

Speech/Music/Silence and Gender Detection Algorithm

Hadi HARB , Liming CHEN, Jean-yves AULOGE

Lab. ICTT Dept. Mathematiques - Informatique, ECOLE CENTRALE DE LYON. 36,
avenue Guy de Collongue B.P. 163, 69131 ECULLY Cedex, France

{Hadi.harb, Liming.chen, Jean-yves.Auloge}@ec-lyon.fr

Abstract

A speech – music – silence discrimination and a gender detection algorithm is presented in this paper. First silence segments are extracted from audio stream by using energy and ZCR features. The speech are detected using the energy envelope and harmonic features. Music segments are then classified using energy envelope, and harmonic features too. For gender detection, we propose a feature that we used to discriminate between men and women's voices. The proposed algorithm needs no training phase, as in Gaussian Mixture Models based algorithms, and it classifies audio stream into 4 classes, speech, music, silence, and else with a delay of 1 s. Once speech is extracted, gender detection could be applied, and the detection could be in real time. As an evaluation of the proposed algorithm, we applied it on 10 min of audio extracted from CNN programs, 93% of classification accuracy for speech and 84% of classification accuracy for music is achieved. 80% of gender detection accuracy in speech segments is achieved.

1. Introduction

With more and more digital video data added and archived every day, and with the digitalization of archives, one could expect the massive video data existing today. Searching segments of interest within video programs, as broadcast news, TV programs, scenes in film, is very hard. On the other hand, manual indexing of video programs is very costly and slow. Clearly we need powerful automatic methods for content indexing.

As a video program is usually composed of a visual channel and one or several audio channels, an automatic video segmentation process should rely on the visual and audio channel analysis. Till now,

most video segmentation techniques are based on visual analysis [1], we investigate in this paper the audio part of a video for indexing purpose.

A very basic segmentation within audio channel is speech/music/silence/noise discrimination which helps improving scene segmentation when combined with visual based segmentation techniques.

Recently, works show that combining audio and visual analysis for video segmentation improves performance [2], [3].

Other motivations for the audio channel segmentation - classification include :

- 1- giving more reliable data (speech only) to the ASR (Automatic Speech Recognizer) [4], which minimizes word error rates, out of vocabulary and computations on non-speech data. Also when we have reliable text spoken in a video (ASR output) we could extract knowledge from this text and combine it with knowledge extracted from image processing to refine indexing.
- 2- Giving meaningful descriptors such as music, silence, speech, ... as in the MPEG7 specifications [5].
- 3- By extracting speech we could apply speaker recognition techniques for identifying and tracking specific speakers in a video. Which is a very important descriptor for content indexing.
- 4- Improving audio coding, by decreasing the bit-rate for silence segments, already extracted reliably.

Humans classify and segment audio every day without a considerable effort. In fact, recognizing audio classes and discriminating between music speech, and silence, is one of many other pattern recognition cases processed all the time by humans. Also, solving audio recognition problem, as for all

pattern recognition problems, is done in an efficient and reliable manner using our brain and nervous system. An efficiency that the most powerful computers, using current state-of-the-art algorithms for pattern recognition are away from it.

What makes the problem of audio recognition as music/speech discrimination a hard task, is that finding features or mathematical models that describe all the variability of classes efficiently is not evident.

Features used for audio recognition are low level features, such as MFCC or cepstrum or DFT features [6]. These features perform well for speech recognition where small variations are important for phoneme recognition. But when classes are music, or speech or noise, where the definition of classes is more general and greater variability within classes exists, we need some kind of high level features that could describe a class in general way.

Approaches proposed in the literature for audio segmentation-classification can be summarized as follows :

- 1- Model-based approaches: where models for each audio class is created, such as Gaussian Mixture Models [7], [8]. These models are based on low level features in general, such as cepstrum, MFCC. Then after a training phase, audio could be classified as one of these classes. With advantage of classifying and segmenting at the same time. The problems are data and time needed for training.
- 2- Metric-based segmentation: which segments using distances (like KullBack-Leibler distance) between neighboring windows [9] (different acoustic features could be used to create the acoustic vectors, Auto-regressive Gaussian model parameters [10], DFT parameters [11]...) . The advantage of this approach is that there is no need for prior knowledge on audio classes, and it could be applied for real time segmentation . But it gives no information about audio segments.
- 3- Rule based approaches: in these approaches rules are created describing each class, [12]. These rules are based on high and low level features.
- 4- Decoder based approaches: in these approaches the Hidden Markov Model HMM of the speech recognition system is used. HMM are trained to give the class of the audio signal [13].

