Advances on Audio Segmentation and Audio Content Description for Multimedia Documents

Universidad Zaragoza

Diego Castán

Department of Electronic Engineering and Communications

University of Zaragoza

Ph.D. Thesis

Thesis Advisor: Prof. Alfonso Ortega

Zaragoza, November-2014

Para mi compañera de viaje, Conchi. Gracias “pichu”.

Agradecimientos

Desde el inicio hasta el final de este viaje me he visto respaldado y apoyado por un grupo de personas a las que les debo mi más sincera gratitud.

En primer lugar, Alfonso ha demostrado ser un director de tesis excelente. No sólo me ha apoyado a nivel científico, sino también personal cuando los resultados no me acompañaban. Ha sabido motivarme y contagiarme su entusiasmo y es, gracias a él, que este trabajo ha visto la luz.

Gracias también al resto de mis compañeros de laboratorio: a Eduardo por darme esta oportunidad y cuidarnos como a hijos, a David con quien he compartido todo el camino, a Antonio por sus sabios consejos, a Jesús por las batallas libradas en el cluster y a Paola, Jorge y Julia con quienes he compartido descansos y risas. Algunos compañeros han ido escogiendo otros caminos como Carlos, Kike y Óscar a quienes les agradezco infinitamente que me guiaran en los primeros pasos de la investigación. En resumen, gracias a todo ViVoLab por crear un entorno de trabajo tan agradable.

Los amigos también han sido un pilar importante: les debo a Jorge, Pedro, Héctor, Jaime, Kike, Josema y un largo etcétera, momentos de asueto en torno a una cerveza. A Ritu, Jesús y "el Richal" porque, a parte de compañeros de armas, sudor y lágrimas, son amigos y amantes del café.

A toda mi familia que, como no podía ser de otro modo, siempre están presentes: a mis padres, a mi abuela y a mi suegra. Gracias a todos de corazón.

Por último, esta tesis va dedicada especialmente a Conchi. La persona que ha compartido conmigo todo el camino. Espero que lo hayas disfrutado tanto como yo. Sólo nos queda por ver lo que queda por delante y, parafraseando a Bruce Springsteen: "Leave behind your sorrows, let this day be the last, tomorrow there'll be sunshine and all this darkness past".

Abstract

Due to the increase of multimedia content, there is a significant interest in multimedia analysis. This thesis aims at providing information extracted from non-speech audio to retrieve and identify multimedia documents.

The thesis focuses on the detection of complex multimedia events using supervised techniques. For this purpose, the non-speech sounds can be critical since they contain information about the context or the activities where the action is developed. The non-speech sounds are composed of segments of noise or music and small informative sounds known as "Acoustic Concepts" in the literature. The variability of the non-speech sounds, however, is very high and compensation techniques are required.

First of all, we study different supervised techniques for Multimedia Event Detection and propose a solution where the recognition lattices of an HMM-based acoustic concept recognition are used to extract posterior N-gram counts. This approach is compared with an unsupervised technique and merged with a spoken concepts approach. The fusion shows the tremendous importance of having a good segmentation system and a good acoustic concepts detector.

Therefore, secondly we propose a segmentation-by-classification system based on Factor Analysis with two clear advantages. The system does not need class-dependent features with hierarchical structure to classify different classes and the algorithm compensates the within-class variability with high accuracy being able to classify well-defined classes in generic tasks. The proposed method is applied to segment and classify audios coming from TV shows and it is compared with a hierarchical system with specific acoustic features achieving a significant error reduction.

Finally, we study the variability compensation for the detection of acoustic concepts. We compare the performance of the Factor Analysis system proposed for segmentation with baseline approaches widely used. The first approximation to the problem is done by classifying isolated concepts that have been generated artificially. Then, the classification and the detection of the concepts spontaneously generated are studied and we point out the drawbacks of the proposed system.

Resumen

Debido al aumento de los contenidos multimedia, existe un interés significativo en el análisis multimedia. Esta tesis tiene como objetivo proporcionar información extraída del audio sin habla para recuperar e identificar documentos multimedia.

La tesis se centra en la detección de eventos multimedia complejos utilizando técnicas supervisadas. Con este fin, los sonidos que no provienen del habla pueden ser críticos, ya que contienen información sobre el contexto o las actividades donde se desarrolla la acción. Los sonidos sin habla se componen de segmentos de ruido o música y de pequeños sonidos informativos denominados " Conceptos Acústicos " en la literatura. Sin embargo, la variabilidad de los sonidos sin habla es muy alta por lo que se requieren técnicas de compensación.

En primer lugar, se estudian diferentes técnicas supervisadas para detección de eventos multimedia y proponemos una solución en la que se utilizan las celosías (lattices) del reconocimiento de conceptos acústicos basados en HMM para extraer recuentos de N-gramas. Este enfoque se compara con una técnica no supervisada y se fusionó con una solución basada en conceptos hablados. La fusión muestra la tremenda importancia de tener un buen sistema de segmentación y un buen detector de conceptos acústicos.

Por lo tanto, en segundo lugar, se propone un sistema de segmentación por clasificación basada en el análisis factorial con dos ventajas claras. El sistema no necesita características dependientes de la clase ni tampoco precisa de una estructura jerárquica para clasificar las diferentes clases y el algoritmo compensa la variabilidad dentro de la clase con una alta precisión, por lo que es capaz de clasificar clases bien definidas en tareas genéricas.

El método propuesto se aplica para segmentar y clasificar audios provenientes de programas de televisión y se compara con un sistema jerárquico con características acústicas específicas logrando una reducción de errores significativa.

