Projet ANR- 11-IS02-001

MEX-CULTURE/ Multimedia libraries indexing for the preservation and dissemination of the Mexican Culture

Deliverable

Final report on Speech/Audio descriptors

Programme Blanc International II- 2011 Edition

A IDENTIFICATION

Project acronym	MEX-CULTURE
Project title	Multimedia libraries indexing for the preservation and dissemination of the Mexican Culture
Coordinator of the French part of the project (company/organization)	Centre d'Etude et de Recherche en Informatique et Communications – Conservatoire National des Arts et Métiers
Coordinator of the Mexican part of the project (company/organization)	Centro de Investigación y Desarrollo de Tecnología Digital – Instituto Politécnico Nacional
Project coordinator (if applicable)	Michel Crucianu : France Mireya Saraí García-Vázquez : México
Project start date Project end date	01/01/2012* 30/04/2016
Competitiveness cluster labels and contacts (cluster, name and e-mail of contact)	Cap Digital Paris-Région Philippe Roy Philippe.Roy@capdigital.com
Project website if applicable	http://mexculture.cnam.fr/

** The Mexican partners are only financed since November 2012.*

Coordinator of this report	
Title, first name, surname	*Prof. Mireya S. Garcia-Vazquez*
Telephone	*(+52) 664 3 47 21 00*
E-mail	*freemgarciav@gmail.com*
Date of writing	*08/05/2016*

Redactors :	IPN, LaBRI Mireya Saraí García Vázquez (CITEDI-IPN) Alejandro Ramírez Acosta (CITEDI-IPN)

B INTRODUCTION

The rise of the Internet and the world-wide-web, starting in the late 1990s, and the invention of the MPEG-1 video and MP3 audio encoding gave an enormous boost to computational processing of video and audio within the research area of Mutimedia Information Retrieval. The integration of digital technology contributes to the preservation of sound heritage and facilitates access and dissemination to a large number of people simultaneously.

Multimedia Information Retrieval (MIR) is a multidisciplinary research [B.1.1]. Two main approaches to MIR can be discerned: metadata-based and content-based. The more challenging approaches in MIR are thus the ones that deal with the content of the multimedia (video, image, speech, audio, text). Interesting recent methods proposes to extract these characteristics [B.1.2, B.1.3].

The research presented in this report was developped in the context of Mex-Culture project, which aim to develop tools for access to cultural heritage of the Mexican culture. In this task we have designed, developed and implemented methods for processing the speech/audio content of the sound files.

C PROPOSED METHODS FOR DESCRIPTION OF SPEECH/AUDIO ENCODED CONTENT

C.1 SPEECH/MUSIC DETECTION SYSTEM

The LaBRI Speech/Music detection system is built using data from the ESTER evaluation campaigns [C.1.1]. The features used are PLP coefficients which are widely used in speech processing. We train GMM models for each class. Viterbi decoding is used during the test phase. This method has proved to be efficient on the data used in the ESTER evaluation campaign, especially for speech detection (F-measure of 0.93). The same method has been applied to the data of the project. The audio files have been annotated manually for evaluation purposes by several annotators. These annotations were provided by the Mexican partners. There were 5 annotators who annotated a total of 40 hours of audio data (64 files).

C.1.1 THE TRAINING DATABASE

On each file, PLP coefficients [C.1.2] are extracted using the HTK toolkit. These coefficients are computed on 30 ms windows each 10 ms (20 ms overlap). 12 coefficients are extracted and derivatives and acceleration are added, creating a 36-dimensional vector for each frame (see report [ED2-1]).

The models used are Gaussian Mixture Models (GMM). For each of the 34 classes, we estimate a 256 mixtures model. These models are trained using the audioseg library[1].

1 https://gforge.inria.fr/projects/audioseg/

C.1.2 TESTING

During the test phase, viterbi decoding is used. Thus, we obtain automatic segments on which the most probable class is given. As the task is to detect the speech and music segments, labels are automatically simplified to match these two classes. Instead of having the 34 classes, we will only get two files containing respectively speech and non-speech events and music and non-music events. A post-processing step is also carried out to discard very short segments.