In this paper we propose a set of features or combination of features, that aims to describe music, speech and silence classes in a general manner. Using these features we propose an algorithm that discriminates between our audio

classes in an IF-THEN fashion. Then we propose a parameter that could discriminate between men and women's voices.

The rest of this paper is organized as follows, in section 2, we describe Music, Speech and Silence properties and the proposed features. In section 3 we present the gender detection. Then the algorithm is presented in section 4. And the results in section 5. Finally a conclusion and future directions are presented in section 6.

2 Speech, Music, Silence Properties

Before segmentation and classification, a study of the properties of audio classes is essential.

2.1 Silence

Silence is defined as a non perceptual audio signal, for the human ear. Normally the energy level of silence is relatively low. So an energy thresholding could extract silence segments. But there exists other types of audio segments that could be classified as silence if energy level is used alone, such as a low energy music or speech. Fortunately the Zero Crossing Rate ZCR for silence is generally less than ZCR, for other types of audio. By intuition combining these 2 features, energy and ZCR could improve the accuracy of silence detection process.

Let $S_w = E_w \bullet ZCR_w$ where E_w ¹ and ZCR_w are respectively the normalized energy, and the Zero Crossing Rate of the window "W".

Thresholding S_w could extract silence segments.

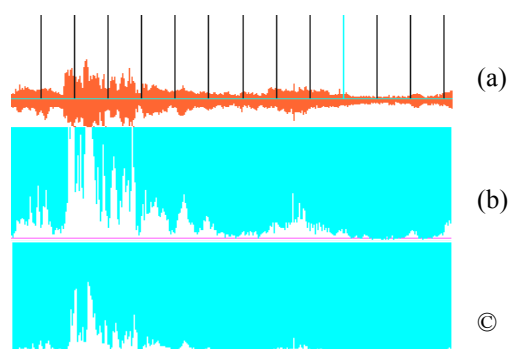


Fig2.1, background noise – music (a), and detection using our variable "S" (b), and by simply thresholding the Energy (c)

¹ See Section 4 for the extraction of the ZCR and the Energy

In Fig2.1, we show a signal containing background noise, and music. By using the energy feature only, we risk to detect a large part of non-silence as silence. But when using our variable “S” based on the energy and the ZCR, we minimize this risk.

2.2 Speech

Speech is defined as a series of words spoken in a continuous fashion, other types of speech, such as a shouting person, is considered as non speech in this paper. One could expect higher energy levels when words are spoken, and lower energy for inter words duration.

Two things characterize speech.

First an alternation in the energy between peaks, corresponding of words, and almost zero energy, corresponding of inter words silence. This characteristic is helpful for detecting speech signals. We proposed to count in windows of 1s the number of times the energy falls below the silence level. We call this feature, the Silence Crossing Rate SCR.

Our experiments show that the SCR is between 5 and 10 for speech. For other types of audio, such as music, it is higher or lower.

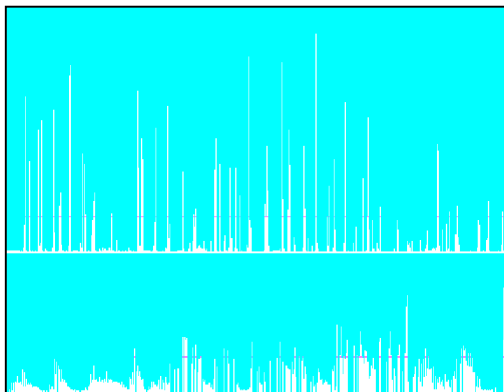


Fig.2.2, the shape of energy envelope for speech (top), and music (bottom), we see the distribution of peaks in the speech signal.

In Fig.2.2 we see the existence of peaks distributed in continuous manner for speech, which is not the case for music.

An other important characteristic for speech, is that Frequency Tracking gives dispersed and short segments.

