Algorithm Selection for Automated Audio Classification based on Content

Ivo Draganov[1] and Krasimir Minchev[2]

Abstract – **In this paper an approach for algorithm selection for automated audio classification is proposed based on content. Three popular algorithms for speech vs. music discrimination are incorporated, namely the zero-crossing rate, average frame power and 4-Hz energy modulation which prove effective at different level depending on the content of the audio being processed. Extensive testing is done over various musical pieces from different countries and in different styles to evaluate the accuracy and time consumption for each one. Then based on registered levels and with statistics from the initial record for particular audio recording the most proper could be switched on for processing in a complete audio classification system. Obtained results re considered promising for future use on a wider scale.**

Keywords – **Audio Classification, Music Speech Discrimination, 4 Hz Energy Modulation, Zero-Crossing Rate, Average Frame Power.**

I. Introduction

Discrimination between speech and music signals is applied in various areas of speech signal processing, such as Voice Activity Detection (VAD). On this occasion, many solutions, both in the time domain and in the frequency, have been proposed. The most common are: 4 Hz energy modulation, entropy modulation, spectral center, spectral flow, and zero-crossing rate (ZCR). Less frequent are spectrum overturning, spectral centroid, spectral flux variation and others [1, 2]. We will look at some of them by striving to discriminate given signals with accuracy we will seek maximum knowledge of the content but at the same time we will discuss the complexity of the calculations. We will focus on the discrimination of speech and sound signals based on their energy. The energy distribution of speech and musical signals will be evaluated by looking at the frequency of zero-crossings, short-term energy, and 4 Hz modulation also considering the Minimum Energy Density (MED), Low Energy Frames (LEF), and the Modified Low Energy Ratio (MLER) [3].

In order to discriminate speech and musical signals, features that are different for both classes are used. A simple look at the waveform of a one-minute excerpt of a voice signal, pop music, classical or opera signal shows great differences between classes. The shape of the signal of the speech signal shows a great difference in energy and amplitude that none of the musical signals show. The highly compressed waveform of a song seems to lack dynamics, while the classic instrumental and the opera performance have a low amplitude peak and show great dynamic deviations.

Even if the classes are easy to identify in wave form, the exact position of the transitions is difficult to detect. An excerpt from pop music shows possibilities of the song dynamics stopping abruptly when vocals appear. The transition is difficult to be spotted [4].

The rock music samples are not the same in comparison to the four examples described above but the most important and interesting thing here are the changes. In all the examples containing music the square root of RMS never goes down to zero and does not deviate significantly from the average value. There are many snippets in the speech signals containing null or near null values of the frames variation and the differences are sharp. Investigating the spectra of the four examples reveals that all musical examples have a higher peak at the low frequencies, although the peak occurs at different frequencies. This peak is mostly responsible for the fundamental frequency of vocal components. The classical music that vocals are missing is not as sharp as the other three examples. The speech signal has more energy in the frequency range around 1 to 3 kHz, unlike the musical examples.

Sandars notes, "It is well known that the energy contour is capable of separating speech from music". Discrimination of speech from music is most likely to be based on the differences of the continuous change in the envelope of the energy curve. In speech signals the vocals and the consonants are clearly distinguishable and on the other hand, the shape of the musical signal that is more stable is also easily detectable. Furthermore, we are aware that the speech signal has a 4 Hz energy modulation characteristic, which coincides with the frequency of the syllabi. Sandars uses a simple method of discriminating speech and musical signals. He found that using statistics calculated on the basis of a zero crossing factor, he could reach a classification of about 90 percent. By adding more information on the energy contour, he upgraded the accuracy to 98% [5].

In this paper we are investigating wide range of sound recordings differing in content which may utilize a proper algorithm for discrimination of music vs. speech with higher reliability and in the same time in some cases with reduced computational complexity based on three audio metrics – the ZCR, average frame power (FPOW) and 4 Hz energy modulation (4Hz). In Section II the selected metrics are described. Then, experimental results follow in Section III with discussion on the overall performance of the tested implementations. A conclusion is made in Section IV.

[1]Ivo Draganov is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: idraganov@gmail.com.