Por último, se estudia la compensación de variabilidad para la detección de conceptos acústicos. Comparamos el rendimiento del sistema de análisis factorial propuesto para la segmentación con soluciones ampliamente utilizadas. La primera aproximación al problema se realiza mediante la clasificación de conceptos aislados que han sido generados artificialmente. A continuación, la clasificación y la detección de los conceptos generados espontáneamente son estudiados y se señalan los inconvenientes del sistema propuesto.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

LIST OF TABLES

Chapter 1

Introduction

Contents

1.1 The Need for Audio Indexation in Multimedia Information Retrieval

Over the last few decades, Information and Communication Technologies (ICT) have evolved dramatically and some of them are worthy of consideration. The number of radio stations and TV channels is increasing day by day and their content is available on the Internet to be accessible from all around the world. Furthermore, the material is replicated on different platforms to ensure its diffusion. Consequently, audio and multimedia materials have rapidly increased their presence due to the popularity of video-sharing websites and audio and video on demand (AVOD) systems. As an example, 100 hours of video are uploaded to YouTube every minute, 1 billion[1] unique users search for content every month and 6 billion hours are watched every month. Likewise, these figures represent a 50% increase over last year's[2].

Videos on sharing websites are defined by text tags chosen by the users to allow other users to access the video they want to watch. In most cases, this information about the video is incomplete and, therefore different methods are needed to improve the retrieval accuracy by giving additional information even when the tags are present. Recently, there has been a growing demand for high-accuracy multimedia indexing and retrieval systems and, due to the nature of the material, research in this field is distributed into different directions which include video information retrieval, audio information retrieval, image search, multimedia indexing and human-computer interactions, among others. According to (Lew2006), all these fields try to meet the two main needs for a multimedia information retrieval (MIR) system: (1) searching for a media item, and (2) summarizing and indexing a media collection. A generic multimedia information retrieval system can be seen in Figure 1.1 where the two main needs are represented.

The difficulty in the tasks of an MIR system can be increased depending on the nature of the material and the final application. For example, applications to search for trademarks, to find pictures with identical visual content or applications to search music to protect the copyright make use of constrained and well-defined material. These fields of applications are defined in the literature as *narrow domains* and are comprised of materials with limited and predictable variability. In this domain, *feature-based*

[1] billion is understood with the English semantic (10^9)
[2] http://www.youtube.com/yt/press/statistics.html

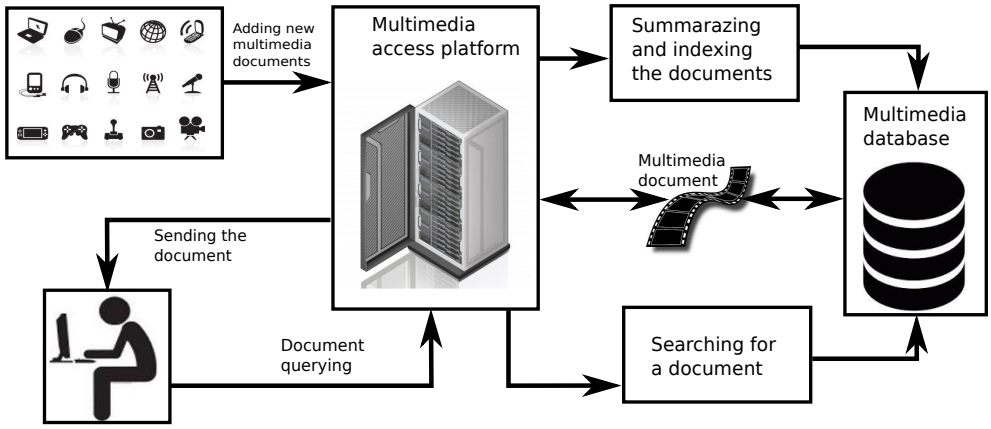


Figure 1.1: Generic block diagram of a multimedia information retrieval (MIR) system.

algorithms describe a homogeneous semantic with a clear interpretation of the features. Antagonistically, these algorithms are not suitable for *broad domains* where material is unconstrained and variability is unpredictable and extremely high. A broad domain can be defined as a set of many domains with very different conditions. This domain presents an important gap between the features and their computational description and, therefore, *low-level feature-based* computational models are not suitable because the number of variables would be very large. This gap is known as *sensory gap* and is defined in (Smeulders2000) as:

> *the gap between the object in the world and the information in a computational description derived from a recording of that scene.*

Moreover, users do not find the exact information that they want because there is a gap between the linguistic description (or the user interpretation) and the information of the data. This gap is known as *semantic gap* and is also defined in (Smeulders2000) as:

> *the lack of coincidence between the information that one can extract from the audiovisual data and the interpretation that the same data have for a user in a given situation.*

Neuroscience researchers know that multiple senses increase human perception as described in (Stein2008). For an MIR task, video features can determine the general content of a video. However, the audio track of the video can also be critical and

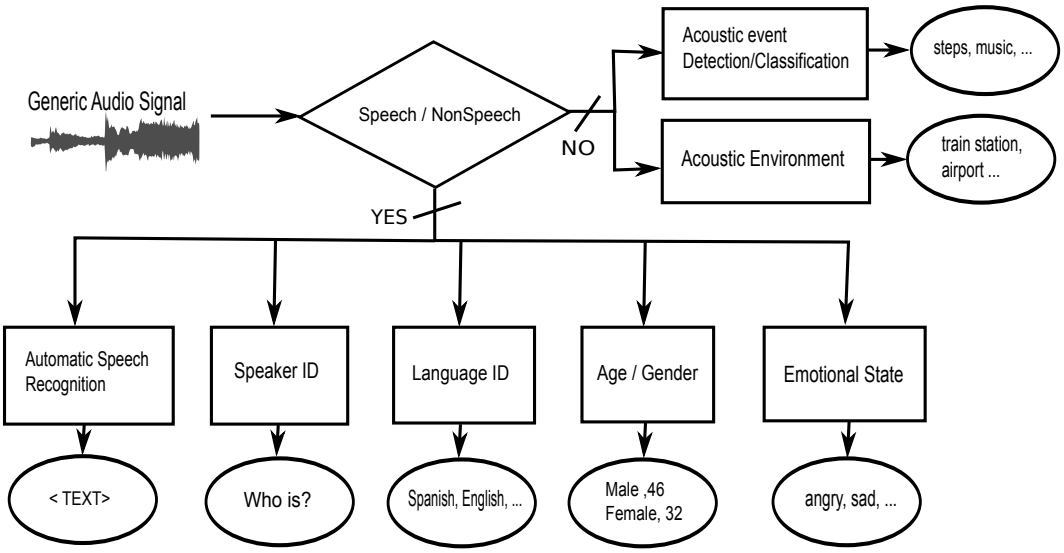


Figure 1.2: Information that can be extracted from a generic audio signal by an audio indexation system. The outputs contribute to provide a *Rich Transcription* of a multimedia document.