The Frequency Tracking is as follows:

In the spectrum of the signal. We try to track the five more important frequencies in the DFT (Discrete Fourier Transform coefficients) vectors. Tracking means to see if these frequencies continue to be the most important frequencies within 100 Hz band in the next DFT vector. When a frequency could not be tracked, we signal a cut. The number of cuts in a window of 1s is the Frequency Tracking FT feature.

The FT for speech signals gives short dispersed segments and it looks like this :

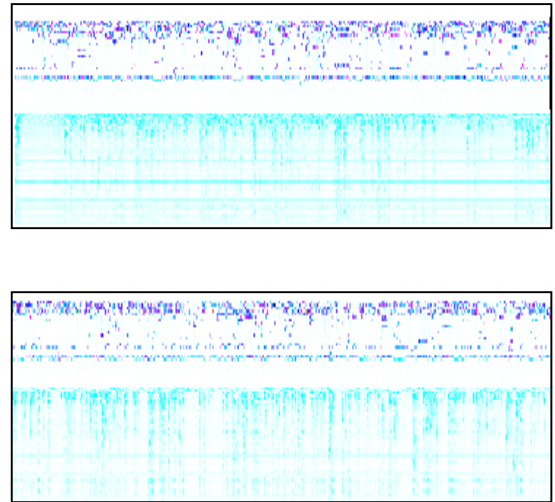


Fig 2.3, spectrum and Frequency Tracking for speech signals

Our experiments show that FT for speech signals is clearly higher than for music ones. For music, the changes in the spectrum are smooth in general.

For music the FT gives long, parallel segments. It looks like:

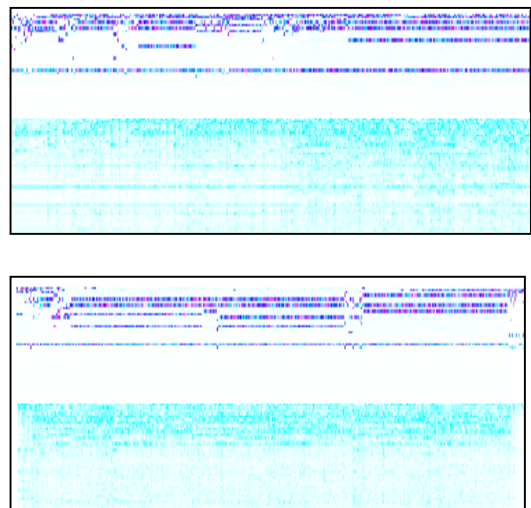


Fig 2.4, Spectrum and Frequency Tracking for music signals

Now a feature based on the combination of the FT and SCR could discriminate between music and speech. Let

$$P_w = g(SCR_w) + f(FT_w) \text{ where } SCR_w$$

and FT_w are the Silence Crossing Rate and the Frequency Tracking.

Functions $f(.)$, $g(.)$ are functions that maps from real numbers to the $[0,1]$ interval. These functions could be linear ones, and they could be defined after experiments to maximize P_w for speech segments.

The same could be done for music, by defining a feature M_w

$$M_w = g1(SCR_w) + f1(FT_w)$$

now we choose functions that maximizes M_w for music.

A basic definition of functions $g(.)$ and $f(.)$ could be as follows:

$$\begin{aligned} g(x) &= 1 && \text{if } 4 < x < 11 \\ &= 0 && \text{otherwise} \\ f(x) &= 1 && \text{if } x > 75 \\ &= 0 && \text{otherwise} \end{aligned}$$

and

$$\begin{aligned} g1(x) &= 1 - g(x) \\ f1(x) &= 1 - f(x) \end{aligned}$$

finally if $P_w=2$ this implies a speech segment
if $M_w = 2$ this implies a music segment
otherwise this is an "else" segment.

In fact, a study on the distribution of the SCR, and the FT for speech and music, to see if they have Gaussian distribution is very important.

3 Gender detection

In videos, a meaningful descriptor that could help improving scenes detection using the visual stream, is the gender of the speaker. In this paper we are interested in the discrimination between man and woman in speech segments already extracted.

Humans discriminate between men and women according to the frequency. Women speak with higher fundamental frequencies than men, and the ZCR for a woman's voice is higher than that for man's voice.

