	11.66	14.02
SR + F0	93.86±2.6	97.28±1.6	95.57±1.5
SR	72.89±4.6	81.97±6	77.43±3.5
Difference	20.97	15.31	18.14
SF + F0	91.30±2.6	96.32±1.7	93.81±1.7
SF	71.26±3.6	80.39±6.8	75.83±4.3
Difference	20.04	15.93	17.98
ZC + F0	93.18±2	97.30±2.2	95.24±1.7
ZC	70.71±4.5	82.97±6.8	76.79±4.4
Difference	22.47	14.33	18.45
MFCC + F0	95.80±3	98.42±0.7	97.11±1.7
MFCC	86.15±3.5	97.22±1	91.69±2
Difference	9.65	1.2	5.42

We have applied the Pittsburgh approach to the SMD problem with the following parameter specifications:

- Initial population: 20 knowledge bases.
- Crossover probability: 0.4.
- Mutation probability: 0.1.
- Stopping condition: 1000 generations.

The results in Table 6 show the good behavior of the proposed two-stage classification scheme for SMD, which implies that GFS can be an interesting component to be used in pattern recognition or classification tasks. As expected, the GFS always leads to a better performance of the speech/music discriminator. From results in Table 6, we can say that an average reduction about 2.35% in the total error rate has been achieved. The GFS leads to similar improvement in the accuracy rate for all considered classifiers, which evidences the robustness of the GFS against the

Table 6 Improvement (%) due to the inclusion of the GFS within the proposed two-stage classification scheme

Feature	Speech (%)	Music (%)	Global (%)
k-NN	90.36±2.6	96.24±1.7	93.30±1.6
k-NN + GFS	94.82±2	97.91±2.1	96.36±1.5
Difference	4.46	1.67	3.06
GMM	89.72±2.9	96.85±1.7	93.28±1.8
GMM + GFS	93.26±2.1	97.55±2.2	95.41±1.5
Difference	3.54	0.7	2.13
NN	92.39±2.5	97.69±1.3	95.04±1.5
NN + GFS	95.31±2.2	98.84±0.9	97.08±1.2
Difference	2.92	1.15	2.04
SVM	92.39±2.7	96.26±2.5	94.33±2
SVM + GFS	95.22±3	97.85±1.3	96.53±2.2
Difference	2.83	1.59	2.2

Table 7 Speech/rap music discrimination: comparison between F0 features and MFCC

Classification scheme	Feature	Speech (%)	Rap music (%)	Global (%)
k-NN	F0 features	88.98±2.8	83.97±2.3	86.47±1.2
	MFCC	77.17±2.6	88.12±4	82.65±2.2
	Difference	11.81	−4.15	3.82
k-NN + GFS	F0 features	89.96±3.3	87.65±1.5	88.81±1.8
	MFCC	79.97±3.5	90.92±4.2	85.44±2.6
	Difference	9.99	−3.27	3.36

type of classifier. The highest classification accuracy percentages correspond to the NN+GFS scheme. An average accuracy percentage above 97% is achieved by such classification scheme.

Finally, comparison between the proposed F0-derived features and MFCC has been performed to discriminate between speech and rap music. The results, presented in Table 7, have been obtained in two cases: with and without using the GFS in the classification scheme. In both cases, the k-NN SPR classifier has been considered.

As can be seen in Table 7, the proposed F0-derived features are especially well-suited to discriminate between speech and rap music, because they rely on the pitch rather than on the timbral texture. From Table 7, it results that the F0-derived features outperform MFCC for speech/rap music discrimination, the differences being close to 4% and 3.5% for the first (k-NN) and second (k-NN+GFS) cases, respectively, which involves that the GFS does not give rise to further difference between the F0-derived features and MFCC. Note that the difference between the F0-derived features and MFCC is about 2% for the general case of SMD. Therefore, the F0-derived features can be an interesting alternative to MFCC, especially for the particular case of speech/rap music discrimination.

To corroborate the results in Table 7, the ROC curves shown in Fig. 20 have been generated, which provide additional information regarding the performance of both features (F0 and MFCC) for discriminating between speech and rap music.