it can be very useful to understand the semantics of a user request (this is called "bridging the semantic gap"). Therefore, audio processing reduces the semantic and the sensory gaps in image/video features and it helps to find some specific events easier than other features. Also, audio processing improves the detection of concepts in MIR with complex backgrounds. Consider the case of a video of a tennis match where a special concept, like a new point, may occur. Audio analysis provide complementary information to detect this specific concept (detecting applause or cheering) that would be significantly more difficult to detect with image/video analysis. In other words, the audio indexation systems in MIR compensate the noise and clutter of multimedia collections in broad domains.

Figure 1.2 provides a general block diagram with the information that can be extracted from the audio track of a video by an audio indexation system. The first step is to determine if the audio has speech content or not. If there is speech, the audio will be processed by an Automatic Speech Recognition (ASR) system to obtain the automatic transcription of the video. Also, other systems can identify the speaker, the language, the age and the gender of the speaker and, even, the emotional state of the speaker. If there is no speech, the audio can still be processed to provide additional insight like the

acoustic environment and the acoustic events presented in the scenario (steps, music, ...). Audio indexation systems collect and store all this information in the form of metadata to be used by MIR systems. The extraction process of all available information from an audio signal is known as *Rich Transcription*, since the transcription is not limited to speech, but also to the extra-information. These rich transcriptions allow video search in the same way that current search engines do with text documents and archives.

In the next section, we discuss the use of *Audio Segmentation and Classification* as part of *Rich Transcription* for *Multimedia Event Detection*.

1.2 Audio Segmentation and Classification for Multimedia Event Detection

Multimedia Event Detection (MED) is an important technology for MIR systems. The goal of MED is to assemble detection technologies into a system that can search for multimedia documents that have not been tagged. Those documents must contain the event that the user wants to retrieve. The systems generate metadata based on the event to search in a multimedia database. Therefore, the metadata must be sufficiently general to be re-used for different events requested by the users. Some of the metadata (but not all of them) are given by the *Rich Transcription* of the audio.

To be able to process the audio in the different subsystems of the *Rich Transcription*, the first step is *audio segmentation*. Given an audio document, *audio segmentation* is the delineation of a continuous audio stream into acoustically homogeneous regions. When the *audio segmentation* is followed by a clustering or classification system, the result is a system that is able to classify an audio stream into different classes chosen for a specific task. This process is known as *audio segmentation and classification*[1] in the literature (Reynolds2005) and it can be defined as the determination of occurrences (in time and class) in an audio signal. The *audio segmentation and classification* can be comprised of several subtasks as *speaker diarization* - it aims at answering the question "who spoke when?" - or *music diarization* - it can be focus on different music styles, instruments or tunes - to give two concrete examples. The segmentation

[1] Also known as audio diarization

and classification systems work in two different fashions: (a) unsupervised and (b) supervised.

Unsupervised segmentation and classification is comprised of a *segmentation* system followed by a *clustering* system without prior knowledge of the classes. In a first step, the segmentation system finds the boundaries between different acoustic sources to delimit homogeneous segments. In a second step, the segments are grouped together into similar classes. If the clusters are generated from one class that is fragmented in more classes, the strategy is known as *top-down clustering*. On the other hand, if the clusters are generated as the agglomeration of classes, the strategy is known as *bottom-up clustering*.

Supervised segmentation and classification systems employ models trained with classes that might be present in the audio signal. The delineation of segments can be carried out in two different ways. In the first one, the segmentation and classification are performed in two steps: first, the segments are delimited by a segmentation system and then, each segment is classified into a predefined class with trained models. This fashion of supervised segmentation and classification is known as *classification-after-segmentation*. However, segmentation can be performed as a result of frame-by-frame classification. Therefore, the boundaries of the segments are defined by the transitions between classes and the method is known as *segmentation-by-classification*. These methods have advantages and disadvantages that will be discussed in the following chapters.

Some of the activities on MED are reflected in a rich variety of sounds and noises that we call *Acoustic Concepts* (ACs) (also known as *Acoustic Events*[1]) produced by humans, objects or animals among others. The segmentation and classification of these concepts (different than speech) may help to detect and describe the multimedia event increasing the robustness of the MIR systems.

This thesis focuses on *audio segmentation and classification* to improve MED. More precisely, the thesis aims at advancing in the *segmentation-by-classification* of broad classes and the detection of ACs to produce metadata to improve the detection of multimedia events on a database comprised of videos from the Internet.

[1]We will call them *Acoustic Concepts* (ACs) in this thesis to avoid confusion with Multimedia Events

1.3 Objectives and Methodology

This thesis aims at providing a set of tools to analyze the audio of multimedia documents as a part of a MED system. Following this idea, the objectives of this work are divided into three tasks as described below.

1.3.1 Audio Processing in Multimedia Event Detection

The main goal is to improve the MED providing audio information. More specifically, the framework proposed is:

- Study the classification of broad ACs on a very noisy and cluttered domain.
- Due to the fact that most of the videos have their ACs overlapped with speech, music or noises, a supervised segmentation is needed to detect between speech, music or ACs.
- State-of-the-art in MED with ACs is considered. The improvements in this area will be validated with previous systems in order to provide a clear idea about our approaches.
- We propose two different approaches to describe the temporal behavior of the multimedia documents based on the transitions between ACs.
- Finally, a fusion of AC approaches with ASR will show how the proposed approaches improve the MED task with non-speech information.