[2]Krasimir Minchev is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: krasimirminchew@gmail.com

II. MEASURES USED

The first method we select for discriminating speech from music signals is the zero-crossing rate (ZCR). The frequency of zero crossings is a simple method of describing the content based on its most energetically pronounced frequency. We observe the definition of zero crossing frequency in the next expression [1]:

$$Zn = \sum_{m=-\infty}^{\infty} |sgn[x(m)]-sgn[x(m-1)]| \, w(n-m), \quad (1)$$

where:

$$sgn[x(n)] = 1, \quad x(n) \geq 0, \quad (2)$$
$$sgn[x(n)] = -1, \quad x(n) < 0. \quad (3)$$

And $w(n)$ is the window comprising of N number of frames:

$$W = 1/2N, \; 0 \leq n \leq N-1, \quad (4)$$
$$W = 0, \text{ otherwise.} \quad (5)$$

The short-term energy method is a little more complicated than the above-mentioned, but it's simpler to apply than finding the 4Hz measure. Given that the amplitude of the non-intuitive segments is noticeably lower than those of the speech segments, the short-term energy of the speech signals reflects the amplitude dispersion. By observing a speech signal, we can notice that the peak of the signal amplitude is noticeable as well as the fundamental frequency in the speech parts of the signal. This suggests that simple time processing techniques could derive useful information about signal characteristics. Most short-term processing techniques that derive features from the time domain Q[n] can be represented mathematically as [6]:

$$Qn = \sum_{m=-\infty}^{\infty} T[x(m)]w(n-m). \quad (6)$$

T is the transposed matrix, which may be linear or non-linear, $x(m)$ represents the information sequence, and $w(n-m)$ represents a time-limited window sequence. The energy of a discreet signal is defined by the expression:

$$E_t = \sum_{m=-\infty}^{\infty} s^2(m), \quad (7)$$

where E_t is the total energy and $s(m)$ is the discreet signal. For the calculation of Short-Term Energy, the signal is considered in short frames, whose size is usually between 10 and 30ms. It is necessary to take all samples in a frame from the signal from $m = 0$ to $m = N$-1, where N is the length of the frame. Then:

$$E_t = \sum_{m=-\infty}^{-1} s^2(m) + \sum_{m=-\infty}^{N-1} s^2(m) + \sum_{m=-\infty}^{\infty} s^2(m). \quad (8)$$

The samples value's are zero outside the frame. Therefore:

$$E_t = \sum_{m=0}^{N-1} s^2(m). \quad (9)$$

From (9) it could be estimated that the total energy of a frame for the signal from 0 to N-1 samples. The short-term energy is defined as the sum of the squares of the samples in a frame according to:

$$e(n) = \sum_{m=-\infty}^{\infty} [s_n(m)]^2. \quad (10)$$

After splitting in frames and windows, the N-th frame of the signal becomes $s(m).w(n-m)$ and therefore Eq. (10) becomes:

$$e(n) = \sum_{m=-\infty}^{\infty} [s(m).w(n-m)]^2, \quad (11)$$

where $w(n)$ is represented with a window function of limited duration, and n is the offset of the frame in number of samples. This shift may be as small as one sample or as large as one full frame.

The most complex for realization of the three methods is 4Hz modulation [7]. This method implies better discrimination and a higher rate of success than the previous two methods, but its implementation goes through more stages. First it is needed to derive MEL coefficients. Pre-emphasis is introduces, then the signal spectrum is re-emphasized and the constant component is removed. A low-order digital filter (most commonly a first order FIR filter) is attached to the input $x(n)$ so it is aligned in its spectrum:

$$H(z) = 1 - az^{-1}, \quad 0.9 < a < 1. \quad (12)$$

Then follows splitting in frames and the spectral analysis is performed on them. This is because human speech does not change much over time and can be treated as a quasi-static process. Very popular frame length is 20-30ms. Hamming window weighting of each frame takes place according to:

$$y(n) = x(n)w(n), \quad (13)$$

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi}{N-1}\right). \quad (14)$$

Every frame undergoes Discrete Fourier transform:

$$X(k) = \sum_{n=0}^{N-1} y(n) e^{-\frac{j2\pi}{N}} \quad 0 \leq n,k \geq N-1, \quad (15)$$

A set of triangular band-pass filters that simulate the characteristics of the human ear are applied to the signal spectrum. This process is called Mel filtering. Human hearing perception analyzes audible spectrum of groups based on the number of overlapping critical bands. These bands are allocated in such a way that the frequency resolution is high in the low frequency area and low in the high frequency area. Mel frequency is found from linear frequency based on:

$$f_m = 2525 \; x \; log\left(1 + \frac{1}{7}\right). \quad (16)$$

The energy of the filter bank is found by:

$$E = \sum_{k-1}^{N} |X(k)|^2 . \psi_i(k) \quad (17)$$

and after DCT over all MEL coefficients, finding their second derivatives and filtering the result in 40 channels with another FIR filter, all the energies contained in them are summed

together and the total energy is normalized with the average energy of the frame. The modulation is characterized by the variation of the filtered energy in dB per a second from the whole signal.