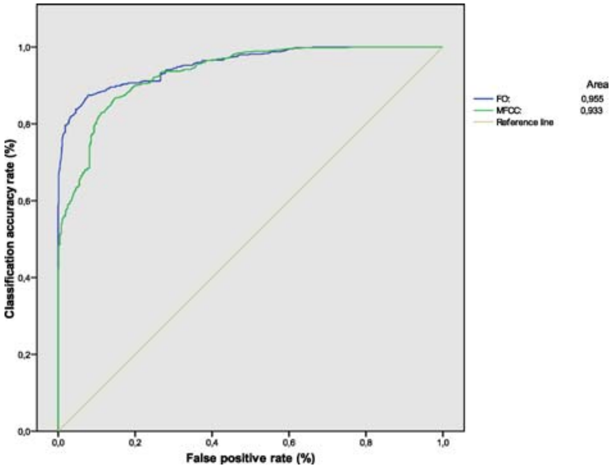


Fig. 20 Speech/rap music discrimination: ROC curves to compare the F0 features with MFCC

Figure 20 confirms the results in Table 7, showing that the F0-based features outperform MFCC, mainly for speech/rap music discrimination.

5 Conclusions

The paper presents an effective and robust approach for SMD based on a new musically-inspired feature set, which is computed from F0 estimation. The approach is completed with a two-stage classification scheme composed of a classifier followed by a GFS. The experimental evaluation compares the proposed F0-based feature set to other features commonly used in audio classification tasks when using different widely used classifiers (k-NN, GMM, NN, SVM). Results show the good performance of the proposed feature set for SMD, since the classification accuracy percentage is higher to that obtained by MFCC (widely used for pattern recognition tasks). An improvement close to 2% is achieved on average. The performance of each single F0-derived feature has also been assessed. To evaluate the performance of the proposed two-stage classification scheme, different widely used classifiers (k-NN, GMM, NN, SVM) have been considered. The two-stage classification scheme (classifier+GFS) provided good results for all considered classifiers. The best results (about 97%) were obtained by the NN+GFS scheme. The classification accuracy percentage achieved by our approach is above 95% for all considered classifiers and a wide range of audio styles, which shows the good performance of the proposed approach to discriminate between speech and music. Finally, the proposed F0-derived features have been evaluated to discriminate between speech and rap music, showing excellent results.

References

1. Booker L (1982) Intelligent behaviour as an adaption to the task environment. Ph.D. Thesis, University of Michigan
2. Burred JJ, Lerch A (2004) Hierarchical automatic audio signal classification. J Audio Eng Soc 52:724–739
3. Carey MJ, Parris ES, Lloyd-Thomas H (1999) A comparison of features for speech, music discrimination. In: Proc. IEEE ICASSP'99, Phoenix, USA. IEEE, Piscataway, pp 1432–1435
4. Cheveigne A, Kawahara H (2002) YIN, a fundamental frequency estimator for speech and music. J Acoust Soc Am 111(4):1917–1930, April
5. Cordon O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems. Evolutionary tuning and learning of fuzzy knowledge bases. Advances in fuzzy systems. Applications and theory, vol 19. World Scientific, Singapore
6. Davis S, Mermelstein P (1980) Experiments in syllable-based recognition of continuous speech. IEEE Trans Acoust Speech Signal Process 28:357–366, Aug
7. Duda R, Hart P, Stork D (2000) Pattern classification. Wiley, New York
8. El-Maleh K, Klein M, Petrucci G, Kabal, P (2000) Speech/music discrimination for multimedia applications. In: Proc. IEEE ICASSP'2000, vol 6. IEEE, Piscataway, pp 2445–2448
9. Every MR (2008) Discriminating between pitched sources in music audio. IEEE Trans Audio Speech Language Process 16(2):267–277, Feb
10. Exposito JEM, Galan SG, Reyes NR, Candeas PV (2007) Audio coding improvement using evolutionary speech/music discrimination. In: IEEE international fuzzy systems conference, (FUZZ-IEEE), July 2007. IEEE, Piscataway, pp 1–6
11. Ezzaidi H, Rouat J (2007) Comparison of the statistical and information theory measures: application to automatic musical genre classification. In: IEEE Workshop on Machine Learning for Signal Processing, August 2007. IEEE, Piscataway, pp 241–246

