Projet ANR- 11-IS02-001

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

Deliverable

MID-term 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* 31/12/2014
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	*Alejandro Ramirez Acosta*
Telephone	*(+52) 664 1 85 46 34*
E-mail	*ramacos10@hotmail.com*
Date of writing	*03/10/2014*

Redactors :	Mireya Saraí García Vázquez (CITEDI-IPN) Alejandro Ramírez Acosta (CITEDI-IPN) Jean-Luc Rouas (LABRI) Henri Nicolas (LABRI)

B INTRODUCTION

B.1 MOTIVATION

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. The main subject of this task is the computational modeling of similarity relation between sound files for retrieval purposes. In this task we must design, develop and implement methods for processing the speech/audio content of the sound files. The main challenge is to formulate the speech/audio problem at hand in terms of the algorithm and corresponding data structure. Examples of such data structures are a sequence of symbols, a vector of feature values, a point set in a geometric space, and so on.

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 deployment and integration of audio processing tools can enhance the semantic annotation of multimedia content, and as a consequence, improve the effectiveness of conceptual access tools.

Recent investigations have shown the feasibility of deploying large vocabulary speech recognition for the generation of multimedia annotations that allow the conceptual querying of video content and the synchronization to any kind of textual resource that is accessible, including other full-text annotation for audiovisual material.

B.2 AUDIO-VISUAL CORPORA

The methods devised in the project will be applied to the large databases of the FONOTECA NACIONAL (National Sound Archive of México), part of CONACULTA (National Council for Culture and the Arts of México) and the Video library (TVUNAM, more than 100,000 hours of video) of the UNAM (National Autonomous University of México). Since these databases will only become progressively available during the project, a large video database provided by INA (sub-contractor of Cnam) will be employed for the early evaluation of the methods devised in this project.

B.3 GENERAL FRAMEWORK FOR DESCRIPTION OF SPEECH/AUDIO ENCODED CONTENT

The integration of digital technology contributes to the preservation of sound heritage and facilitates access and dissemination to a large number of people simultaneously.

The main challenge in the sound files is, on the one hand, preservation, growth and diversification of its users and for other hand, look for other mechanisms to facilitate information retrieval based on audio content. These search mechanisms are based on numerical methods based on the digital processing of the sound signal.

One of our goals in this projet is to devise and avaluate new methods in which automatic speech and audio analysis can contribute to increased granularity of automatically extracted metadata.

C STATE-OF-THE-ART IN DESCRIPTION OF SPEECH/AUDIO ENCODED CONTENT

This section presents a brief overview of the state of the art in the domain of cross-media description of speech/audio content.

C.1 CROSS-MEDIA DESCRIPTION OF SPEECH/AUDIO CONTENT

Multimedia Information Retrieval (MIR) is a multidisciplinary research, contributing disciplines how computer science, information retrieval, video/audio engineering, multimedia signal processing, music theory, photograph theory, library science, cognitive science and others [C.1.1].

Two main approaches to MIR can be discerned: metadata-based and content-based. In the former, the issue is mainly to find useful information describing the multimedia item; these informations are expressed in text. Hence, existing text-based retrieval methods can be used to search those descriptions. The more challenging approaches in MIR are thus the ones that deal with the content of the multimedia (video, image, speech, audio, text), e.g., for video/image: color, shape, texture and movement; for speech/audio: pitch, melody and rhythm. For the content-based speech/audio, the humans in general have well-developed abilities to extract features from a acoustic signal: they can distinguish pitches, melodies and harmonies, rhythms and beat patterns, they can identify instruments, and at times they are strongly moved by the emotions these features evoke. Extracting these features from speech and audio through signal processing and using them to let people retrieve the information they like seems the obvious thing to do. This however has proved to be very difficult but no imposible. Interesting recent methods proposes to extract these characteristics [C.1.2, C.1.3, C.1.4].

C.2 SPEECH AND MUSIC DETECTION SYSTEM

The LaBRI Speech/Music detection system is built using data from the ESTER evaluation campaigns [C.2.1]. The features used are PLP coefficients which are widely used in speech processing. We train GMM models with 256 components for each class. Viterbi decoding is used during the test phase.

C.3 MEXICAN SONOROUS CONTENT IDENTIFICATION OF FONOTECA NACIONAL MÉXICO

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

Sonorous file cataloging is one of the most difficult task, it requires a cataloging personal with a generous culture and in some cases, be a specialist in some areas.

Fonoteca Nacional México provided a database of 25 hours of sonorous files. These sonorous files belonging to her five classes.

The type of content in this database is :

- Any type of music : Mexican popular and traditional music , music of natives peoples, romantic music, Pre-columbian music, children's music, christmas music, ranchera music, latin-american popular music, protest song , mexican jazz, nueva trova, mexican music concert , etc.
- Discours, interviews , opening ceremony , roundtable , etc.
- Newspapers spoken (emissions mexican radio broadcast), videos with narration mexican customs, reportage, etc.
- Music, sound experimentation, sheet music, etc.
- Endangered sounds, field recording, nature sounds , animal sounds, etc.

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, animals sound. In addition, there is different leve of noise in all sonorous files of the database.

According to the database characteristics mentioned above, the descriptors that will characterize the sonorous content of the database must take into account both general characteristics of the acoustic signal based in the music and the voice, but at the same time, specific characteristics in details of the acoustic signal. Also, they must take into account the presence of noise, like background noise, noise channel, environmental noise, etc. Another feature is that the sonorous files have different characteristics of record; can have professional recordings as amateur recordings with bad quality.

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. Then, we will 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 will have a better indexing of these five classes.

For this first identification we are inspired on acoustics landmarks works [C.3.2, C.3.3, C.3.4, C.3.5]. We implemented descriptors that we call them as landmark-based sonorous recognition.