C.1.3 EVALUATION

In Speech/music detection the performances obtained are a little but lower than what we had on the ESTER database, with F-measure for speech detection varying from 0.82 to 0.88 according to the annotator (between 11% and 13% of error).
Unfortunately, the results for music detection are not very satisfactory, with F-measures ranging from 0.56 to 0.67 according to the annotator, corresponding to error rates between 34% to 40%.

C.2 AUDIO SEGMENTATION METHOD

In order to obtain the best performance in classifying and identifying audio content for the Mex - Culture project´s tasks, a method for audio segmentation was developed.

The objective of this method is to classify into three different categories (voice, music and silence), fragments of audio file content. This is based on two techniques called Root mean square and Zero crossings. Employing the results of this segmentation method an improvement is done by creating new files that contain one of each categories.

Segmentation of an audio file into different categories like music and voice is a process that involves several stages as it shows in the diagram on figure C2.1.

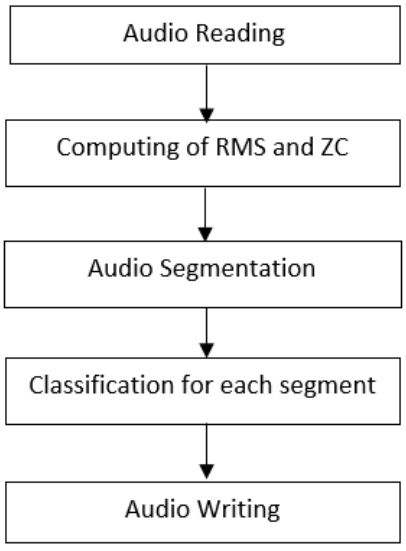

Figure C2.1. Audio segmentation stages.

C.2.1 ROOT MEAN SQUARE ANS ZERO-CROSSING

The root mean square (RMS), is defined as the square root of the arithmetic mean of the squares of a set of numbers. RMS can also be defined for a continuously varying function in terms of an integral of the squares of the instantaneous values during a cycle.

Zero crossings (ZC), is a point where the sign of a mathematical function changes (e.g. from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function. It is a commonly used term in electronics, mathematics, sound, and image processing [C.2.1].

The process stars by applying Root mean square (RMS) and Zero crossings (ZC) techniques. RMS is applied on data values fragments that represent the audio signal. It gets the average value that is later used to analyse the audio signal.

ZC is used as a technique that can determine if an audio signal is voice or music. Certainly, the voice varies or has more pauses than music, causing more crossings on the x-axis [C.2.2]. This is done using iterations that go from the start of the audio to the end of its duration. In the first iteration there is a verification to see if the audio has one or two channels. In the case that two channels exist, it is needed to create a single channel to process only one signal using RMS and ZC. After this, the RMS and ZC are calculated using the values corresponding to 20 milisecons (msec) pieces of audio because it is an interval where the voice is more stable, which makes it easier to detect as said in [C.2.3].

C.2.2 AUDIO SEGMENTATION

In this stage the algorithm segments the entire audio file. The method detects if a change is presented between the frames. It only detects the change if the frames are too differents.

During the segmentation a general sampling to detect where there are changes within the audio is done. In this method a vector P is obtained (probability vector). The probabilities of change are stored, once the vector gets the average of the vector. Utilizing the probability function, there is a computation of the changes that occurs in the 20 msec of signal. The stronger changes or those that take a larger part of the 20 msec signal are the ones that are considered to be part of one type of categorie.

C.2.3 CLASSIFICATION OF SEGMENTS

Each segment defined is classified in one of the three categories: silent, voice and music. The technique starts reading the frequency segment. Then it takes into account the RMS and past zero to determine their classification. After the verification of each segment we concatenate the part of the original audio vector corresponding to said segment to an exclusive final vector (voice, music, silence or music and voice), and the process continues until all the signal is analyzed. If the segment belongs to either music or voice it its stored into the finalMV vector, and next to the vector that contains only voice (finalVoz) or music (finalMusica).