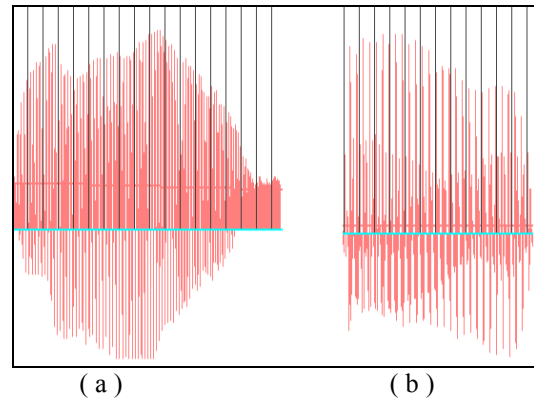


Fig 3.1, the same word "close" spoken by a woman (a) and a man (b)

Fig 3.1 shows the higher ZCR (or the number of times the signal crosses the 0 axis) for woman's voice.

Another important thing in man's and woman's voices, is that the center of gravity of the spectrum is, for man's voices close to low frequencies; and to higher frequencies for women's ones. The center of gravity of an acoustic vector could be calculated as follows

Let X the acoustic vector containing frequency coefficients. And G the center of gravity of this vector.

$$G = \frac{\sum_f X_f \cdot f}{\sum_f f} \quad \text{where } X_f \text{ is the coefficient}$$

corresponding of the frequency "f" in the vector X , and "f" is the frequency index.

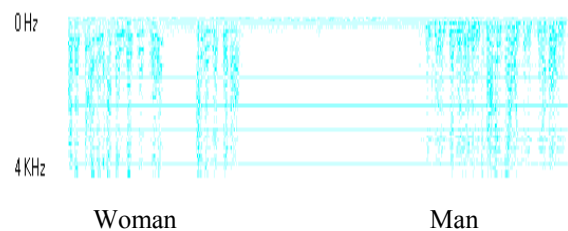


Fig3.2, frequency distribution for a woman and man's voice

Fig 3.2 shows the frequency distribution for women's and men's voices. We see clearly that the frequency distribution for men's voices is closer to low frequencies than those of women's voices. The proposed variable to discriminate between man and woman is the following :

$$W = \frac{\sum_{f=1}^5 X_f}{\sum_{f=35}^{40} X_f} \cdot \frac{1}{\text{Mean}(ZCR)} \cdot G$$

Mean(ZCR) is the mean of ZCR in 1s

W could be calculated for every acoustic vector in a speech segment, and refined for every 1 s by Mean(ZCR). This variable should be higher for men's voices.

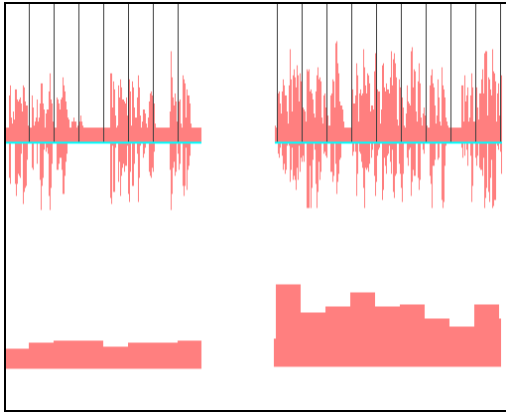


Fig3.3 the variable W generated for woman's voice "left", and man's voice "right"

4 Discrimination algorithm

- 1- For every 10 ms an acoustic vector containing DFT coefficients for frequencies 100 Hz to 4KHz is extracted, the energy is then calculated by the following formula:

$$E_t = \sum_{f=1}^{40} X_f^2(t) \cdot \frac{1}{40} \quad \text{"t" means that the}$$

energy is calculated for every 10 ms (acoustic vector).

- 2- For every 10 ms calculate ZCR. ZCR is calculated by the following formula :

$$Z_n = \frac{1}{2} \sum_m [sgn[x(m)] - sgn[x(m-1)]]w(n-m)$$

with

$$sgn[x(n)] = \begin{cases} 1 & \text{if } x(n) \geq 0 \\ -1 & \text{if } x(n) < 0 \end{cases}$$

w(n) is a rectangular window of length n, 10ms in our case.

- 3- Extract silence segments using the feature "S"
- 4- Discard silence segments in further studies.
- 5- Extract speech using the feature "P"

- 6- Discriminate between men's and women's voices in speech segments, using the feature "W".