C.4 MEXICAN INDIGENOUS LANGUAGES RECOGNITION

The automatic language identification is a relatively new task, which is becoming increasingly important, although it is still far from solving the problem to an acceptable level of identification success rate in a real environment.

The task in this part of the work of Mex-Culture project is to implement the necessary resources developing tools that can help us to identify an indigenous language of Mexico among several other indigenous languages of Mexico. This identification will be done by analyzing sound files that we call sound documents.

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

The type of content that will focus in this research is:

- Radio News (emissions Mexican indigenous radio).
- Music.
- Autochtones Tales.
- Videos with narration Mexican indigenous customs.

We will do as a first step, the identification of languages indigenous based on landmark-based speech fingerprinting descriptors, this as a first separation of the languages, based on the time-frequency informations of the concentrate of higher frequencies.

C.5 Speech/speaker recognition

The speech recognition and speaker work was addressed in two stages. As first part, we addressed speech recognition implementation. In this part we focus on implementing algorithms that have the best performance in systems reported in the literature. In this way, we do optimizations to get the best performance around Mex-Culture project conditions.

The type of content that will focus in this research is:

- Radio News (the emblematic emissions of Mexican radio broadcast).
- Keynote speech of some important figures in Mexican life.

We will describe the recognition of isolated words implementation; Once validated it, these work will be included in the development of the continuous speech recognition.

D Proposed methods for description of speech/audio encoded content

D.1 Architecture

Indexing and retrieval of multimedia information is based on architecture that we called architecture for scalable search. This architecture will allow to store and organize multimedia information in a scalable manner, allowing a multimedia resource search in large databases scale easily and quickly.

In Figure D1.1 we can see that the decomposition tasks are reflected in this scalable search architecture. It shows then the interrelationship between the different activities of each partner.

Tasks 1, 2 and 3 all develop software components (for extracting content descriptions, for performing content summarization and for scalable search). All software integration activities were grouped in Task 4.

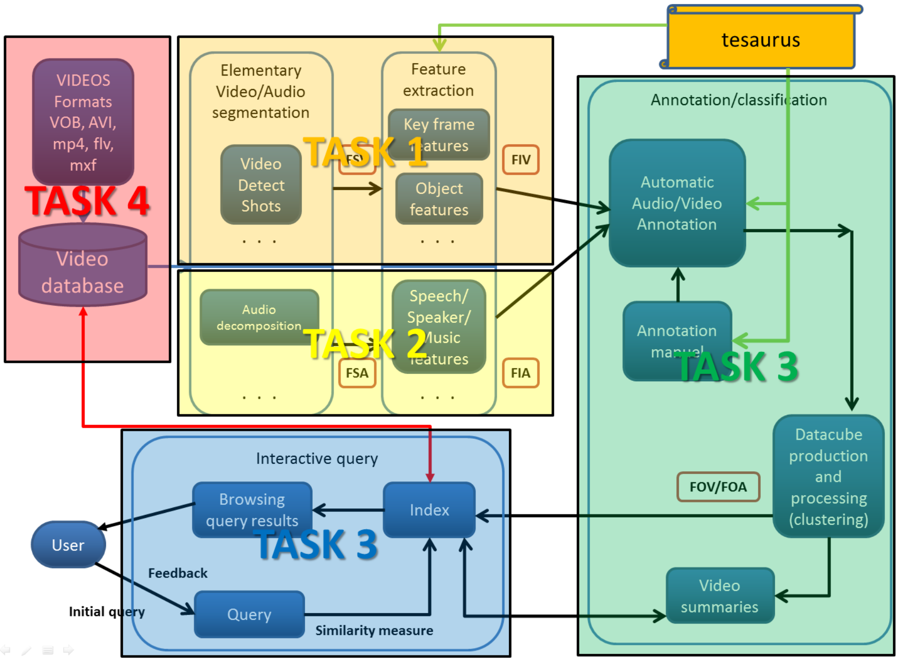


Figure D1.1. Task in the architecture for content-based audiovisual scalable indexing and retrieval.

This report is focused only on the task 2.

D.2 Speech/music detection system

The LaBRI Speech/Music detection system is built using data from the ESTER evaluation campaigns [C.2.1]. 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. The details of the system are given below.

D.2.1 The training Database

The ESTER database is manually annoted in several classes with different annotation tiers for each, which means that some overlap may occur. The classes are :

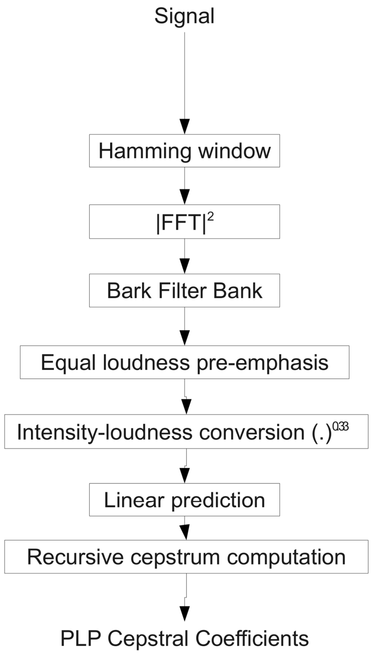
- acapella
- advertising
- applause
- jingle
- laughter
- multiple speech (overlapping speech)
- music
- other
- speech

For the design of our system, we decided to create models using overlapping classes. So, the first step is to merge the annotations, creating 34 classes. The duration for each of these classes is given in the table below :

Category	Training duration (minutes)
jingle+speech	1,8
applause+other+speech	3,4
jingle+music+other	2,7
advertising+other	2,9
advertising+music+other	2,9
multiplespeech2+music+speech	3,2
acappella+music	3,5
laugh+music+other+speech	3,7
multiplespeech1+music+other+speech	3,8
applause+other	5,5
applause+music+other	4,9
advertising+music+other+speech	4,5
multiple_speech2+other+speech	4,8
multiple_speech1+other+speech	5,4
laugh+music+other	6,3
advertising+other+speech	7,0
advertising+music	9,5
multiple_speech1+speech	9,9
multiple_speech1+music+speech	10,0
jingle+music+speech	11,3
laugh+other+speech	18,5
laugh+other	22,8
multiple_speech2+speech	18,8
advertising+speech	20,5
jingle+music	23,5
music+other+speech	27,9
Null	34,1
advertising+music+speech	40,9
music+other	46,0
Other	50,6
other+speech	123,8
music+speech	218,8
Speech	429,7
Music	823,3

D.2.2 Features

On each file, PLP coefficients [D.2.1] 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. The general framework used to obtain these coefficients are given in the figure below.