12. Fujihara H, Kitahara T, Goto M, Komatani K, Ogata T, Okuno HG (2006) F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search acoustics. In: Proc. IEEE int. conf. on acoustic, speech and signal processing (ICASSP), May 2006, vol 5. IEEE, Piscataway, pp 14–19
13. Garau G, Renals S (2008) Combining spectral representations for large-vocabulary continuous speech recognition. IEEE Trans Audio Speech Lang Process 16(3):508–518, March
14. Garcia Arnal Barbedo J, Lopes A (2007) Speech/music discriminator based on multiple fundamental Frequencies Estimation. IEEE Latin America Trans 5(5):294–300, Sept
15. Gong C, Xiong-wei Z (2006) The application of speech/music automatic discrimination based on gray correlation analysis. In: 5th IEEE international conference on cognitive informatics (ICCI), July 2006, vol 1. IEEE, Piscataway, pp 68–72
16. Harb H, Chen L (2003) Robust speech music discrimination using spectrum's first order statistics and neural networks. Proc IEEE Int Symp Signal Process Appl 2:125–128
17. Hess W (1983) Pitch determination of speech signals. Springer, Berlin
18. Hess WJ (1992) Pitch and voicing determination. In: Furui S, Sohndi MM (eds) Advances in speech signal processing. Marcel Dekker, New York, pp 3–48
19. Hirose K, Iwano K (2000) Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition. In: Proc. IEEE int. conf. on acoustics, speech, and signal processing (ICASSP), June 2000, vol 3. IEEE, Piscataway, pp 1763–1766
20. Ji-Soo Keum, Hyon-Soo Lee (2006) Speech/music discrimination using spectral peak feature for speaker indexing. In: International symposium on intelligent signal processing and communications (ISPACS), Dec. 2006. IEEE, Piscataway, pp 323–326
21. Karneback S (2001) Discrimination between speech and music based on a low frequency modulation feature. In: European conf. on speech comm. and technology, Alborg, 3–7 September 2001, pp 1891–1894
22. Kawahara H, Masuda-Katsuse I, de Cheveigne A (1999) Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Commun 27:187–207
23. Logan B (2000) Mel frequency cepstral coefficients for music modeling. In: Proc. int. symp. music information retrieval (ISMIR), Plymouth, 23–25 October 2000
24. Lu L, Zhang H, Jiang H (2002) Content analysis for audio classification and segmentation. IEEE Trans Speech Audio Process 10(7):504–516, October
25. Malik H, Khokhar A, Ansari R, Cappe de Baillon B (2002) Predominant pitch contour extraction from audio signals. In: IEEE International Conference on Multimedia and Expo (ICME), August 2002, vol 2. IEEE, Piscataway, pp 257–260
26. Matsunaga S, Mizuno O, Ohtsuki K, Hayashi Y (2004) Audio source segmentation using spectral correlation features for automatic indexing of broadcast news. In: Proc. EUSIPCO, Vienna, Sep 2004, pp 2104–2106
27. Minami K, Akutsu A, Hamada H, Tonomura Y (1998) Video handling with music and speech detection. IEEE Multimed 5(3):17–25
28. Molla KI, Hirose K, Minematsu N, Hasan K (2007) Voiced/unvoiced detection of speech signals using empirical mode decomposition model. In: Int. Conf. on Information and Communication Technology (ICICT), March 2007. IEEE, Piscataway, pp 311–314
29. Muñoz-Exposito JE, Ruiz-Reyes N, Garcia-Galan S, Vera-Candeas P (2006) New speech/music discrimination approach based on warping transformation and ANFIS. J New Music Res 35:237–247, Dec
30. Muñoz-Exposito JE, Ruiz-Reyes N, Garcia-Galan S, Vera-Candeas P (2007) Adaptive network-based fuzzy inference system vs. other classification algorithms for warped LPC-based speech/music discrimination. Eng Appl Artif Intell 20:783–793, Sep
31. Panagiotakis C, Tziritas G (2005) A speech/music discriminator based on RMS and zero–crossings. IEEE Trans Multimedia 7:155–166, Feb
32. Paradzinets A, Kotov O, Harb H, Chen L (2007) Continuous wavelet-Like transform based music similarity features for intelligent music navigation. In: International workshop on content-based multimedia indexing (CBMI), Bordeaux, June 2007, pp 165–172
33. Politis D, Linardis P, Tsoukalas I (2000) An audio signatures indexing scheme for dynamic content multimedia databases. In: 10th Mediterranean electrotechnical conference (MELECON), vol 2. IEEE, Piscataway, pp 725–728
34. Qiao RY (1997) Mixed wideband speech and music coding using a speech/music discriminator. In: Proc. IEEE TENCON. IEEE, Piscataway, pp 605–608

