FUSION OF DESCRIPTORS FOR SPEECH / MUSIC CLASSIFICATION

Julie Mauclair[1] *and Julien Pinquier*[2]

[1]Laboratoire d'Informatique de l'Université du Maine - CNRS FRE 2730
Avenue René Laennec, 72085 Le Mans cedex 09, FRANCE
phone: ++33 (0)2 43 83 38 43 - fax: +33 (0)2 43 83 38 68
email: mauclair@lium.univ-lemans.fr - http://www.lium.univ-lemans.fr

[2]Institut de Recherche en Informatique de Toulouse - CNRS UMR 5505
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
phone: ++33 (0)5 61 55 88 35 - fax: ++33 (0)5 61 55 62 58
email: pinquier@irit.fr - http://www.irit.fr/recherches/SAMOVA/

ABSTRACT

This work addresses the soundtrack indexing of multimedia documents. We present a speech/music classification system based on three original features: entropy modulation, stationary segment duration and number of segments. They were merged by basic score maximisation with the classical 4 Hertz modulation energy. We validate this fusion approach with the use of the probability theory and the evidence theory. The system is tested on radio corpora. Systems are simple, robust and could be improved on every corpus without training or adaptation.

1. INTRODUCTION

Methods of indexing in audio (and video) are mainly manual: a human operator must read, listen to and/or look at the numerical document in order to select required information. This task of indexing must be automated because the volume of data increases enormously and the treatment of several requests becomes extremely tiresome. To index an audio document, keywords or melodies are semi-automatically extracted, speakers are detected... Nevertheless all these detection systems presuppose the extraction of elementary and homogeneous acoustic components.

Several methods of speech/music discrimination were described in the literature. We observe two tendencies. On one hand, the musician community has given more importance to features which increase a binary discrimination [11]. For example, the zero crossing rate and the spectral centroid are used to separate voiced speech from noisy sounds [12], [14] whereas the variation of the spectrum magnitude (the spectral "Flux") attempts to detect harmonic continuity [13]. On the other hand the automatic speech processing community has focused on cepstral features [6]. Three concurrent classification frameworks are usually investigated: Gaussian Mixture Models, k-nearest-neighbors [4] and Hidden Markov Models.

In a previous paper, we have studied a basic fusion (by score maximisation) of these methods [10]. In this paper, we decide to use the probability theory and the evidence theory in order to validate our approach of fusion.

This paper is divided into three parts: a presentation of our classification system, a description of our fusion methods and test experiments performed on radio documents.

2. CLASSIFICATION SYSTEM

Two detection subsystems have been previously studied: one for speech detection and another for music detection (Figure 1).

- For the speech detection subsystem, we have used entropy modulation and 4 Hz modulation energy.
- For the music detection subsystem, an automatic segmentation has given the number of segments by time unit and segment duration average.

In fact, we have two classifications for each second of input signal: the speech/non-speech one and the music/non-music one. Of course, at the end, we can merge them.

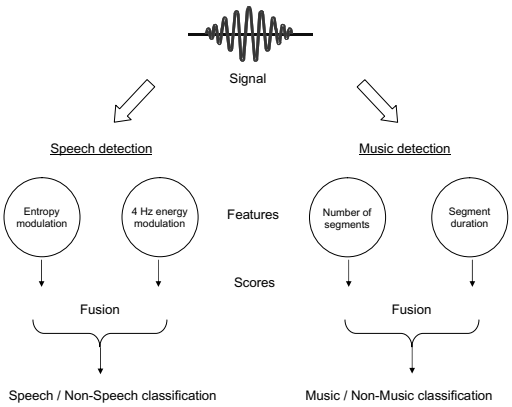

Figure 1: Speech/Music classification system.

2.1 Speech detection subsystem

- 4 Hz modulation energy

Speech signal has a characteristic energy modulation peak around the 4 Hz syllabic rate [7].In order to model this property, the classical procedure is applied: the signal is segmented in 16 ms frames. Mel Frequency Spectrum

Coefficients are extracted and energy is computed in 40 perceptual channels. This energy is then filtered with a FIR band pass filter, centered on 4 Hz. Energy is summed for all channels, and normalized by the mean energy on the frame. The modulation is obtained by computing the variance of filtered energy in dB on one second of signal. Speech carries more modulation energy than music.