Signal
Hamming window
$|FFT|^2$
Bark Filter Bank
Equal loudness pre-emphasis
Intensity-loudness conversion $(.)^{0.33}$
Linear prediction
Recursive cepstrum computation
PLP Cepstral Coefficients

D.2.3 Models

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

D.2.4 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.

D.2.4 Evaluation

All times are given in seconds.

Target time is the number of seconds where the target class is present in the test set.

Non-target time is the number of seconds where the target class is absent in the test set.

Missed time is the number of seconds where the system has not identified the target class although it was present in the signal.

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

Insertion time is the number of seconds where the system has wrongly decided the target class was present.

% Error is defined as : $\%error = 100\text{x}\,\frac{MissedTime + InsertionTime}{TargetTime + NontargetTime}$

% Miss is defined as : $\%miss = 100\text{x}\,\frac{MissedTime}{TargetTime}$

%FalseAlarm is defined as : $\%FalseAlarm = 100\text{x}\,\frac{InsertionTime}{NontargetTime}$

%Precision is defined as : $\%Precision = 100\text{x}\,\frac{TargetTime - MissedTime}{TargetTime - MissedTime + InsertionTime}$

%Recall is defined as : $\%Recall = 100\text{x}\left(1 - \frac{TargetTime - MissedTime}{TargetTime - MissedTime + MissedTime}\right)$

F-measure is defined as : $F - measure = 2\text{x}\,\frac{PrecisionxRecall}{Precision + Recall}$

	Target time	Non-target time	Missed time	Insertion time	% Error	% Miss	% False alarm	% Recall	% Precision	F-measure
Speech	32279,30	35534,85	957,72	3335,60	6,3	2,9	9,3	97,0	90,3	0,9359
Music	46164,75	25240,11	5477,73	2884,09	11,7	11,8	11,4	88,1	93,3	0,9068

Overall, these results are consistent with the literature. Results available on this database show that our system performs a bit worse than the best system on speech detection (F-measure of 0,9359, best system is 0,9942) but the performance on music detection is better (F-measure of 0,9068, best system was 0,7885).

The system has yet to be evaluated on the Mex-Culture database, for which we are waiting for the manual annotations.

D.3 MUSICAL PROPERTIES OF THE AUDIO

A large part of collections of audio-visual contents consists in audio-visual musical streams, i.e. concert recordings and music video playlists broadcasted through internet services or TV channels. The problem we focus on is the estimation of the temporal boundaries (start time, end time) of western popular music pieces occurring in such streams [D.3.1]. Such an estimation can be useful to navigate within the stream (automatic chaptering) and extract statistical informations from it (e.g. providing the number of music pieces and their occurrences). Moreover, it can help the cross-referencing of music pieces from different multimedia documents for copyright protection.

Estimating the boundaries of music pieces within an audio stream is a difficult problem: instrumental breakdowns can be introduced on purpose by the band during concerts, or by the video producer for scripting issues. On the opposite, music pieces can be played successively without any pause between them, keeping locally similar properties such as a stable timbre or a constant tonality.

We describe the music video stream as a sequence of audio and visual features. As they are extracted from different modalities with different time resolutions, we choose to express them at a common time-scale, empirically set to a sampling period of 0.5 seconds.

We consider musical properties of the audio through the use of Mel-Frequency Cepstral Coefficients (MFCC) and Chroma vectors. A vector of MFCCs is obtained by filtering the log-power spectrum of a signal with bandpass filters whose frequency responses are regularly spaced at the Mel frequency scale. This filtered spectrum is then decomposed with a discrete cosine transform. The resulting set of coefficients roughly describes the spectral envelope of the input signal [D.3.2] and it is often considered as a way to describe its overall musical timbre [D.3.3].

We also consider another feature related to harmony. A Chroma vector is a set of coefficients which quantizes the energy associated to the twelve semi-tones of the chromatic scale over the signal's whole spectrum in western music theory [D.3.4]. They constitute a description of the tonal content of the input signal. An homogeneous sequence of chroma vectors over time can be interpreted as the use of local key.

These audio features are not original but the experiments we performed show that they capture significant musical information related to timbre and harmony. Furhter experiments have to been performed to evaluate and confirm these first results.