III. Experimental results

Experimental testing relies on a custom built database comprising of 5 speech and 10 music recordings in non-compressed format. Sampling frequency is 44100 Hz with resolution of the samples of 16 bits and all being stored in single channel wav files. The duration for all sounds captured is 1 minute.

Table I contains the ZCR, FPOW and 4Hz values obtained from the speech signals where *speechf* labels identify recording with a female voice and *speechm* – a male one.

Table I
Speech signals measures

Test set	ZCR	FPOW	4Hz
speechf1	0.0947	0.0018	45.5400
speechm2	0.0671	0.0034	45.3575
speechf3	0.0743	0.0036	36.4281
speechm4	0.0853	0.0035	41.4678
speechm5	0.0793	0.003	43.5407
Average	0.0801	0.0031	42.4668

In Table II the resulting values for the three parameters are given when found over the 10 music recordings some of which are typical folklore works from different geographical locations.

Table II
Music signals measures

Test set	ZCR	FPOW	4Hz
Rock	0.1971	0.0024	0.6245
Jazz	0.0971	0.0016	13.8121
Latino	0.2086	0.0035	0.5937
Folk	0.0899	0.0067	5.4985
Hindi	0.0916	0.0056	5.6968
Nordic	0.1083	0.0023	2.5773
African	0.0888	0.000837	64.8428
Chinese	0.0421	0.0036	25.4098
Russian	0.0612	0.0038	14.1327
Classic	0.0688	0.0029	11.7992
Average	0.1053	0.0033	14.4987

In order to discriminate speech and music signals based on these methods using full validation, we must use the average values for all files of a given type - musical or speech signals, and by them to calculate a threshold to use. To compute the threshold of a method, we collect the two average values of the musical and speech signals and divide them into two, and thus we get the threshold that will discriminate against the signals. For the zero crossing frequency, the values below the threshold, i.e. the signals with a lower frequency of zero crossings, will be defined as speech signals and those with higher values - as musical. Related values found for the thresholds are t_{ZCR} = 0.0928, t_{FPOW} = 0.0032, and t_{4Hz} = 28.4830.

The results from full validation of the speech database are shown in Table III.

Table III
Classification accuracy of speech signals

Measure	ZCR		FPOW		4Hz	
Threshold	0.0928		0.0032		28.4830	
Classify	Right	Wrong	Right	Wrong	Right	Wrong
speechf1	0	1	1	0	1	0
speechm2	1	0	0	1	1	0
speechf3	1	0	0	1	1	0
speechm4	1	0	0	1	1	0
speechm5	1	0	1	0	1	0
Accuracy,%	80	20	40	60	100	0

Let's take a look at the data in the second and third columns referring to the ZCR method. We determine the accuracy of the validation against the correct classification of the speech signal due to the threshold of the different methods that are used according to the method described above. After the classification of all speech files for the zero crossing frequency method, we obtain accuracy of 80% at full validation. Taking into account the simplicity of this method, its accuracy is very satisfactory. We continue with the examination of the fourth and fifth columns where the short-term energy method is described. When it is classified after all the speech files, we obtain the accuracy of the 40% full validation, which is an unsatisfactory result. The following method is observed in columns six and seven, where the dispersion method of 4 Hz modulation energy is described. For it, we get 100% validation accuracy, which is the highest possible result we are looking for. By comparing the three methods, we can categorically define the 4Hz modulation energy method as the best method for determining speech signals, given its complexity compared to the other two methods, the result is expected. But the simplest method - the frequency of zero crossings is more effective for determining speech signals than the more sophisticated method - the short-term energy method.

The first wrong classification we notice for the ZCR method in the first female voice record. Given the simplicity of the method, it is not the most reliable classifier since the values of the source data do not differ dramatically from the wrong classification and may be due to the low energy in this record and the presence of more consonant letters or silence mixed with some noise. Let's look at the short-term energy method, and we see that all but one values are very close. We see that wrong classification is in both female and male speech records. Their values are so close that we can not identify features that are clear and suggest an increase or decrease in energy to a subsequent misclassification. These close values may be due to the peripheral devices used, the non-isolated environment, etc. So the accuracy of the short-term energy method is unsatisfactory for end-user applications.

The most effective and accurate method that end-user can use is the dispersion of 4 Hz modulation energy. Depending

on the priority, if it is the accuracy and not time consumption, the best and most appropriate is this method.

The classification accuracy for the music signals is presented in Table IV.