• Entropy modulation

Music appears to be more "ordered" than speech considering observations of both signals and spectrograms. To measure this "disorder", we evaluate a feature based on signal entropy:

$$H = \sum_{i=1}^{k} -p_i log_2 p_i$$

with p_i=probability of event i

The signal is segmented in 16 ms frames, the entropy is computed on every frame. This measure is used to compute the entropy modulation on one second of signal. Entropy modulation is higher for speech than for music.

2.2 Music detection subsystem

The segmentation is provided by the "Forward-Backward Divergence algorithm"[3] which is based on a statistical study of the acoustic signal. Assuming that the speech signal is described by a string of quasi stationary units, each one is characterized by an auto regressive Gaussian model. The method consists in performing a detection of changes in the auto regressive parameters. The use of an *a priori* segmentation partially removes redundancy for long sounds, and a segment analysis is relevant to locate coarse features. This approach have given interesting results in automatic speech recognition: experiments have shown that segmental duration carry pertinent information [2].

Figure 2a - Segmentation on about 1 second of speech.

Figure 2b - Segmentation on about 1 second of music.

• Number of segments

The duration feature is the consequence of the application of the segmentation algorithm described above. The speech signal is composed of alternate periods of transient and steady parts (steady parts are mainly vowels). Meanwhile, music is more constant, that is to say the number of changes (segments) will be greater for speech (Figure 2a) than for music (Figure 2b). To estimate this feature, we compute the number of segments on one second of signal.

• Segment duration

Segments are generally longer for music (Figure 2b) than for speech (Figure 2a). We use the segment duration as feature. We decide to model it by a Gaussian Inverse law. The pdf is given by:

$$p(g) = \sqrt{\frac{\lambda}{2\pi g^3}} * e^{\frac{-\lambda(g-\mu)^2}{2\mu^2 g}}, g \geq 0$$

with μ = mean value of g and $\frac{\mu^3}{\lambda}$ variance of g.

2.3 Reference system

For each subsystem, a statistical model is estimated. We make the decision regarding to the maximum likelihood criterion and we called it "reference system". Then, we propose to validate this approach by using other fusion methods like probability theory and evidence theory.

3. DATA FUSION

In the scientist community, data fusion becomes a very attractive domain [5]. In fact, sources are not always reliable and we need to compensate ones' weakness with others. Probability theory and evidence theory provide some solutions to address the problem of merging information coming from several sources.

3.1 Probability theory

Fusion methods were first considered in the Bayesian approach. Effectively, in probability theory, the Bayesian theorem is used to estimate the probability of a future event occurrence given the probability events in the past. The decision is made with the a posteriori maximum criterion. This theory has the disadvantage to require a perfect knowledge about the event probability, particularly for *a priori* probability.

To avoid this problem of ignorance, we can rely on measures of confidence, based on two information: expert and class [9]. Let α be the expert confidence measure and β be the class confidence measure. The expert manages to discriminate Class C from Non-Class NC with a rate α_e which is the one's complement from the error rate.

Here, experts are the four parameters of the classification system (entropy modulation, 4hz modulation energy, number of segments and segment duration). From each expert, we can extract a confusion matrix Class/Non-Class (Speech/Non-Speech or Music/Non-Music). β becomes:

$$\beta_e C = \frac{P(y = C|C)}{P(y = C|C) + P(y = NC|C)}$$

$$\beta_e NC = \frac{P(y = NC|NC)}{P(y = C|NC) + P(y = NC|NC)}$$

where y is the observation extracted every second.