As we will show in Chapter 3, the ACs performance and the segmentation is basic for MED. Therefore, this work explores methods to improve the segmentation of a continuous audio stream and the AC detection in a spontaneous generation environment.

1.3.2 Segmentation-by-Classification

We propose an approach to be able to segment and classify and audio stream in broad classes based on Factor Analysis (FA) with two clear advantages: 1) the system uses common audio features and 2) the algorithm compensates the high variability present in each class. The proposed framework is evaluated in a broadcast news domain. This domain is very challenging because it has a very large number of speakers in different environments with background noises. The objectives are:

- State-of-the-art with classic HMM segmentation-by-classification system.
- Explanation and evaluation of the proposed approach based on FA.
- Comparison of the proposed approach with a hierarchical approach with specific features for each class.

1.3.3 Audio Concepts Detection

As we have stated in the last subsection, the ACs are often overlapped with speech, music, noises or other ACs. As a result, the difficulty of detecting the concepts correctly increases. We propose this framework to study AC detection:

- State-of-the-art with isolated ACs.
- AC classification and detection with spontaneous generation. The evaluation of the proposed approaches will be evaluated in a meeting-room domain.

1.4 Outline

The remainder of the thesis is organized as follows.

Chapter 2 describes the state-of-the-art in MED and more specifically the audio technologies in the MED task. The chapter shows the importance of the audio segmentation and classification of acoustic concepts for MED and presents the most common algorithms to model the classes on this task.

Chapter 3 presents a set of systems based on ACs for MED. Firstly, the chapter shows the difficulty in classifying and segmenting the ACs that can be found in MED due to the high variability in the videos from the Internet. Secondly, the chapter describes a baseline system with fixed segmentation and GMMs. This system is compared with a natural evolution system where the segmentation is produced as a result of the transitions among GMM/HMM models. Finally, a novel MED system based on the lattice counts of the ACR is proposed. This system improves the detection of the multimedia events and is also compared with unsupervised approaches for the same dataset. Finally, the results of this system are merged with an ASR approach to take advantage of the performance of both systems since the information is complementary.

Chapter 4 shows a novel segmentation-by-classification system based on FA with two clear advantages. The system does not need class-dependent features with hierarchical structure to classify different classes and the algorithm compensates the within-class variability with high accuracy being able to classify well-defined classes in generic tasks as MED. The proposed method is applied to segments and classifies audios coming from TV shows, and it is compared with a hierarchical system with specific acoustic features, achieving significant error reduction.

Chapter 5 considers the particular problem of the ACD that can be found in a meeting room. The chapter shows the performance of the FA system proposed for segmentation and compares the system with baseline broadly used approaches. The first approximation to the problem is done by classifying isolated ACs that have been generated artificially. Then, the chapter shows the classification and the detection of spontaneously generated ACs. Even if the boundaries of the ACs are given, the systems achieve an extremely high error rate since the ACs are overlapped with speech and have low SNR.

Finally Chapter 6 concludes the work. The main conclusions are summarized in this chapter and several future research lines are highlighted.

Chapter 2

State of the Art

Contents

2.1 Chapter Overview

This chapter compiles and summarizes the most relevant works in multimedia event detection (emphasizing the approaches based on audio information), audio segmentation, acoustic event detection and the most relevant technology employed in these areas.

The sections of this chapter are organized as follows: section 2.2 presents the MED task, a review of the most important datasets and evaluations and a general structure of a MED system with special interest in audio processing. Section 2.3 shows the most important approaches for audio segmentation and classification, speech/non-speech classification and acoustic concept recognition and the relevant technology (features and statistical modeling) for these fields. Finally, a chapter summary is presented in Section 2.4.

2.2 Multimedia Event Detection

Multimedia event detection (MED) is a challenging task due to the high variability in the databases used to generalize the task (for example, videos on the Internet). The goal of the task is to detect events of interest for users in very large video collections. A good definition of an event was given by Jiang et al. in (Jiang2012a) as:

> *a long-term spatially and temporally dynamic object interactions that happen under certain scene settings.*

Each event is defined as high-level content like "wedding ceremony" or "landing a fish" and are comprised of a complex collection of objects, people or sounds. Detecting the desired events is an extremely difficult problem especially from *wild* videos (recorded under unconstrained conditions by non-professional users). For example, Figure 2.1 shows three different images from TRECVid MED 2011 database (NIST2011) representing the same event: "feeding an animal". As it can be seen, the event can be represented by different animals and the action may take place indoor or outdoor. The high variability comes, not only from the different visual/audio elements present in the clip, but also because the videos are recorded with different cameras/microphones and in different scenarios. In addition, the videos can be edited by users with filters that can modify the original colors or add effects and background music. All these artificial elements increase intra-class variability and, therefore, the difficulty of the problem.

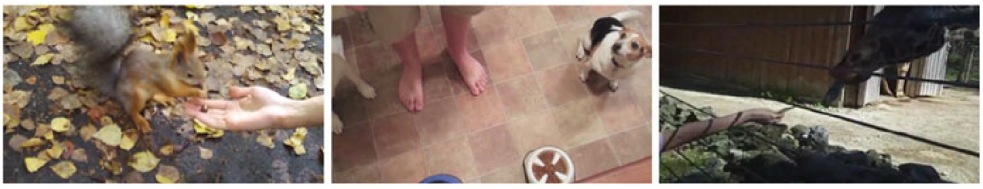


Figure 2.1: Three images of different animals (a squirrel, a dog and a horse) representing the same event "feeding an animal" defined in TRECVID MED 2011 (NIST2011). The clips shows the high intra-class variability present in the task. The pictures have been extracted from (Jhuo2013).