35. Rentzos D, Vaseghi S, Qin Yan, Ching-Hsiang Ho (2004) Voice conversion through transformation of spectral and intonation features. In: IEEE international onference on acoustics, speech, and signal processing (ICASSP), May 2004, vol 1. IEEE, Piscataway, pp 21–24
36. Richard G, Ramona M, Essid S (2007) Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), April 2007, vol 2. IEEE, Piscataway, pp 461–464
37. Saitou T, Goto M, Unoki M, Akagi M (2007) Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices. In: IEEE workshop on applications of signal processing to audio and acoustics, October 2007. IEEE, Piscataway, pp 215–218
38. Saunders J (1996) Real-time discrimination of broacast speech/music. In: Proc. IEEE ICASSP'96, Atlanta, May 1996, pp 993–996
39. Scheirer E, Slaney M (1997) Construction and evaluation of a robust multifeature speech/music discriminator. In: Proc. IEEE ICASSP'97, Munich, April 1997, pp 1331–1334
40. Smith SF (1980) A learning system based on genetic adaptive algorithms. Ph.D. thesis, University of Pittsburgh
41. Tancerel L, Ragot S, Ruoppila VT, Lefebvre R (2000) Combined speech and audio coding by discrimination. In: Proc. IEEE workshop on speech coding. IEEE, Piscataway, pp 17–20
42. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Trans Speech Audio Process 10(5)
43. Venturini G (1992) SIA: a supervised inductive algorithm with genetic search for learning attribute based concepts. In: Proc. European conference on machine learning (ECML'92), Viena. Springer, Heidelberg, pp 280–296
44. Wang WQ, Gao W, Ying DW (2003) A fast and robust speech/music discrimination approach. In: Proc. 4th pacific rim conference on multimedia, vol 3. IEEE, Piscataway, pp 1325–1329
45. Wang J, Wu Q, Deng H, Yan Q (2008) Real-time speech/music classification with a hierarchical oblique decision tree. In: IEEE international conference on acoustics, speech and signal Processing (ICASSP), March 2008. IEEE, Piscataway, pp 2033–2036
46. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
47. Zhang T, Kuo J (2001) Audio content analysis for online audiovisual data segmentation and classification. IEEE Trans Speech Audio Process 9(4)

N. Ruiz-Reyes was born in Linares (Jaen), Spain, in 1967. He received the MSc and PhD degrees in Telecommunication Engineering from the Technical University of Madrid (UPM) and the University of Alcala, in 1993 and 2001, respectively. Since 1998, he is Associate Professor in Signal Processing and Communications at the Telecommunication Engineering Department of the University of Jaen. His areas of research interest are Signal Processing and its Applications to Communications, Speech and Audio Analysis, Electrical and Biomedical Engineering and Ultrasonic NDT. He is co-author of more than 100 papers, and is involved in research projects of the Spanish Ministry of Science and Education, European Commission, and private companies.

P. Vera-Candeas was born in Madrid, Spain, in 1976. He received the M.S. degree in Telecommunication Engineering from the University of Málaga (UMA) in 2000 and a Ph.D. degree from the University of Alcala in 2006. Since 2000, he has worked at the Telecommunication Engineering Departament of the University of Jaén. Nowadays, he is an Associate Professor in Signal Processing and Communications Area. His areas of research interest are Signal Processing and its Applications to Audio Analysis and Ultrasonic NDT. He has been involved in research projects of the Spanish Ministry of Science and Education (MEC) and private companies.

J. E. Muñoz was born in Estepona (Málaga), Spain, in 1970. He received the MSc degree in Telecommunication Engineering from the University of Malaga in 1995. Since 2003, he is assistant professor in Telematics at the Telecommunication Engineering Department of the University of Jaen. His area of research interest is Speech and Audio Analysis. He is involved in research projects of the Spanish Ministry of Science and Education.

S. García-Galán was born in Lahiguera (Jaen), Spain, in 1969. He received the MSc and PhD degrees in Telecommunication Engineering from the University of Malaga (UMA) and the Technical University of Madrid (UPM), in 1995 and 2004, respectively. Since 1995, he is with the Telecommunication Engineering Department of the University of Jaen. His areas of research are engineering applications and artificial intelligence.

F. J. Cañadas was born in Linares (Jaén), Spain, in 1977. He received the M.S. degree in Telecommunication Engineering from the University of Málaga (UMA) in 2004. During 2004-2006, he worked as engineer in an Europe Research Project (INTUITION Network Excellence). Nowadays, he is an assistant professor at the Telecommunication Engineering Departament of the University of Jaén. His areas of research interests include automatic music transcription, multi-pitch estimation and sound source separation in polyphonic music signals. His PhD Thesis is focused on multi-pitch estimation techniques and single-channel source separation.