Bayesian strategy gives us a decision function for each expert:

$$s_e^*(y) = min\left\{\left\{(1-\beta_{NC}) * \frac{Pr(y|C)}{P(y)}\right\}, \left\{(1-\beta_C) * \frac{Pr(y|NC)}{P(y)}\right\}\right\}.$$

Finally, we produce the decision with the expert e maximizing:

$$\alpha_e * (1 - s_e^*(y)).$$

3.2 Evidence theory

As we have seen, solutions in combining multiple classifiers are numerous but each of them has weaknesses. Most treat imprecision, but uncertainty and reliability are ignored. Evidence theory allows to use uncertain data [8] and [1].

Let θ be a set of N classes: $\theta = \{C_1, ..., C_N\}$.
Here, $\theta = \{S, M, SM, N\}$ with S for speech, M for music and N for noise.

From this confusion set, we can define another set:

$$2^\theta = \{A | A \subseteq \theta\}$$
$$2^\theta = \{\emptyset, \{S\}, \{M\}, \{SM\}, \{N\}, \{S \cup M\}, ..., \theta\}.$$

This set is used as a referential to evaluate the truth of a proposition. Supposing an information coming from an expert (or any other source), it expresses an opinion over elements in 2^θ that is to say over single hypothesis or disjunctions of those. Opinions over the system can be illustrated with belief degrees on hypothesis. These degrees are described with a belief function $m_\theta : 2^\theta \rightarrow [0,1]$ with:

1. $m_\theta(\emptyset) = 0$
2. $\sum_{A \subseteq \theta} m_\theta(A) = 1$

Such a description permits to distribute our knowledge on the 2^θ set and $m_\theta(A)$ is the part of the belief degree on proposition A. Every expert provides its own function which is combined with others using the Dempster-Shafer's rule [8]. The result gives a final mass distribution to access reliable information.

We have four experts (parameters of the classification system) providing four belief functions (m_1,...,m_4). To obtain $m_e(\theta)$, which represents ignorance mass, we have to consider expert's mistake. For other hypothesis, every expert gives an opinion resulting from *a priori* probabilities.

For example, expert 1 (4 Hz modulation energy):

$m_1(y \in \{S \cup SM\}) = m_1(S \cup SM) = P(y|Speech)$
and $m_1(M \cup N) = P(y|Non - Speech)$.

So, we have four mass distributions to combine with Dempster-Shafer's rule. We make a decision with the maximum of plausibility, which cares about all hypothesis disjunctions' weights:

$$Pl_\theta(A) = \sum_{B \cap A \neq \emptyset} m_\theta(B)$$

Thus, belief theory makes it possible to manage combined hypothesis and to reduce *a priori*'s part in the description's problem.

4. EXPERIMENTS

4.1 Corpus

Our corpus corresponds to a database made of records from RFI (Radio France Internationale, CNRS' RAIVES project). It is composed of songs, reports, sports, adverts, news, interviews... sampled at 16 kHz. It contains long periods of speech, music, 'mixed' zones and noise. Speech is recorded in different conditions (phone call, outdoor), with different speakers and many languages. We have also different kind of music like pop songs, opera, orchestral music...

4.2 Results

The classification system (cf. 2) gives results by decomposing Speech/Non-Speech from Music/Non-Music. To evaluate our model, we separate experts 1 and 2 (Speech/Non-Speech, Table 1) from experts 3 and 4 (Music/Non-Music, Table 2) and we compare the results with those from the classification system.

Table 1: *Speech/Non-Speech classification*

Speech/Non-Speech sub-system	Accuracy
4 Hz Modulation energy	87.3 %
Entropy modulation	87.5 %
Reference system (max)	90.5 %
Probability theory	**90.7 %**
Evidence theory	**90.9 %**

Table 2: *Music/Non-Music classification*

Music/Non-Music sub-system	Accuracy
Number of segments	86.4 %
Segments duration	78.1 %
Reference system (max)	89 %
Probability theory	**84.8 %**
Evidence theory	**86.9 %**

The reference system is our previous work [10] where we have used a basic fusion (by score maximisation).

Theory of probabilities offers a model which gives similar results to the reference system. By the way, this theory is at the base of this system so this model confirms previous results. Confidence measures provide improvement in Speech/Non-Speech discrimination.