In the Figure C2.2 it can see the results of the audio at the moment of analyze. It can be the music and Spanish speak. In the Figure C2.3 the audio is also segmented but in this case, it is used an audio of indigene language (Pai-pai) and it segmented also properly ant it detects the changes of the frequency.

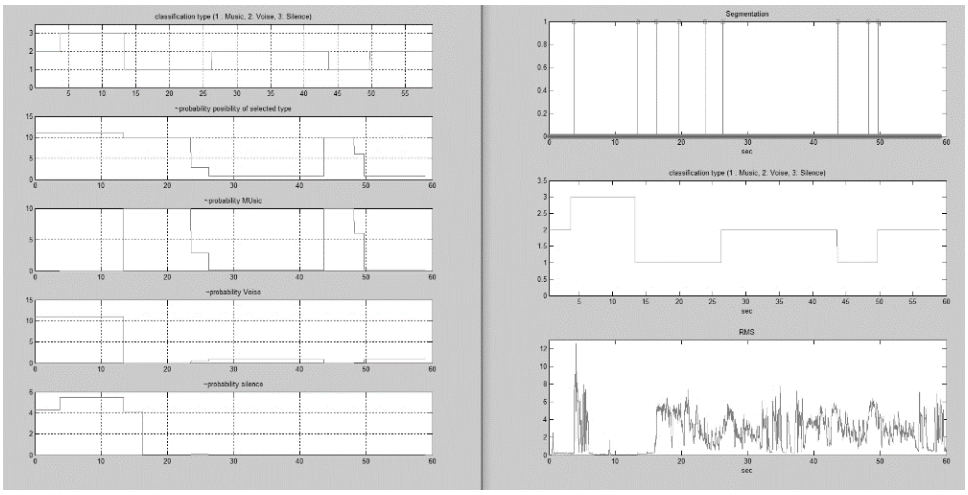

Figure C2.2. Audio segmentation. Music and speak in Spanish.

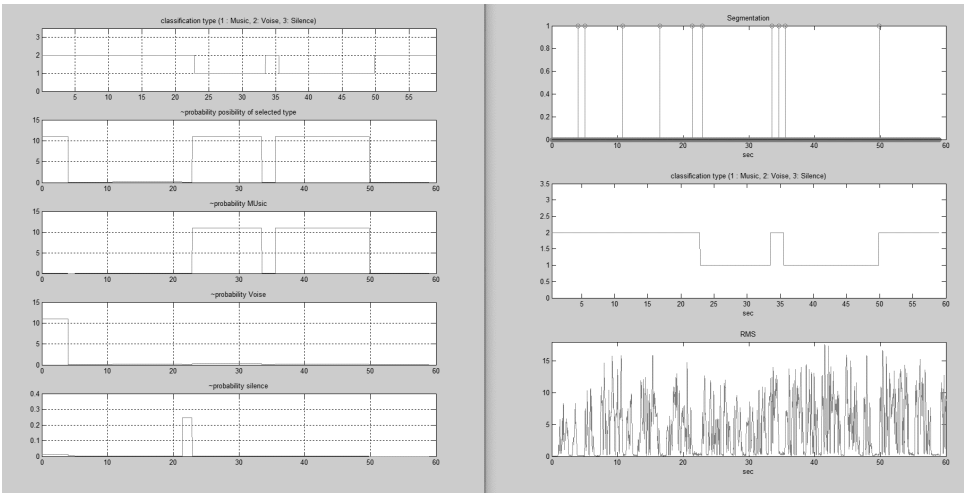

Figure C2.3. Audio segmentation. Indigene language (Pai Pai).

C.3 IDENTIFICATION OF FIVE CLASES MEXICAN SONOROUS CONTENT BASED ON FONOTECA NACIONAL MÉXICO CLASSIFICATION

Fonoteca Nacional México [C.3.1] makes the classification of its sonorous content in five classes: Sonorous Art, Music, Landscape Sound, Radio and Voice. These five classes are a significant part to make her catalog of sonorous files.