D.4 IDENTIFICATION OF FIVE CLASES MEXICAN SONOROUS CONTENT BASED ON FONOTECA NACIONAL MÉXICO CLASSIFICATION

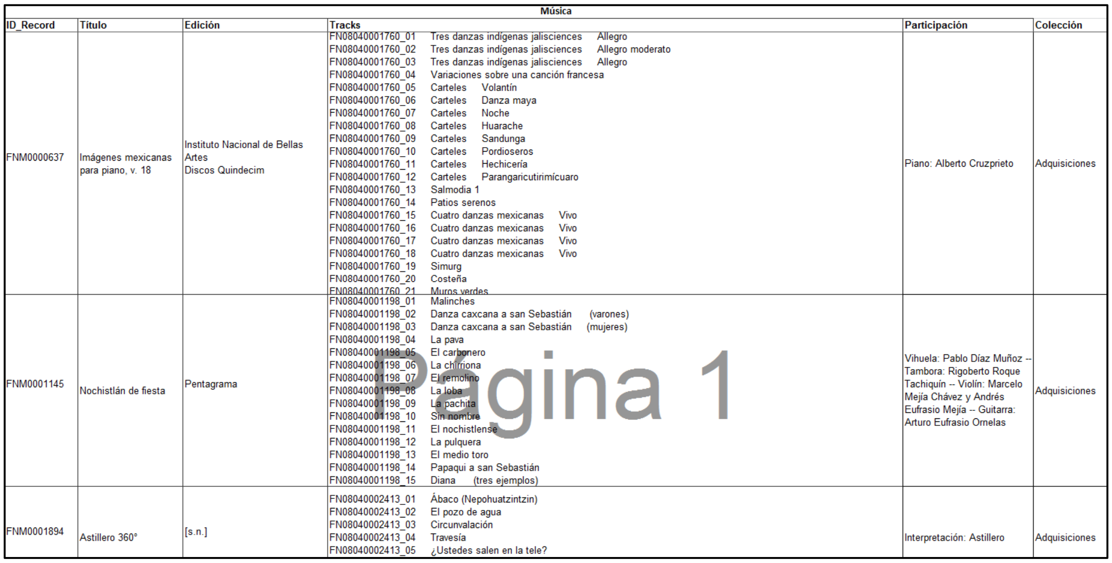
Fonoteca Nacional México [C.4.1] makes the classification of its sonorous content in five classes: Sonorous Art, Music, Landscape Sound, Radio and Voice. These five classes are a significant part to make her catalog of sonorous files.

Sonorous file cataloging is one of the most difficult task, it requires a cataloging personal with a generous culture and in some cases, be a specialist in some areas; for example, analysis of a musical genre, history, sonorous art, soundscape, among others. It is important to observe that the existence of relational database should allow the generation of metadata for each sonorous file; this is part of cataloging called analytical.

The macro-processes, acquisition, preservation and access to the Fonoteca National México is based on the OASIS model (Organization for the Advancement of Structured Information Standards). One of the essential parts is the cataloging which is based on these five classes. Therefore we started developing identification stage of five classes based on the sonorous content.

Fonoteca Nacional México provided a database of 25 hours of sonorous files. These sonorous files belonging to her five classes.

The figure D4.1 shows an overview of the classes of Fonoteca Nacional México. This database has been provided by this institution after a user agreement and responsibility to use the audio content provided. It took several months to Citedi-IPN and Fonoteca personal group to get this database. The work consisted of administrative procedures as well as meetings with the Fonoteca technical staff for the technical specifications for the database.



			Música		
ID_Record	Título	Edición	Tracks	Participación	Colección
FNM0000637	Imágenes mexicanas para piano, v. 18	Instituto Nacional de Bellas Artes Discos Quindecim	FN08040001760_01 Tres danzas indígenas jalisciences Allegro FN08040001760_02 Tres danzas indígenas jalisciences Allegro moderato FN08040001760_03 Tres danzas indígenas jalisciences Allegro FN08040001760_04 Variaciones sobre una canción francesa FN08040001760_05 Carteles Volantín FN08040001760_06 Carteles Danza maya FN08040001760_07 Carteles Noche FN08040001760_08 Carteles Huarache FN08040001760_09 Carteles Sandunga FN08040001760_10 Carteles Pordioseros FN08040001760_11 Carteles Hechicería FN08040001760_12 Carteles Parangaricutirimícuaro FN08040001760_13 Salmodia 1 FN08040001760_14 Patios serenos FN08040001760_15 Cuatro danzas mexicanas Vivo FN08040001760_16 Cuatro danzas mexicanas Vivo FN08040001760_17 Cuatro danzas mexicanas Vivo FN08040001760_18 Cuatro danzas mexicanas Vivo FN08040001760_19 Simurg FN08040001760_20 Costeña FN08040001760_21 Muros verdes	Piano: Alberto Cruzprieto	Adquisiciones
FNM0001145	Nochistlán de fiesta	Pentagrama	FN08040001198_01 Malinches FN08040001198_02 Danza caxcana a san Sebastián (varones) FN08040001198_03 Danza caxcana a san Sebastián (mujeres) FN08040001198_04 La pava FN08040001198_05 El carbonero FN08040001198_06 La chirriona FN08040001198_07 El remolino FN08040001198_08 La loba FN08040001198_09 La pachita FN08040001198_10 Sin nombre FN08040001198_11 El nochistlense FN08040001198_12 La pulquera FN08040001198_13 El medio toro FN08040001198_14 Papaqui a san Sebastián FN08040001198_15 Diana (tres ejemplos)	Vihuela: Pablo Díaz Muñoz -- Tambora: Rigoberto Roque Tachiquín -- Violín: Marcelo Mejía Chávez y Andrés Eufrasio Mejía -- Guitarra: Arturo Eufrasio Ornelas	Adquisiciones
FNM0001894	Astillero 360°	[s.n.]	FN08040002413_01 Ábaco (Nepohuatzintzin) FN08040002413_02 El pozo de agua FN08040002413_03 Circunvalación FN08040002413_04 Travesía FN08040002413_05 ¿Ustedes salen en la tele?	Interpretación: Astillero	Adquisiciones

Figure D4.1. Examples of classes.

For comparing sonoruous files is to extract an abstract description of the acoustic signal which reflects the perceptually relevant aspects of the acoustic signal, followed by the application of a distance function to the extracted information. Usually, an acoustic signal is segmented into short, possibly overlapping frames which last short enough such that there are no multiple distinguishable events covered by one frame.

