New Feature For Automatic Speech/Music Discrimination

Jayme Garcia Arnal Barbedo[1], Amauri Lopes[1]

[1]Department of Communications – FEEC – UNICAMP
C.P. 6101 – CEP 13.081-970 – Campinas – SP – Brazil
{jgab,amauri}@decom.fee.unicamp.br

Abstract. *This paper presents a new mechanism for automatic speech/music discrimination (SMD). Such feature is based on the concept of multiple fundamental frequencies. The performance of the feature in terms of correct classifications is evaluated for a wide variety of audio signals, and factors such as computational complexity and robustness are also investigated. The results are compared to those ones reached by previous techniques.*

1. Introduction

In the last decade, the demand for techniques able to automatically discriminate between music and speech signals has risen dramatically. There are several technologies that can benefit from the advances achieved in this area, as Automatic Speech Recognizers and Automatic Music Transcriptors, which must be fed with the appropriate signals. Other applications for speech/music discrimination techniques are the hearing devices and the automatic selection of FM radio stations.

The first researches in SMD have reached about 95% of accuracy in their experiments [Saunders 1996], [Scheirer and Slaney 1997]. Several works have followed those early proposals [Carey et al. 1999], [Cho et al. 2003], [El-Maleh et al. 2000], [Jarina et al. 2002], [Lu et al. 2002], most of them presenting accuracy between 92% and 98%.

There are two characteristics that are common to all speech/music discriminators: 1) the great number of features extracted from the signals and 2) the use of techniques such as Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) and k-Nearest Neighbors (KNN) to combine such features. These approaches have a number of drawbacks associated: high computational and programming complexity and a large number of degrees of freedom, reducing the robustness of the approach. The technique presented here overcomes most of such limitations, since it is extremely simple to implement, requires little computational resources and is composed of only one feature, meaning that it is very robust to a wide range of situations.

2. Feature Extraction

Before the feature extraction itself, the signal must be properly formatted to fit the process requirements. The first step is to identify if the signal is monophonic or has more than one channel. In the first case, no action is taken; otherwise, the channels must be combined using a simple arithmetic average. In this work, the signals are sampled at 48 kHz and divided into frames of 1,024 samples, corresponding to time intervals of 21.3 ms. The frames are 50% superposed and are weighted by a Hanning window.

The strategy presented here is based on the signal main fundamental frequencies

(*f0*) detection. Since the signals analyzed here have several sound sources, some kind of processing is necessary to ease the detection of the *f0* of each sound source. Most of the techniques described in the following were inspired in the multipitch analysis model presented in [Tolonen and Karjalainen 2000]. The strategy is illustrated in Figure 1.

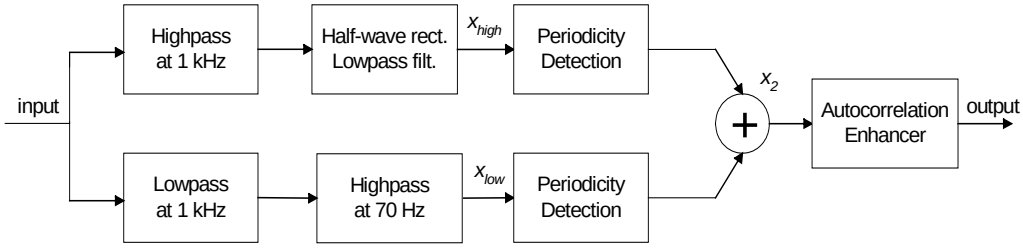

Figure 1. Strategy to estimate multiple fundamental frequencies.

In Figure 1, the input consists of the signal frames, and is divided into two bands (low and high frequencies) by a filtering process with cut-off at 1 kHz. The low frequency portion is also submitted to an extra filtering to block frequencies below 70 Hz. The high frequency portion is then submitted to a half-wave rectification. After that, it is lowpass filtered with a filter similar to that used to determine the low frequency portion.

The periodicity detection, which results in x_2 in Figure 1, is based on the concept of "generalized autocorrelation", and is given by

$$x_2(n) = \mathrm{IDFT}\left[\left|\mathrm{DFT}\left(x_{low}(n)\right)\right| + \left|\mathrm{DFT}\left(x_{high}(n)\right)\right|\right], \quad (1)$$

where DFT and IDFT represent the Discrete Fourier Transform and its inverse, respectively, and n is the time index.

The peaks of the autocorrelation given by $x_2(n)$ are good indicators of potential fundamental frequencies. However, since the signals have multiple sound sources, $x_2(n)$ can show lots of spurious information that can potentially lead to wrong estimations. To reduce the amount of unwanted information, a peak pruning technique is applied. Firstly, a half-wave rectification is applied to clip negative values of $x_2(n)$. The resulting function is expanded in time by a factor-two oversampling and subtracted from the clipped autocorrelation function. This procedure eliminates all peaks with twice the time lag of a higher amplitude reference peak. The technique also removes near-zero values of the autocorrelation function. In the present work, the procedure was applied to eliminate peaks with twice and three times the time lag of the reference peaks.