Due to the challenging task of MED, some evaluations have been recently proposed. The evaluations can be divided into two groups: the first group is comprised of datasets captured in controlled environments such as KTH (Schuldt2004), UCF-Sports (Rodriguez2008a), Hollywood2 (Marszalek2009), Weizmann (Gorelick2007) or MSR action (Yuan2009). These datasets are very useful to advance the technology on MED but some approaches may be very specific for a database and may present limitations in unconstrained environments. Therefore, the second group is comprised of unconstrained videos. MediaEval is a cluster of evaluations dedicated to test new algorithms for multimedia access and retrieval. Social event detection is a task inside the MediaEval 2012 (Papadopoulos2012) and MediaEval 2013 (Reuter2013) to discover social events and detect related media items. The evaluations proposed some challenges such as supervised clustering, finding technical events, finding soccer events or classifications between specific events (concert, conference, exhibition or protest as some examples). Columbia consumer video (CVV) dataset (Jiang2011) was created in 2011 with Internet consumer videos without professional edition. The database is comprised of 20 categories divided into 3 groups: objects (such as "dog" or "birds"), scenes (such as "beach" or "parade") and events (such as "biking" or "birthday"). Probably, the most famous datasets for MED are sponsored by the National Institute of Standards and Technology (NIST) in the TRECVid Evaluations (Smaeton2006). These evaluations have been carried out from 2003 until today. The challenge difficulty and the events increase year after year and, therefore, it provides a good benchmark for MED. The TRECVID MED 2011 database (NIST2011) is used in this thesis and it will be described in more detail in the next chapter.

State-of-the-art technologies in MED are presented in following subsections. Nowadays MED systems have a typical structure based on audio/video features, clustering, different classification approaches and a fusion. All these elements will be presented in 2.2.1 divided into thematic subsections. More specifically, an in-depth review about the audio processing in MED will be shown in 2.2.2 to present the most recent approaches in supervised and unsupervised audio concepts recognition.

2.2.1 MED systems

A common factor in all competitive MED systems is the need for multimodal approaches. Most of the videos are recorded with an audio track that can provide useful audio patterns for event detection and, quite often, these audio patterns provide clear information about an event that could be fuzzy, using only video features. On the other hand, some events have clear video characteristics like predominant color distributions or distinctive objects. A considerable amount of literature has been published on MED recently (Jiang2010a, Jiang2012, Jiang2012a, Ballan2010, Jiang2012b, Tamrakar2012, Natarajan2012, Smith2003). All these approaches have a widespread structure based on multimodal feature extraction, clustering/classifying methods and fusion methods as can be seen in Figure 2.2.

2.2.1.1 Multimedia Features in MED

The behavior of a video event can be captured choosing the right features. Good features are robust against variability and are able to describe the same event under different conditions of noise or scenario and therefore good feature selection becomes critical in MED. The features can be divided into two groups depending on their source of information: video or audio. The visual channel has information related to color, movement and texture. The features to describe the visual source can be divided into *frame-based visual features* (also called appearance features) and *spatio-temporal visual features* (also called motion features). The audio channel may contain important information about object interactions (known as acoustic concepts), environmental sounds, music or speech that can be described with *audio features*. This subsection summarizes a collection of representative features of both information sources used in multimedia tasks.

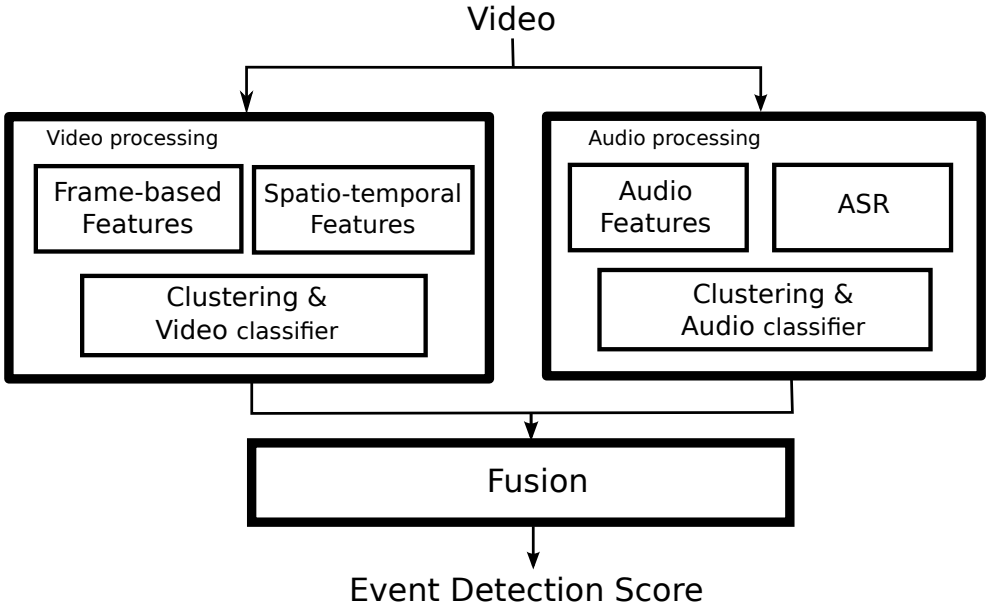


Figure 2.2: An MED generic multimodal system extracts the information from video and audio sources. The typical process presents a feature extraction block for each source, a codification of the features and a classification block. Finally, a fusion block merges the information from the video and the audio.

- **Frame-based Visual Features:**

The frame-based visual features are extracted from frame to frame without considering the temporal relations between them. The features can be computed to represent the whole frame providing a global representation of the frame (global features) or they can represent local regions to provide stable patches that can be employed to identify an event (local features).

The global features represent the distribution of colors and textures. As an example, GIST feature (Oliva2001) is one of the most popular global features for video analysis. The feature is a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represents the dominant spatial structure of a scene (spatial envelope). The feature is computed in a grid-structure to take into account the spatial distribution of the scene.