In raport [ED2-1], we saw the type of sonorous content that Fonoteca database has. After analyzing the sonorous content that is in each of the five classes, we conclude that : in general, each class is composed of a mixture of sonorous signals: music with different types of genres, only voice, voice with music, voice with different sounds in backgrounds, nature

sounds, animal sounds. In addition, there is different level of noise in all sonorous files of the database.

The idea is to have a first classification of sonorous files with regard to the five classes of the cataloging of the Fonoteca Nacional México. Wold et al. [C.3.2] list some features that are commonly extracted from acoustic frames with a duration between 20 and 40 milliseconds (see raport [ED2-1]). Initially, we cannot use that very specific descriptors such rhytms or descriptors for speech (pitch, formants,…) [C.3.2]; but using combinations of specific descriptors of music and speech may be an option, but it must be considered that these descriptors are not too heavy (see raport [ED2-1]). Our contribution is an interesting classification of sonorous files with regard to the five classes of the cataloging of the Fonoteca Nacional México. Then, we do a deeper analysis of each class. This, in order to generate new descriptors that will provide information on the content in more detail. With these descriptors we have a better indexing of these five classes.

We implemented descriptors that we call them as landmark-based sonorous recognition. Landmark-based sonorous recognition analysis provides a means to relate acoustic events to sonorous behavior of the acoustic content of each of the five classes, thereby allowing comparison of the query with the content of each classes along a set of distinct acoustic parameters (see raport [ED2-1]). A hash table is constructed to store all the locations of each landmarks-based sonorous recognition codebooks hash value. Landmarks-based sonorous recognition codebooks locations are stored in the table with an identification number from the originating sonorous class and a time offset value.

C.3.1 SONORUOS DATABASE OF FONOTECA NACIONAL MÉXICO

We generated a database for testing the identification system of five classes.

In the training database each class consist of five sound files selected from the database. In total 25 sounds files. The set comprises 2.2 hours of sonorous signal. All sounds files are sampled at 44100 Hz.

The test database is composed of ten sound files selected from the database and these files are not part of the training database. In total 10 sounds files. The set comprises 1.2 hours of sonorous signal. All sounds files are sampled at 44100 Hz.

C.3.2 MATCH EVALUATION

The integration of developed segmentation method (see section C.2) voice / music / silence, allows useful information to process the audio content. The sound files for each sonorous class was processed obtaining theirs landmarks-based sonorous recognition descriptors. After this, it is processed the cluster algorithm k-means obtaining landmarks-based sonorous recognition codebook of size 512. As we look in Table C3.1, we have implemented the landmarks-based sonorous recognition codebook option and we had a 40% efficiency, the database of training was processed with the integration segmentation method and stored in the hash table. It has obtained an identification efficiency of 91%.

In the test database, they are two sound files for each sonorous class. The classes are marked as: 01 – Sonorous Art; 02 – Music; 03 – Landscape sonorous; 04 – Radio; 05 – Voice.

Test audio files	**k-mans 512**		**Segementation method with k-means 512**	
	identification	Class identified	identification	Class identified
01A	x	04	✓	01
01B	✓	01	✓	01
02A	x	03	✓	02
02B	x	05	✓	02
03A	✓	03	✓	03
03B	x	05	✓	03
04A	x	03	x	03
04B	✓	04	✓	04
05A	✓	05	✓	05
05B	x	01	✓	05

Table C3.1. Match evaluation of the sounds files.

C.4 RECOGNITION OF MEXICAN INDIGENOUS LANGUAGES SOUND DOCUMENTS

Our contribution is the identification of languages indigenous based on landmark inspired on speech fingerprinting descriptors (see raport ED2-1). The separation of the languages is based on the time-frequency information of the higher frequencies concentration.

We worked on the automatic identification of four indigenous languages of México: 01 – Maya -Yucatec ; 02 – Nahuatl - Central ; 03 – Otomi variant Hñahñu ; 04 – PaiPai.

- *Yucatec Maya language*

Is a Mayan language spoken in the Yucatán Peninsula and northern Belize. Maya remains many speakers' first language today, with approximately 800,000 to 1.2 million speakers.