The next step is to identify the three main peaks of the enhanced autocorrelation function for each frame. Those three peaks are taken as the *f0* of the three main sound sources of the frame. If less than 3 sources are present, only one or two peaks will be identified. The estimated frequencies are then converted to the MIDI scale, according to the procedure described in [Tzanetakis and Cook 2002] and given by

$$m = 12\log_2(f/440) + 69, \quad (2)$$

where f is the frequency in Hz and m is the MIDI number. All frequencies with same MIDI number are counted over all frames, generating a histogram whose bins are the MIDI notes.

It was observed that most of the analyzed speech signals have frequencies whose corresponding MIDI numbers are equal or greater than 100, while music signals rarely present such high frequencies. This probably occurs because of short speech segments with low energy and high frequency whose period is defined enough to be detected by the detection procedure. Next section will describe how this information is explored in order to provide a reliable differentiation between speech and music signals.

3. Tests and Results

The database used in the tests is composed by 2,587 wav-format audio files sampled at 48 kHz and quantized with 16 bits, and it is divided into speech and music files.

As commented before, it was observed that speech signals often presents higher *f0* than music signals. Therefore, the first task is determining the proportion of high frequencies that leads to the best discrimination between speech and music signals. It was observed that the following rule led to the best results: given a histogram, if the proportion of MIDI values equal or higher than 100 is over 0.1%, the signal is considered speech; otherwise, the signal is considered music. Table 1 summarizes the results obtained.

Table 1. Results

Group	Right Classification Percentage
Speech (all files)	94.01%
Speech (only files without environmental noise)	96.05%
Music (all files)	93.63%
Music (without rap files)	94.87%

As can be seen, the percentage of right classifications lies between 93 and 96%. In the case of speech signals, the performance is very good even when strong environmental noise (street, office, nature sounds) is present, indicating that the strategy is very robust to extreme conditions. In the case of music, it was observed that for some musical genres, like classical and rock, the percentage of right classification is near 100%, while for musical genres that have several elements of actual speech, like rap, the correctness can drop to values below 80%.

Comparing the proposed procedure with some methods in the literature, one can conclude that there are previous techniques presenting slightly better discrimination accuracy. The best results were achieved by [Lu et al. 2002], which used a very complex strategy to reach a precision of about 98%. However, when the comparison is done taking into account not only the discrimination but also the robustness of the discriminator to unexpected situations and the computational effort demanded by the method, the conclusion is that the strategy here presented is clearly superior. As commented before, since this proposal depends only on one feature, it presents a great robustness to unexpected situations, as can be observed in Table 2. Additionally, it demands low computational effort, indicating that the procedure can be used in real time applications, even when the available computational resources are scarce.

4. Conclusions

This paper presented a new strategy to discriminate between speech and music signals.

The technique consists of the extraction of a single feature based on the concept of multiple fundamental frequencies.

The performance of the strategy in terms of correct estimates is competitive with previous works. Additionally, it presents a clearly advantage in terms of robustness and computational complexity. The characteristics of this technique make it appropriate to be used in applications where potentially problematic conditions, like degradations and environmental noise, are expected. Finally, it can be used in real-time applications.

There are several possible directions for future research. A possible enhancement can be achieved with the improvement of the process used to estimate the multiple fundamental frequencies. Another interesting line of research is trying to combine the strategy here presented with another successful techniques. At last, some new features based on the histograms generated in the fundamental frequency estimation can be created and combined, in such a way the results are improved without adding significant computational effort. Two novel features that have already been implemented and that have shown good results are the ratio between the amplitude of the histogram peak and the histogram sum, and a measure to the variation between the bin amplitudes of consecutive MIDI notes.

Acknowledgements

The authors would like to thank Fapesp for supporting this research under grant n. 04/08281-0.

References

Carey, M.J.; Parris E.S.; Lloyd-Thomas, H. (1999) “A comparison of features for speech, music discrimination”, *Proc. of ICASSP99*, pp. 149-152.

Cho, Y.-C.; Choi, S.; Bang, S.-Y. (2003) “Non-negative component parts of sound for classification”, *Proc. IEEE Int. Symp. Signal Processing and Information Technology*, Darmstadt, Germany.

El-Maleh, K.; Samouclian, A.; Kabal, P (1999). “Frame-Level Noise Classification in Mobile Environments”, *Proc. IEEE Conf. Acoustics, Speech, Signal Proc.*, Phoenix, AZ, USA.

Jarina, R.; O'Connor, N.; Marlow, S. (2002) “Rhythm Detection for Speech-Music Discrimination in MPEG Compressed Domain”, *Proc. of the IEEE 14th International Conference on Digital Signal Processing 2002*, Santorini, Greece, pp. 129-132.

Lu, L.; Zhang, H.-J.; Jiang, H. (2002) “Content Analysis for Audio Classification and Segmentation”, *IEEE Trans. on Speech and Audio Proc.*, vol. 10, no. 7, pp. 504-516.

Saunders, J. (1996) “Real-Time Discrimination of Broadcast Speech/Music”, Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, pp 993-996.

Scheirer, E.; Slaney, M. (1997) “Construction and Evaluation of a Robust Multifeature Speech-Music Discriminator”, *Proc. of ICASSP97*, pp. 1331-1334, Munich, Germany.

Tolonen, T.; Karjalainen, M. (2000) “A Computationally Efficient Multipitch Analysis Model”, *IEEE Trans. on Speech and Audio Proc.*, vol. 8, no. 6, pp. 708-716.