Wold et al. [D.4.1] list some features that are commonly extracted from acoustic frames with a duration between 25 and 40 milliseconds:

- **Attack** : the duration from a zero to a maximum amplitude,
- **Decay** : the duration from the initial maximum amplitude to a stable state amplitude,
- **Sustain** : the level of the steady state amplitude,
- **Release** : the duration from the steady state to its final zero amplitude,
- **Zero crossing rate** : the rate of sign-changes along a signal,
- **Spectral centroid** : the average frequency, weighted by amplitude, of a spectrum.

Features include frequencies:

- **Loudness** : can be approximated by the square root of the energy of the signal,
- **Pitch** : the Fourier transformation of a frame delivers a spectrum, from which a fundamental frequency can be computed with an approximate greatest common divisor algorithm,
- **Tone** (**brightness** and **bandwidth**) : brightness is a measure of the higher-frequency content of the signal; bandwidth can be computed as the magnitude weighted

average of the differences between the spectral components and the centroid of the short-time Fourier transform, it is zero for a single sine wave, while ideal white noise has an infinite bandwidth,

- **Mel-filtered cepstral coefficients** (**MFCCs**) : can be computed by applying a mel-spaced set of triangular filters to the short-time Fourier transform, followed by a discrete cosine transform, it transforms the spectrum into perception-based acoustic characteristics, a mel is a unit of measure for the perceived pitch of a tone, the human ear is sensitive to linear changes in frequency below 1000 Hz and logarithmic changes above, mel-filtering is a scaling of frequency that takes this fact into account,
- **Derivatives** : since the dynamic behaviour of sound is important, it can be helpful to calculate the instantaneous derivative (time differences) for all of the features above.

Many sounds retrieval systems compare vectors of such features in order to find sound files that sound similar to a given query.

In Section C.3, 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 leve of noise in all sonorous files of the database.

Initially, we cannot use that very specific descriptors such rhytms or descriptors for speech (pitch, formants,…); 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. One option is to use descriptors that must take into account fairly general characteristics of the acoustic signal, but at the same time specific characteristics. Also, they must take into account the presence of noise, like background noise, channel noise, environmental noise, etc. For this, we are looking for descriptors that give us information in space and time at the same time.

Another point to note in the Fonoteca database is that there are sound files with a very short time duration (minimum 5 seconds). Another feature is that the sonorous files have different characteristics of record; we can have professional recordings as amateur recordings with bad quality.

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. Then, we will 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 will have a better indexing of these five classes.

After an analysis detailed audio content of the database provided; we tested different descriptors, this, to get an overview of the features that are important to take into account and that should have a good identifier of sonorous classes. Therefore, the soundtrack matching problem has similarities with that of identifying identical musical recordings in the presence of noise and channel variations. In both cases, we expect to see a lot of invariant underlying structure (e.g. spectral peaks) in the same relative time locations, but possibly corrupted with different channel effects and mixed with varying levels and types of noise. This problem is addressed by a number of prior works in audio fingerprinting [D.4.2].

Audio fingerprinting systems can find matching audio files even when the query contains added noise. Some audio fingerprinting algorithms are better at identifying different types of queries: queries that are short, or have a large amount of noise present in the signal.

An audio fingerprint is a content-based compact signature that summarizes an audio file [D.4.3]. For example, all phone-based systems for identifying popular music use some form of audio fingerprinting. A feature extractor is used to describe short segments of recordings in a way that is as robust as possible against the typical distortions caused by poor speakers, cheap microphones, and a cellular phone connection, as well as background noise like people chatting in a site public. These audio fingerprints, usually just a few bytes per recording segment, are then stored in a database index, along with pointers to the recordings where they occur. The same feature extractor is used on the query, and with the audio fingerprints that were extracted from the query, candidates for matching recordings can be quickly retrieved. The number of these candidates can be reduced by checking whether the fingerprints occur in the right order and with the same local timing.

For this first identification we are inspired on acoustics landmarks works [C.3.2, C.3.3, C.3.4, C.3.5]. 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. The descripteurs are inspired on the approach of [D.4.4, D.4.5] which uses the locations of pairs of spectrogram peaks as robust features for matching. These descriptors create the distinguishing landmarks-based sonorous recognition. The landmarks-based sonorous recognition correspond to concentrated energy localized in time and frequency. A similar approach based on matching pursuit (MP) was presented in [D.4.5] to group similar but nonidentical audio events.

D.4.1 Landmarks-based sonorous recognition

For each sonorous class are calculed and identified the locations of pairs of spectrogram peaks. With this information we obtain a group of landmarks-based sonorous recognition descriptors corresponding to the most energetic points in the acoustics signals of the sounds files of each sonorous class. The cluster algorithm is based in the k-means [D.4.6]; the k-means algorithm iteratively selects a landmarks-based sonorous recognition codebook corresponding to the most energetic points in the acoustics signals of each sonorous class. A landmark-based sonorous recognition is defined by their two center frequencies and the time difference between their temporal centers. The values of landmark-based sonorous recognition codebook are quantized to allow efficient matching between landmarks-sonorous.

The landmarks-based sonorous recognition codebooks values as quantized above can be described as a unique hash of 20 bits. 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.

D.4.2 Sonorous class identification for a sound file query

To find at which of five sonorous classes belong a sound file in a query, it is decomposed in landmarks-based sonorous recognition descriptors and then they are hashed as described above. The hash table is queried for each of the landmarks-sonorous found in the sound file query.

D.4.3 Sonoruos Database of Fonoteca Nacional México

We generated a database for testing the identification system of five classes (see section C.3).

Database : training
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.

Database: test
It 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.

D.4.4 Match evaluation

The sound files for each sonorous class was processed obtaining theirs landmarks-based sonorous recognition descriptors; after this was processed the cluster algorithm k-means for obtained the landmarks-based sonorous recognition codebook of size 512. The database of training was processed as above and stored in the hash table.

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.