- *Central Nahuatl language*

Known informally as Aztec, is a language or group of languages of the Uto-Aztecan language family. Varieties of Nahuatl are spoken by an estimated 1.5 million Nahua people, most of whom live in Central Mexico.

- *Otomi – Hñahñu language*

It is spoken in the Mexican state of Hidalgo, especially in the Mezquital Valley, by 100,000 people.

- *Paipai language*

Paipai is the native language of the Paipai peoples. It is part of the Yuman language family. There are very few speakers left because most Paipai now live in Kumeyaay villages.

Each automatic identification system language, accepts a voice file, as input, and makes the membership decision. The output represents the similarity values for each language.
Each reference langage indigenous is described by the many hundreds of descriptors landmark-based speech fingerprinting, it contains the times at which they occur. This information is held in an inverted index, which, for each of hundred distinct descriptors landmarks-based speech fingerprinting, lists the language indigenous it belongs and when they occur in those structure acoustic.

C.4.1 SONOROUS DATABASE OF LANGAGES INDIGENOUS OF MÉXICO

We generated a database to test the system for identification of four indigenous languages (see section C.4). The four indigenous languages were selected on the criteria suggested by the experts group INAH-DL (National Institute of Anthropology and History - Linguistic Department). Among the most significant criteria are : different family; phonetically different family and languages in danger of extinction.

The database training comprises 1:02:08 hours of speech files. All sounds files are sampled at 44100 Hz.

The database test is composed of 14 speech files selected in the database and they are not part of training database. In total 14 speech files. The set comprises 00:23:33 hours of speech signal. All sounds files are sampled at 44100 Hz.

C.4.2 MATCH EVALUATION

The integration of developed segmentation method (see section C.2) voice /silence, allowed generate useful information to process the speech content.

The speech files for each langage indigenous was processed for obtained theirs landmarks-based on speech fingerprinting descriptors. They are parameterized by the frequencies of the peaks and the time in between them. Then, it was processed by the cluster algorithm k-means for obtained the landmarks-based speech fingerprinting codebook of size 512. The training database was processed as above and stored in the hash table.

In the test database, they are three speech files for each langage indegenous. The indigenous languages are marked as: 01 – Maya Yucatec; 02 – Nahuatl of center; 03 – Otomi - Hñahñu; 04 – Paipai.

Test speech files	**k-mans 512**		**Segementation method with k-means 512**	
	identification	Language indegenous identified	identification	Language indegenous identified
01A-Test	✓	01	✓	01
01B-Test	x	03	✓	01
01C-Test	x	02	✓	01

01D-Test	x	03	✓	01
02A-Test	x	03	✓	02
02B-Test	✓	02	✓	02
02C-Test	✓	02	✓	02
03A-Test	✓	03	✓	03
03B-Test	✓	03	✓	03
03C-Test	x	01	✓	03
03D-Test	✓	03	✓	03
04A-Test	x	03	✓	04
04B-Test	x	02	x	02
04C-Test	x	01	✓	04

Table C4.1. Test identification results.

As we look in Table C4.1, we have obtained an identification efficiency of 42.86% for the system with clusters. Then, we implemented the option segmentation method and we had a 93% efficiency.

C.5 SPEECH RECOGNITION

In this part we focus on implementing algorithms that have the best performance in systems reported in the literature and doing optimizations to get the best performance around Mex-Culture project conditions. Speech recognition is a challenging problem on which much work has been done the last decades. The hidden Markov model (HMM) is the best technique when working with speech processing [C.5.1].

In this stage we worked on optimizing the parameters of HTK toolkit and the test configuration for Markov hidden models with the aim to improve the performance in the recognition task.

A isolated-word speech recognition system trains one hidden Markov model for each word that it should be able to recognize. The models are trained with labeled training data, and the classification is performed by passing the features to each model and then selecting the best match. The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The number of coefficients needed is 13. To increase the information of the human perception, the first and second time derivative are calculated.

Number	Feature
12	MFCC
1	Energy
13	Delta-cepstral
13	Delta-delta cepstral

Table D6.1. Total features per short-frame.