The local features locate interest points in a first step. This step is known as *interest point detection*. Once the interest points are detected, the next step is to link them in a useful way to describe the frame in an invariant manner and, therefore,

the description is robust against rotations, illumination changes or partial occlusions. Several local features can be read in the literature (Mikolajczyk2005) with different *interest-point-detection* algorithms or different approaches to describe the points. The most widely used is the *scale-invariant feature transform* (SIFT) (Lowe2004) that has recently been used in many MED systems (Jiang2012, Natarajan2012, Jiang2010a, Marin-Jimenez2013, Snoek2007) among others. The feature divides an interest point region into equal-sized grids, each described by the histogram of gradient orientations so the interest point is represented relative to its dominant orientation. Some other local features have also been recurrently used in MED context. For example, the *histogram of oriented gradients* (HOG) (Dalal) represents the edge distributions in images or video frames. *Local binary pattern* (LBP) (Ojala2002) is another texture feature that compares the value of each pixel of a frame with its neighborhood pixels.

Frame-based visual features are discriminative enough if the videos do not have rapid content changes. However these features do not model the temporal information or the motion, which are very important in video tasks.

- **Spatio-temporal Visual Features:**

Unlike frame-based features, spatio-temporal features use time as another dimension to describe the motion and this information is critical to detect the events because the motion is invariant to changes of color and lighting. The motion is often represented as a histogram/phase correlation or model parameters for global motion description.

Some spatio-temporal features have been developed following the same procedure as described for frame-based local features but extending the concepts to work in 3D (x, y, t). These features are known as *spatio-temporal descriptors*. For example, SIFT (Lowe2004) has been adapted to 3D in (Scovanner2007). Another feature to locate space-time interest points (STIPs) is described in (Laptev2005). This feature has been frequently used in MED systems. The feature computes space-time volumes in which pixel values have significant variations in both space and time. The spatio-temporal behavior of a video can also be described by tracking frame-based local features. The procedure is based on the detection of an interest point and sustains the detection of a period of time. The main disadvantage of these procedures is the expensive computational cost. A famous tracking approach can be read in (Wang2008) where the

authors extract SIFT key-point trajectories, and create a feature to be able to model the motion between every trajectory pair.

- **Audio Features:**

As we have stated in previous paragraphs, audio information is very useful especially when the video is recorded under realistic environments. There are considerable number of audio features to describe the acoustic signals in short-term windows and long-term windows and several works have summarized the most recurrent features throughout history as in (Mckinney2003, Liu1998, Peeters2004, Mporas2007, Lavner2009). Each of the audio features describes the sound in different ways but, due to the unconstrained dataset presented in MED, the sound taxonomy to describe the acoustic environment is very chaotic. For that reason it becomes very difficult to select a group of features to characterize the audio in a general context (as in MED) without the undesirable phenomena of the "curse of dimensionality". Moreover, the authors in (Eronen2006) evaluated a set of audio features using realistic audio context and the conclusion was that the Mel-frequency cepstrum coefficients (MFCC) perform better than the rest of the features.

Therefore, there are not so many differences between the audio feature extraction methods in MED. The most popular audio features are the MFCCs. A complete description with a comparison of different implementations of MFCCs can be found in (Zheng2001). MFCCs will be described in 2.3.3.1 as part of the commonly used audio features in speech and audio technologies. An important number of MED systems are using MFCCs as the main audio features. In (Baillie2003), the authors propose an audio-based MED system to detect events in sports broadcast data. In (Jiang2010a), the MFCCs are computed to detect events in TRECVid 2010 database and, more recently, the systems presented in (Myers2013, Oh2013, Jhuo2013) merge MFCCs with visual features to detect events in TRECVid 2011.

2.2.1.2 Bag-of-Words

The features presented in the last subsection (2.2.1.1) have different dimensionality and size due to their dependence on the video content, complexity or duration which increases the complexity to compare similarities. One solution for that is to segment

the information and directly match the local features, but this approach becomes computationally expensive.

A well-known approach to compare elements is *bag-of-words* (BoW) that was adapted from text retrieval to video processing in (Sivic2003) by treating features as words. In text retrieval, a bag-of-words is a vector representing the distribution of words over the vocabulary. The adaptation for multimedia can be seen as the distribution of features over an "audio/visual vocabulary" (also called codebook) that is generated by clustering features and treating each cluster as a "word".

The implementation of BoW has been performed in different ways, which has been proposed in several works. A quite famous variation of BoW was proposed in (Jiang2007) where the authors proposed a soft-weighting scheme to reduce the quantization error due to the hard assignment of each segment to the clusters. Other works like (Philbin2008) or (VanGemert2010) propose other approaches to alleviate the effect of quantization loss by soft assignment or using kernel codebooks, respectively.

The BoW is known as bag-of-audio-words (BoAW) or bag-of-frames in speech technologies and it has been recently proposed by Liu in (Liu2010) to detect video copies by the audio information on TRECVid 2009. The process is similar to the process used in video but with some differences due to the information source. For example, the video scene persists a few seconds but the audio signal changes very quickly and it is considered stable only for a few milliseconds. Therefore, firstly the audio features are segmented into frames and each frame is considered an audio word. The number of codewords is determined by a clustering algorithm. This parameter is known as the codebook size, and its value is a trade-off between system effectiveness and efficiency. If the codebook size is very small, the computational time to generate the codewords decreases and the codebook becomes more general. On the contrary, a large codebook increases the computational cost but it is more discriminative. Finally, the audio is represented by the histogram of the codebook (which is known as the bag) where each element represents the count of occurrences of a word. Histogram normalization is a common step because the audio signals vary in length. Figure 2.3 shows a general framework of BoAW representation from the audio signal extraction to the histogram with a codebook size of 4 words.