As we look in Table D4.1, we have obtained an identification efficiency of 20% for the systems with landmarks-based sonorous recognition descriptors. Then, we have implemented the landmarks-based sonorous recognition codebook option and we had a 40% efficiency.

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

Table D4.1. Match evaluation of the sounds files.

D.5 Recognition of mexican indigenous languages sound documents

The Automatic Language Identification (ALI) identify by means of computer systems, as fast as possible, the language used by a speaker at a given time. Several scientific approaches have been developed for the construction of such sytems. The automatic language identification is a relatively new task, which is becoming increasingly important, although it is still far from solving the problem to an acceptable level of identification success rate in a real environment.

The task in this part of Mex-Culture project is to implement the necessary resources developping tools that can help us identify an indigenous language of México among several other indigenous languages of México. This identification will be done by analyzing sound files that we call sound documents.

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

Optimization of automatic systems that model the languages, they allow confirming or contrasting structural concepts from linguistic theories about the differences between languages. The automatic language identification can be divided according to the nature of handled information [D.5.1-D.5.5]:

- **Acoustic-phonetics** : phonetic inventories differ from language to language. Even when languages have identical phone, the frequencies of occurrence of phones differ across languages.
- **Phonotactics** : phonotactics refers to the rules that govern the combinations of the different phones in a language. There is a wide variance in phonotactic rules across languages.

- **Prosodics** : languages vary in terms of the duration of phones, speech rate and the intonation (pitch contour). Tonal languages such as Mandarin and Vietnamese have very different intonation characteristics than stress languages such as Emglish.
- **Vocabulary** : conceptually the most important difference between languages is that they use different sets of words – that is, their vocabularies differ. Thus, a non-native speaker of English is likely to use the phonemic inventory, prosodic partterns and even (approximately) the phonotactics of her/his native language, but will be judged to speak English if the vocabulary used is that of English.

A ssuccessful language identification algorithm would exploit information from all of the above sources to arrive at its identification decision.

Each automatic identification system language, accepts a voice file, as input, and makes the membership decision. The output represent the similarity values for each language.

As we have seen, the work has as main goal that is the automatic identification of four indigenous languages of México. First, we will do, the identification of languages indigenous based on the descriptors landmark-based speech fingerprinting, this as a first separation of the languages, based on the time-frequency informations of the concentrate of frequencies higher; then we will do analysis based on acoustic-phonetics, phonotactic and prosodic for have an automatic identifier indigenous languages of México more robust.

D.5.1 Description indigenous languages of México

D.5.1.1 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.

Maya - Yucatec	
ISO 639-3 Language Code	yua
ISO 639-3 Language Name	Yucateco
Native speakers	770,000
Population	735,000 in Mexico (2000 INALI). Population total all countries: 766,000. 58,800 monolinguals (2007).
Classification family	Mayan • Yucatecan ○ Yucatec-Lacandon ▪ Yucatec
Where Yucateco is spoken	Belize Guatemala Mexico

Table D5.1. Maya-Yucated language.

D.5.1.2 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.

Nahuatl - Central	
ISO 639-3 Language Code	nhn
ISO 639-3 Language Name	Central nahuatl
Native speakers	1.5 million
Population	40,000 (1980 census). All Nahuatl variety speakers: 1,380,000. 1,000 monolinguals (1990 census). Ethnic population: 63,000 (1986).
Classification family	Uto-Aztecan • Aztecan ○ General Aztec ▪ Nahuatl
Where Central Nahuatl is spoken	Hidalgo Puebla State of Mexico Tlaxcala

Table D5.2. Nahuatl – Central language.

D.5.1.3 Otomi – Hñahñu language

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

Otomi del valle del Mezquital - Hñahñu	
ISO 639-3 Language Code	ote
ISO 639-3 Language Name	Mezquital
Native speakers	1.5 million
Population	130,000 (1990 census)
Classification family	Oto-Manguean • Oto-Pamean ○ Otomian ▪ Northwestern Otomi
Where Otomi - Hñañu is spoken	Hidalgo

Table D5.3. Otomi – Hñahñu language.

D.5.1.4 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.

Paipai	
ISO 639-3 Language Code	ppi
ISO 639-3 Language Name	paipai
Native speakers	100 (2007)
Population	---
Classification family	Yuman • Core Yuman ○ Pai ▪ Paipai
Where Paipai is spoken	Ensenda

Table D5.4. Paipai language.

D.5.2 IDENTIFICATION INDIGENOUS LANGUAGES OF MÉXICO

In this stage we only us speech files for each langage. In order to identify indigenous languages, we will based on the calculation of the landmark-based speech fingerprinting descriptors. It is inspired on the works in acoustic landmarks [C.3.2-C.3.5]. These landmark-based speech fingerprinting descriptors are the prominent onsets concentrated in frequency. This descriptors are the onsets formed into pairs, they are parameterized by the frequencies of the peaks and the time in between them. These descriptors are quantized to give a realtively large number of distinc landmark-based speech fingerprinting hashes. The identification procedure follows almost the same pattern as in the recognition of sonorous classes (Section D.4).

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.

To identify a query, it is similarly converted to descriptors landmarks-based speech. Then, the database is queried to find all the reference langages indigenous that share descriptors landmarks-based speech with the queries, and the relative time differences between where they occur in the query and where they occur in the reference structure acoustic of the langage indigenous. Once a sufficient number of descriptors landmarks-based speech have been identified as coming from the same reference structure acoustic langage indigenous, with the same relative timing, a match can be confidently declared.

D.5.3 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 - DL INAH (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.

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

Database: test.
This 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.