TABLE IV
PAGE LAYOUT DESCRIPTION

Measures	ZCR		FPOW		4Hz	
Threshold	0.0928		0.0032		28.4830	
Classify	Right	Wrong	Right	Wrong	Right	Wrong
Rock	1	0	0	1	1	0
Jazz	1	0	0	1	1	0
Latino	1	0	1	0	1	0
Folk	0	1	1	0	1	0
Hindi	0	1	1	0	1	0
Nordic	1	0	0	1	1	0
African	0	1	0	1	0	1
Chinese	0	1	1	0	1	0
Russian	0	1	1	0	1	0
Classic	0	1	0	1	1	0
Accuracy,%	40	60	50	50	90	10

After the necessary calculations for all the music signals for the zero crossing method, we get the accuracy of the full validation of 40%. The accuracy of this method of recognizing musical signals is not satisfactory. Some of the errors that are introduced in classifying using this method may be due to the type of music used. We observe a proper classification for rock, jazz, etc., while in classical, Russian, Chinese music, etc., we observe a wrong classification, which may be due to the instruments and pauses contained in a certain type of music. For example, classical and Chinese music is experiencing energy accumulation at low and high frequencies, as are the moments of silence caused by instruments used in this type of music such as string, wind, keyboard instruments - flute, violin, piano.

The next method that we are looking at is the short-term energy method for it after the classification of all the music files we get the accuracy of the full validation 50%. The accuracy of this method for recognizing a musical signal is better than the accuracy of the zero crossing frequency method and its accuracy of speech recognition but is still unsatisfactory and the use of this method is not very reliable.

After the classification of all music files with 4Hz method, modulation energy has a 90% accuracy of full validation. The only error that has been made is with African music, yet this method is the closest to the maximum accuracy we are looking for. By comparing the three methods, we can conclude that the method of 4 Hz modulation energy dispersion is the most reliable method for recognizing musical signals. But this time for the classification of musical signals, the short-term energy method is more applicable than the zero crossing frequency method. Observing Table IV we can see that the wrong classification, which is present in the 4Hz method, is not correctly classified in the other two methods.

In Fig. 1 the overall performance is given for the three algorithms of classifying the test audio content into speech and music with the complete variability in different styles.

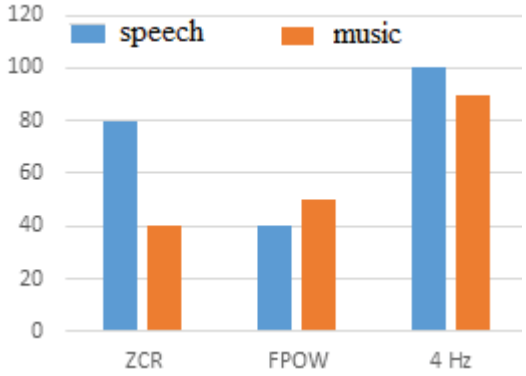

Fig. 1. Overall accuracy of classification for the three tested algorithms

IV. CONCLUSION

The best and most accurate method that most satisfies the discriminatory condition and can be used by an end user in an objective environment with publicly available and relatively cheap peripherals for speech and music discrimination is the 4Hz method. The following method, which certifies the discrimination condition to some extent and can be used is the method of the zero crossing rate, although it is not very reliable, but the calculation is much simpler, which makes it an ideal choice if the main goal is speeding-up the process rather than achieving accuracy close to 100%. For future improvement of the zero crossing method and the short-term energy method, professional peripheral devices and soundproofing environment can be used to increase their classification accuracy. End user use can use those along with other methods or features to improve their performance.

REFERENCES

[1] Carey, M., E. Parris, S. Eluned, S., H. Lloyd-Thomas. A comparison of features for speech, music discrimination. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 149-152, 1999.

[2] Lu, L., H. Jiang, H. Zhang, A robust audio classification and segmentation method. In Proceedings of the 9th ACM International Conference on Multimedia, pp. 203-211, October 2001.

[3] Velayatipour, M., B. Mosleh, A review on speech-music discrimination methods, International Journal of Computer Science & Network Solutions, February 2.2, 2014.

[4] Ericsson, L., Automatic speech/music discrimination in audio files. Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan, 2010.

[5] Spina.S. Analysis and transcription of general audio data, PhD Thesis. Massachusetts Institute of Technology, 2000.

[6] Shete, D., S. Patil, S. B. Patil, Zero crossing rate and Energy of the Speech Signal of Devanagari Script. IOSR-JVSP, 4.1: 1-5, 2014.

[7] Logan, B. Mel Frequency Cepstral Coefficients for Music Modeling. In: ISMIR, p. 1-11, 2000.