This representation has been used in several works for MED, where the implementation choices slightly differ like in (Mertens2011, Natarajan2012, Pancoast2012,

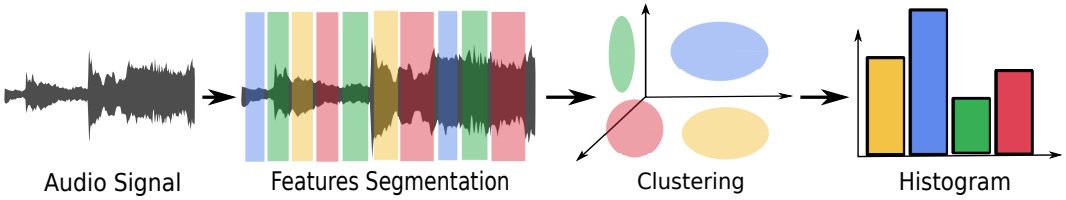


Figure 2.3: Bag-Of-Audio-Words system: after the feature extraction, the vectors are segmented into frames. These frames are clustered and the clusters are represented by histograms.

VanHout2013, Myers2013) among others. Some systems slightly modify the representation of the bag-of-words like in (Lee2010, Aucouturier2007) where the authors used Gaussian Mixture Models (GMM) to represent the bag of words or in (Lu2008) where words were generated using spectral clustering.

Once the event is represented by the features, a set of annotated events are given to train a model for the classification step like a typical machine learning process. Applying the "No free lunch" theorem (Wolpert1997) to classifiers we can conclude that there is not one classifier that outperforms all the others. Depending on the data, the task or the final application, a classifier may be beaten by another classification approach and the classifier selection is given for the interpretability, the easy-to-use algorithm and the performance. The end of this chapter summarizes the most widely used classification algorithms in this field.

2.2.1.3 Fusion

Fusing multiple information sources is generally useful because the video is analyzed from different aspects and thus they may complement each other. In other words, audio and visual features are not independent except in some cases where the users edit the original video with soundtrack or audio effects. In these cases, the original audio channel can be replaced by an entirely different audio content or the original audio can be overlapped with other audio. In the last case, the audio still contains some useful information but it is very difficult to process. On the other hand, if the original audio has not been modified, the co-occurrence or the correlation between audio and video can be exploited to perform a better multimedia analysis.

The fusion can be done in two ways, which are known as *early fusion* and *late fusion*. In *early fusion*, the feature vectors from different sources are concatenated into a long

vector. For example, in (Beal2003), the authors combine audio and visual variables to be used in graphical models. This method was developed for a constrained environment and may not be applicable to MED where the videos are unconstrained. In (Jiang2009), Jiang et al. proposed an approach to create audio-visual joint representation called audio-visual atom (AVA). An AVA is an image region trajectory associated with both regional visual features and audio features. This approach differs from the simple concatenation of audio-visual features but it can be considered *early fusion* because the unconstrained videos are represented by a bag of AVA before the classification. The approach was extended in (Jiang2011a) with an audio-visual grouplet (AVG). The sets of audio and visual codewords are grouped together if there is a temporal correlation between them. These approaches are computationally expensive because a foreground/background separation is required. An alternative for them was proposed by Jhuo in (Jhuo2013) where a bi-modal audio-visual codewords are generated using a normalized cut on a bipartite graph which captures the co-occurrence relations between audio and visual words.

Early fusion may amplify "the curse of dimensionality" (Bellman1957) problem and, therefore, *late fusion* has been more frequently used for MED. As an example, in (Jiang2010) and (Inoue2011), SIFT, STIP, and MFCC features were lately fused. Equal weights (average fusion) were used in (Jiang2011b) and (Natarajan2011). There are several late fusion methods involved with a weighted average of scores from the individual classifiers. The choice of a specific fusion method depends on the application problem. A good compilation of fusion models is described and evaluated in (Myers2013). The conclusion was that the simplest fusion methods were very effective compared with more complex fusion methods.

2.2.2 Audio Processing in MED task

As we have stated in the previous sections, the audio component of videos and the potential contribution of audio content analysis is critical in a MED task due to the unconstrained domain of the video.

The classification of the audio features (low-level features) described in 2.2.1.1 can not provide the structure or the semantic understanding present in a complex event. For example, the event "landing a fish" may have a sequence of semantic sounds like "wind blowing", "water splashing", some "laughter", etc. Therefore, a complex event

can be represented by a sequence of semantic units of sounds that we know as *acoustic concepts*. These concepts can characterize a scene (beach, park, etc.), a moving object (chair moving, a car, etc.) or a certain audio sound (metallic noise, cheering, etc). In contrast to the direct classification approaches of low-level features, the concept-base classification improves the detection of complex events. In addition, the event is divided into semantically meaningful entities where the variability of the low-level features is lower. (Haitsma2002)

Audio concept extraction approaches explored under different multimedia retrieval and content analysis projects, such as *multimedia event detection* (MED), can be grouped into two categories: (1) unsupervised and (2) supervised approaches from the perspective of modeling acoustic concepts. In the first group, one popular unsupervised approach is the Bag-of-Audio-Words (BoAW) method which has been described in 2.2.1.2. In this approach, all frame-level features are clustered via vector quantization (VQ), and then VQ indices are used as features within a classifier to model audio content ((Pancoast2012, Li2012, Pancoast2013)). Other unsupervised approaches focus on segmenting the audio track, and clustering the segments to form atomic sound units and then word-like units (Byun2012, Chaudhuri2012), or on modeling the segments with total variability factors (Zhuang2012) or GMM super-vectors (Mertens2011), which are methods borrowed from speaker identification.