C.5.1 Results

In this stage, we used the same database to see the performance of new configurations of Markov models.

The training database is formed by the english words of digits one to five, spoken by 28 females and 18 males speakers. There are two repeats for each word by all speakers. Total words for the training database are 460. There are 92 samples of each word.

The test database is formed by the english words of digits one to five in english, spoken by 28 females and 28 males speakers. There are two repeats for each word by all speakers. Total words for the test database are 560. There are 112 samples of each word. The training results in five models (one for each word). These models are trained using the HTK toolkit. These models represents a statistical model (HMM) of each word. The structure of the model is the left-right model. We varied the number of states to analyze performance over the state of the model number. The HMM for each word has 6 to 10 states.

To evaluate the performance of the speech recognizer the trained models were used, recognizing the utterances one to five. The recognition rates are presented in Table C5.1.

States	Utterances one to five
6	97.85
7	99.30
8	99.58
9	99.51
10	99.54

Table C5.1: word recognition rates %.

This is fairly good results. The better performance is obtained for the model HMM with 8 states.
This corroborates that the HMM model with 8 states, will have the best performance of recognition, since in tests reported in ED2-1 it also had a good performance.

D Conclusion and Perspectives

Based on the results section D.4, we conclude that increasing the number of sound files in the training database and increasing the size of the number of clusters can give us better results.

The assumptions based on the use of the locations of pairs of spectrogram peaks as robust features for matching is a good option. So we propose that it must be analyzed in detail the window size used for peer of spectrogram peaks and see the possible use of a segmentation of events such as music and speech.

Another part that we have noted, is that the cataloging of these five class made by the Fonoteca Nacional staff and it is based on experience and knowledge of Mexican music. So we can say that there are sound files belong to more than one sonorous class. This aspect is important to consider in the selection and calculation of descriptors that we want to develop. The robustness of this approach is that only a few of the acoustic landmarks have to be the

same in the refererence and query examples to allow a match. If the query example is noisy, or filtered strangely, or truncated, there's still a good chance that enough of the hashed landmarks will match to work.

E REFERENCES

[B.1.1] Downie, J.S., Music Information Retrieval. Annual Review of Information Science and Technology 37 (2003) 295–340.

[B.1.2] Bosma, M., Veltkamp, R.C., Wiering, F.: Muugle: A Music Retrieval Experimentation Framework. In: Proceedings of the Ninth International Conference on Music Perception and Cognition, Bologna 2006, 1297–1303.

[B.1.3] Pickens, J., Bello, J.P., Monti, G., Sandler, M., Crawford, T., Dovey, M., Byrd, D.: Polyphonic Score Retrieval Using Polyphonic Audio Queries: A Harmonic Modeling Approach. Journal of New Music Research 32 (2003) 223–236.

[C.1.1] Galliano, S.; Geoffrois, E.; Gravier, G.; Bonastre, J.-F.; Mostefa, D. & Choukri, K. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news Language Resources and Evaluation Conference (LREC), 2006.

[C.1.2] Hermansky, H. Perceptual Linear Predictive (PLP) analysis of speech Journal of the Acoustical Society of America, 1990, 87, 1738-1752.

[ED2-1] Mid-term report on speech/audio descriptors tools. Mex-Culture project, 2014. ANR-CONACYT.

[C.2.1] Zero crossing, Wikipedia, 2016. https://en.wikipedia.org/wiki/Zero_crossing.

[C.2.2] Elizabeth Cano, Osvaldo Pérez. 2016. Report of Speech/Music Discriminator using RMS and zero-crossings. Internal report for Mex-Culture project.

[C.2.3] C. Panagiotakis and G. Tziritas, 2005. A speech/music discriminator based on RMS and zero-crossings, IEEE Trans. Multimedia, vol. 7, no. 1, pp. 155-166.

[C.3.1] Fonocata Nacional de México, Conaculta : http://www.fonotecanacional.gob.mx/

[C.3.2] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search, and retrieval of audio. IEEE Multimedia, 3(3):27–36, 1996.

[C.5.1] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, vol. 77, pp. 257-286, Feb. 1989.