We explain the score in Music/Non-Music with the expert confidence rate of "segment duration" which is not very high ($\alpha_4 = 53\%$). Thus, this parameter is not relevant in final fusion, only the parameter "number of segments" matters.

For evidence theory, using mass distributions improve results. We explain the score in Music/Non-Music like previous theory.

5. DISCUSSION

We describe in this paper four features based on different properties of the signal. All those features considered separately are relevant in a speech/music classification task, and the correct classification rates vary from 76 to 84 %. Then, we propose fusion theories that permit to raise the correct classification rate up to 90 %.

Probability theory allows using our *a priori* knowledge about the system to merge information with confidence measures. This theory produces satisfying results to discriminate Speech from Non-Speech. So, it validates this subsystem's

model. For Music/Non-Music classification, we obtain lower scores than with the parameter "number of segment" alone. That shows how "segment duration" is not the best support to "number of segment" in this discrimination.

Evidence theory gives the best results in Speech/Non-Speech classification. For Music/Non-Music's ones, we can improve them by transforming our distribution masses' calculation for the two segmentation features. Thus, this theory appears to be the best solution to formalize the classification system.

Those two theories fit with a shape recognition problem in sound indexation. The problem's modelization can be refined by using other techniques for combining and calculating confidence measures and distribution masses. Those fusion's techniques may be relevant for other domains like language identification where we can, for example, merge models describing different sources: acoustic, phonotactic and prosodic information. Many errors in automatic classification are due to a bad discrimination in Speech/Singing Voice. We hope those theories will give results in this sound indexation domain.

REFERENCES

[1] H. Altinçay and M. Demirekler. Speaker identification by combining multiple classifiers using dempster-shafer theory of evidence. *Speech Communication*, 2003.

[2] R. André-Obrecht and B. Jacob. Direct identification vs. correlated models to process acoustic and articulatory informations in automatic speech recognition. In *International Conference on Audio, Speech and Signal Processing*, pages 989–992, Munich, Germany, 1997. IEEE.

[3] R. André-Obrecht. A new statistical approach for automatic speech segmentation. *IEEE Transactions on Audio, Speech, and Signal Processing*, 36(1), January 1988.

[4] M. J. Carey, E. J. Parris, and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. In *International Conference on Audio, Speech and Signal Processing*, pages 149–152, Phoenix, USA, March 1999. IEEE.

[5] D. Dubois and H. Prade. La fusion d'informations imprécises. *Traitement du signal*, 11(6), 1994.

[6] J. L. Gauvain, L. Lamel, and G. Adda. Systèmes de processus légers : concepts et exemples. In *International Workshop on Content-Based Multimedia Indexing*, pages 67–73, Toulouse, France, October 1999. GDR-PRC ISIS.

[7] T. Houtgast and J. M. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, 77(3):1069–1077, 1985.

[8] F. Janez. *Fusion d'informations définies sur des référentiels non-exhaustifs différents*. PhD thesis, Université d'Angers, 1996.

[9] P. Leray, H. Zaragoza, and F. d'Alché Buc. Pertinence des mesures de confiance en classification. In *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, Paris, December 2000.

[10] J. Pinquier, Jean-Luc Rouas, and R. André-Obrecht. A fusion study in speech / music classification. In *International Conference on Audio, Speech and Signal Processing*, Hong-Kong, China, April 2003.

[11] S. Rossignol, X. Rodet, J. Soumagne, J. L. Collette, and P. Depalle. Automatic characterization of musical signals: feature extraction and temporal segmentation. *Journal of New Music Research*, 28(4):281–295, December 1999.

[12] J. Saunders. Real-time discrimination of broadcast speech/music. In *International Conference on Audio, Speech and Signal Processing*, pages 993–996, Atlanta, USA, May 1996. IEEE.

[13] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *International Conference on Audio, Speech and Signal Processing*, pages 1331–1334, Munich, Germany, April 1997. IEEE.

[14] T. Zhang, C. Kuo, and C. J. Hierarchical system for content-based audio classification and retrieval. In *Conference on Multimedia storage and Archiving Systems III*, volume 3527, pages 398–409. SPIE, November 1998.