- 7- Discard speech segments in further studies.

- 8- Extract music segments using feature "M"

- 9- Label unclassified segments as "else"

The segmentation of audio stream is done after this classification. Segment boundaries are placed where their is a class change.

5 Results

To evaluate the proposed algorithm we used it for the classification of 10 min of audio data collected from CNN programs, and 10 min from film "un indien dans la ville". In these streams music segments, speech (spoken by men and women), silence and noise exist.

Results for speech detection accuracy are listed in table 5.1

Insertion rate is the percentage of non-speech segments classified as speech. Deletion rate is the percentage of speech segments classified as non-speech.

Program	Deletion rate %	Insertion rate %	Man/Woman discrimination error %
10 min Film	4	10	23
10 min CNN	8	4	17

Table 5.1 deletion, insertion and gender detection error rates, for speech

And the accuracy for music classification on 10 minutes extracted from different songs, and 10 min from instrumental music.

Instrumental	Songs
88%	80%

Table 5.2 classification accuracy for music.

6 Conclusion

An algorithm for speech/music/silence and gender detection algorithm is presented in this paper.

First a study on the audio classes we used, speech, music and silence are done. We proposed a set of features that could reflect the properties of the defined audio classes. By thresholding these

features a classification – segmentation could be done. We showed that our method for silence detection is robust for noise, or background music as in [12]. Then the problem of gender detection is presented, and a feature for this task is proposed.

We showed that the proposed features could discriminate between our audio classes, in experiments we did.

With no prior training phase, as for GMM based algorithms, and by simply thresholding the proposed features 90% of classification accuracy is achieved.

The problem with this algorithm stills the use of thresholds. As future directions, modeling these features by GMM for example will be done, to eliminate the use of thresholds, and to increase the robustness. Speaker detection, and merging audio information with visual studies already done [1] to improve scene detection in video programs, will be investigated.

7 References

[1] W. Mahdi, M. Ardebilian, L. Chen, *Automatic Scene Segmentation Based on Spatial-Temporal Clues and Rhythm*, to appear in International Journal of Networking and Information Systems, Vol. 5 September 2001.

[2] D. Pye, [N. Hollinghurst](#), [T. Mills](#), [K. Wood](#), *Audio Visual segmentation for content based retrieval*, the international conference on spoken language processing (ICSLP 98), Sydney, Australia, December 98

[3] John M. Gauch, Susan Gauch, Sylvain Bouix, Xiaolan Zhu, *Real time video scene detection and classification*, Information Processing and Management 35 (1999), pp 401-420

[4] Michelle S. Spina, *Analysis And Transcription of General Audio Data*, Phd thesis, MIT, june 2000

[5] ISO/IEC draft MPEG7 Audio specifications, <http://www.darmstadt.gmd.de/mobile/MPEG7/Documents.html>, N3802.

[6] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, *Strategies for automatic segmentation of audio data*, Proc. icassp2000 ISTANBUL pp 1423-6, 2000

[7] Gethin Williams, Daniel Ellis, *Speech/music discrimination based on posterior probability features*, Eurospeech 1999

[8] P. C. Woodland, T. Hain, S. Johnson, T. Neisler, A. Tuerk, S. Young, *Experiments in Broadcast news transcription*, Proc. ICASSP 1998, Seattle, May 1998

[9] M. Seigler, U. Jain, B. Raj, R. Stern, *Automatic segmentation, classification, and clustering of Broadcast news audio*, Proc. Of the DARPA speech recognition workshop, February 1997.

[10] R. André-Obrecht , *A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals*, IEEE on Trans. on ASSP, vol. 36 , n 1, January 1988.

[11] H. Harb, Husseiny, Said, *Isolated Words Recognition Using Neural Networks*, Proc. Of the 7th IEEE ICECS 2k, pp 349-352, Kaslik, Lebanon 2000.

[12] Francis Kubala, Hubert Jin, Spyros Matsoukas, Long Nguyen, Rich Schwartz, John Makhoul *The 1996 Bbn Byblos Hub-4 Transcription System*, 1996

[13] Tong Zhang and C.-C. Jay Kuo, *Heuristic approach for generic audio data segmentation and annotation*, ACM Multimedia Conference, pp. 67-76, Orlando, Nov. 1999