D.5.4 MATCH EVALUATION

The speech files for each langage indigenous was processed for obtained theirs landmarks-based speech fingerprinting descriptors; then, it was processed the cluster algorithm k-means for obtained the landmarks-based speech fingerprinting codebook of size 512. The database of training 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	Whitout k-mans		k-means 512	
	identification	Language indegenous identified	identification	Language indegenous identified
01A-Test	x	03	✓	01
01B-Test	x	04	x	03
01C-Test	x	04	x	02
01D-Test	✓	01	x	03
02A-Test	x	04	x	03
02B-Test	✓	02	✓	02
02C-Test	x	04	✓	02
03A-Test	x	01	✓	03
03B-Test	x	04	✓	03
03C-Test	x	04	x	01
03D-Test	x	01	✓	03
04A-Test	x	02	x	03
04B-Test	✓	04	x	02
04C-Test	✓	04	x	01

Table D5.5. Test identification results.

As we look in Table D5.5 ; we have obtained an identification efficiency of 28.57% for the system without clusters ; then, we implemented the option clusters and we had a 42.86% efficiency.

D.6 Speech recognition and speaker recognition

The job for speech recognition and speaker was addressed in two stages. As first part, we addressed the implementation of speech recognition. In this part we will 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.

We will describe the work of the implementation of the recognition of isolated words; once validated, these work be included in the development of the continuous speech recognition.

D.6.1 Speech recognition

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 [D.6.1]. This is the technique used to implement the isolated word speech recognition system. This stochastic signal model is trying to characterize only the statistical properties of the signal. In the HMM design there is a need for solving the three fundamental problems, the recognition, optimal state sequence and the adjustment. Generally, human speech conveys much information such as gender, emotion and identity of the speaker.

The speech recognition consist of two phases:

- **The training** : each speaker has to provide samples of their speech so that the reference template model can be build,
- **The testing**: to ensure the input test speech is match with stored reference template model and recognition decision are made.

First the feature extraction from the speech signal is done by a parameterization of the wave formed signal into relevant feature vectors. This parametric form is then used by the recognition system both in training the models and testing the same. 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.

D.6.2 Feature extraction

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The source speech is sampled at 8000 Hz and quantized with 16 bits. The signal is split up in short frames of 80 samples corresponding to 10 ms of speech. The frames are overlaped with 20 samples on each side. The idea is that the speech is close to stationary during this short period of time because of the relatively limited exibility of the throat.

The method used to extract relevant information from each short frames is the mel-cepstrum method (Mel-Frequency Cepstrum Coefficients, MFCC). MFCC is perhaps the best known and most popular, and will be utilised here. MFCC is based on known variation of the human ear's critical bandwidth with frequency [D.6.2]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz.

MFCC consists of four steps. Each step has its function and mathematical approaches. See figure D6.1. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFCC's are shown to be less susceptible to mentioned variations.

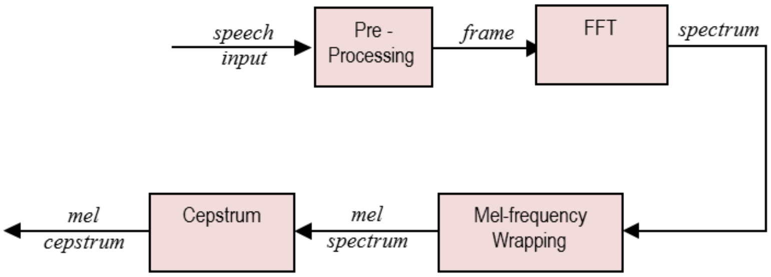


Figure D6.1. MFCC processor.

In the pre-processing step the continuous speech signal is blocked into frames of *N* samples; after is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. Typically the Hamming window is used.

The FFT step converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples {xn}, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,\ldots,N-1$$

In general X_k's are complex numbers and we only consider their absolute values (frequency magnitudes).

The aim in the mel-frequency wrapping step is to have one approach of the spectrum of the human perception of sounds for speech signals. Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the "mel" scale. The mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel-scale. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The number of mel spectrum coefficients, *K*, is typically chosen as 20.

The cepstrum step convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, these can be converted to the time domain using the Discrete Cosine Transform

(DCT). 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.

D.6.3 Feature matching HMM

As mentioned before the technique used to implement the isolated words speech recognition system was the Hidden Markov Model, HMM. In the hidden Markov Model the state at each time t must be inferred from observations. An observation is a probabilistic function of a state [D.6.1].

The observable output from a hidden state is assumed to be generated by a multivariate Gaussian distribution, so there is one mean vector and covariance matrix for each state. The state transition probabilities are independent of time, such that the hidden Markov chain is homogenous. The notation for a hidden Markov model is : total number of N states, an element $a_{ss'}$ in the transition probability matrix A denotes the transition probability from state s to state s', and the probability for the chain to start in state s is π_s. The mean vector and covariance matrix for the multivariate Gaussian distribution modeling the observable output from state s are μ_s and Σ_s respectively. For an observation o, $b_s(o)$ denotes the probability density of the multivariate Gaussian distribution of state s at the values of o. We will sometimes denote the collection of parameters describing the hidden Markov model as
$\lambda = (\pi, A, B)$ [D.6.1]. Where,

π = initial state distribution vector.
A = State transition probability matrix.
$B = \mu, \Sigma$ = continuous observation probability density function matrix.

The three fundamental problems in the Hidden Markov Model design are the following [D.6.1]:

Problem one - Recognition
Given the observation sequence O = (o_1, o_2,...,o_T) and the model λ = (π, A, B), how is the probability of the observation sequence given the model, computed? How is P(O|λ) computed efficiently?.

Problem two - Optimal state sequence
Given the observation sequence O = (o_1, o_2,...,o_T) and the model λ = (π, A, B), how is a corresponding state sequence, q = (q_1, q_2,...,q_T), chosen to be optimal in some sense?.

Problem three – Adjustment
How are the probability measures, λ = (π, A, B), adjusted to maximize P(O|λ)?.

D.6.3.1 The training of a model

Given a N number of observation sequences of a word O_N = {o_1 o_2 o_3 ... o_T}. How is the training of that model done to best represent the word. This is done by adjusting the parameters for the model λ = (π, A, B). The adjustment is an estimation of the parameters for the model λ = (π, A, B) that maximizes P(O|λ). The solution for this is the solutions of the first and third HMM problem [D.6.1]. The signal used for training purposes are ordinary utterances of the specific word, the word to be recognized.

The training is a combination of both supervised and unsupervised techniques. We train one hidden Markov model per word with already classified speech signals. One important choice is the number of different states in each model. The goal is that each state should represent a phoneme in the word. The clustering of the Gaussians is however unsupervised and will depend on the initial values used for the Baum-Welch algorithm.

D.6.3.2 The testing of an observation

When comparing an observation sequence **O** = {o_1 o_2 o_3 ... o_T} with a model **λ** = (**π**, **A**, **B**) we need to find the solution to problem two [D.6.1]. The solution is about finding the optimal sequence of states **q** = {q_1 q_2 q_3 ... q_T} to a given observation sequence and model. There is different solutions depend on what is meant by optimal solution. In the case of most likely state sequence in its entirety, to maximize P (q|O, λ) the algorithm to be used is the Viterbi Algorithm [D.6.1], state transition probabilities has been taken into account in this algorithm. The testing is done in such matter that the utterance to be tested is compared with each model and after that a score is defined for each comparison. The highest score is naturally the highest probability that the compared model has produced the given test utterance. The signals used for testing purposes are ordinary utterances of the specific word, the word to be recognized.

D.6.4 Results

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 D6.2.

States	Utterances one to five
6	98.95
7	99.20
8	99.48
9	99.40
10	99.44

Table D6.2: Word recognition rates %.

This is fairly good results. The better performance is obtained for the model HMM with 8 states.

Experimentation indicated that the most important parameter were the number of hidden states, N.

E 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.

F References

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

[C.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.

[C.1.3] Haitsma, J., Kalker, T.: A Highly Robust Audio Fingerprinting System. Proceedings ISMIR 2002, 107–115.

[C.1.4] 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.2.1] 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.3.1] Fonocata Nacional de México, Conaculta : http://www.fonotecanacional.gob.mx/

[C.3.2] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," Acoustical Society of America, vol. 111, no. 4, pp. 1872–1891, 2002.

[C.3.3] Sharlene A. Liu. Landmark detection for distinctive feature-based speech recognition. Journal of the Acoustical Society of America, 100(5):3417{3430, Nov. 1996.

[C.3.4] Andrew Wilson Howitt. Vowel landmark detection. In Proc. ICSLP, 2000.

[C.3.5] Marilyn Chen. Nasal landmark detection. In Proc. ICSLP, pages 636-639, 2000.

[D.2.1] Hermansky, H. Perceptual Linear Predictive (PLP) analysis of speech Journal of the Acoustical Society of America, 1990, 87, 1738-1752.

[D.3.1] G. Sargent, P. Hanna, H. Nicolas. Segmentation of Music Video Streams in Music Pieces through Audio-Visual Analysis, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14), Florence, Italy, May 2014.

[D.3.2] Logan, B. Mel Frequency Cepstral Coefficients for Music Modeling. Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR) 2000.

[D3.3] Paulus, J. and Muller, M. and Klapuri, A. Audio-based music structure analysis Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), 2010, pages 625—636.

[D.3.4] Bartsch, M. A. and Wakefield, G. H. Audio Thumbnailing of Popular Music Using Chroma-Based Representations. IEEE Transactions on multimedia, vol 7,, n°1, pages 96-104.

[D.4.1] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search, and retrieval of audio. IEEE Multimedia, 3(3):27–36, 1996.

[D.4.2] P. Cano, E. Batlle, T. Kalker, and J. Haitsma., A review of audio fingerprinting. J. VLSI Sig. Proc., vol. 41, no. 3, pp. 271–284, 2005.

[D.4.3] P. Cano, E. Batlle, T. Kalker, J. Haitsma. A Review of Algorithms for Audio Fingerprinting. Proceedings of the International Workshop on Multimedia Signal Processing 2002.

[D.4.4] A. Wang. The Shazam music recognition service. Comm. ACM, vol. 49, no. 8, pp. 44–48, Aug. 2006.

[D.4.5] C. Cotton and D.P.W. Ellis. Finding similar acoustic events using matching pursuit and locality-sensitive hashing. In Proc. WASPAA, 2009, pp. 125–128.

[D.4.6] Linde, Y., Buzo, A., Gray, R. An Algorithm for Vector Quantizer Design. IEEE Trans on Comm., v. 28, (1980).

[D.5.1] Wang, L., E. Ambikairajah, and E.H.C. Choi. Multi-lingual Phoneme Recognition and Language Identification Using Phonotactic Information. In IEEE International Conference on Pattern Recognition. p. 245-248. HongKong. 2006.

[D.5.2] Navratil, J., Spoken Language Recognition-A Step Toward Multilinguality in Speech Processing. IEEE Transactions on Speech and Audio Processing, 9: p.678-685. 2001.

[D.5.3] Tong, R., et al. Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification. in IEEE International Conference on Acoustics, Speech and Signal Processing. p. 205-208. 2006.

[D.5.4] Zissman, M.A. Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models. in IEEE International Conference on Acoustics, Speech and Signal Processing. p. 399-402. 1993.

[D.5.5] Nakagawa, S., Y. Ueda, and T. Seino. Speaker-independent, Text-independent Language Identification by HMM. in International Conference on Spoken Language Processing. p. 1011-1014. 1992.

[D.6.1] 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.

[D.6.2] E.C. Gordon. Signal and Linear System Analysis. John Wiley & Sons Ltd., New York, USA,1998.