In the second group of approaches, audio concept/event models are trained using annotated data. These approaches have been successfully applied for other tasks. For example, in (Haitsma2002), the authors propose an audio retrieval system based on a robust audio fingerprint approach. However, the audio fingerprint is sensitive to noise and, therefore, it may not be suitable for representing concepts. In (Reed2009), the authors propose an approach based on music segment models for music genre classification. In (Tsao2010), the authors propose acoustic segment models to describe the temporal information between units for speaker recognition. In multimedia, the audio information has been modeled recently with these techniques. For example, in (Jiang2010), authors model acoustic concepts by training Support Vector Machines (SVM) on 10-sec audio segments, which are annotated with generic concept labels (e.g., indoor vs. outdoor), and they use detected acoustic concept labels as features for multimedia event detection task. In (Pancoast2012a), fixed-duration segments are represented with segmental-GMM vectors where each element in the vector is a GMM score

calculated from a pre-trained GMM that corresponds to an annotated concept label. More recently, Oh proposed HMM models to capture the diverse temporal structure of audio concepts in (Oh2013).

Speech is another important source of audio information in MED and automatic speech recognition (ASR) can be considered as an audio supervised approach where the words play the role of the annotated concepts. ASR started becoming an important part of multimedia projects since spoken content provides information that is both discriminative and complementary to video imagery features. Before MED, speech recognition was applied to constrained domains where the audio was fairly homogeneous in terms of acoustic conditions. Speech information, however, is very difficult to extract since the speech comes from several speakers in noisy environments. In addition, speech may be overlapped with other sources and it is generated in a very natural way. All these drawbacks increase the difficulty of training suitable and robust models to transcribe speech. Segmentation and classification systems greatly help with the speech recognition. For example, segmentation would allow searching for words spoken by a speaker or aiding speaker adaptation techniques for the speech recognition system. Sources may also be non-speech concepts like music, where segmentation could help find a structure or be used by speech recognition systems to skip sections for faster processing. Despite the difficulty, any properly recognized speech can be critical to detect a multimedia event. An early work can be seen in (Chang1996) where the authors studied the importance of speech understanding and the conclusion was that speech is even more useful than video analysis for sport event recognition. ASR has been extensively used in TRECVid evaluations because it may provide extra information about some events. Natarajan et al. (Natarajan2011) found ASR helpful for a few events (e.g., narrative explanations in procedural videos) but not for general videos in TRECVid 2011. Some systems employ a combination of different approaches like in (VanHout2013), where authors combine automatic speech recognition with broad-class acoustic concepts.

Although unsupervised approaches have the advantage of not requiring labeled acoustic event/concept data, these approaches do not present semantic labels to allow for semantic searches. This is an important aspect for tasks such as multimedia event detection when the number of examples for multimedia event types becomes quite small. Therefore supervised acoustic concept detectors are useful to tackle this problem because these approaches provided detailed information about the event. On

the other hand, an inaccurate detection of the concepts could worsen the detection of the multimedia event. An essential process to make the supervised approaches more robust is audio segmentation and classification, which identifies the segments of speech or acoustic concepts to characterize the multimedia events and, therefore, it will be analyzed in depth in this thesis.

2.3 Audio Segmentation and Classification

Audio segmentation and classification is the process of delimiting an input audio stream into temporal regions with predefined information to identify specific sources in time and class. These sources can include speakers, music, background noises or other characteristics.

As mentioned above, this is useful to search for and index audio archives and it is very important in the characterization of multimedia events through audio. As an example, an audio segmentation and classification for MED can detect the audio streams containing speech to extract the words being spoken, or meta-data like acoustic concepts that provide context and information beyond the words.

There are four application domains for audio segmentation and classification research, which can be sorted from highest to lowest entropy: multimedia sharing websites, broadcast news domain, recorded meetings and telephone conversations. The data from each domain differ in the amount of audio sources, the quality of sound, the number of environments, the number of speakers and the style of speech. Even if the challenges are different for each domain, the system techniques tend to generalize well and this permits choosing the domain to study a specific technique in more controlled domains. For example, telephone conversations are suitable to study the speaker diarization of two speakers like in (Vaquero-AvilesCasco2011) while recorded meetings are suitable to study the detection of acoustic concepts like in (Temko2009a, Butko2011c). The most relevant information about the audio segmentation and classification approaches and their applications can be found in (Reynolds2005).

This section summarizes relevant works in two relevant areas. The first one is audio segmentation and classification for the broadcast news domain where the main goal is to detect segments of speech (pure speech, speech with music, speech with noise, etc...) versus non-speech (music, noise acoustic concepts or silence). The second one is

to detect acoustic concepts in meeting rooms where the presence of non-speech sounds can help to describe the scene.

2.3.1 Speech/Non-speech Classification

Many different approaches have been proposed for speech/non-speech audio classification because it can be useful in many ways. The detection of speech segments may help speech recognition systems aiding speaker adaptation techniques if the speech is overlapped with noise or music. On the other hand, non-speech segments (most of them comprised of music, noise and silence) can provide a document structure or enrich the transcription of the documents with meta-data. In (Lavner2009), the authors present a review of several solutions for speech/music classification that make use of different acoustic features and classification methods. The studies that can be found in the literature focus on either the feature extraction method or on segmentation/classification strategies.

Audio segmentation/classification systems can be divided into two different groups depending on how the segmentation is performed. The first group detects the boundaries in a first step and then classifies each delimited segment in a second step. We refer to them as *classification-after-segmentation* approaches. For example, in (Nguyen2011), an approach using a temporally weighted fuzzy C-means algorithm was proposed. The *Bayesian Information Criterion* (BIC) is used extensively in many studies, like that of (Chen1997) to generate a break-point for every speaker change and environment/channel condition change in BN domain. The main idea is to compute the distance between two continuous segments to determine if there is a change between them. (Wu2006a) and (Kotti2008) also utilize BIC to identify mixed-language speech and speaker change respectively. However, BIC has several shortcomings that have to be considered. It can only set one break-point for each analysis window, so a small window involves more precision but the Gaussian estimation may be inaccurate due to the scarcity of data. Although BIC is the most popular segmentation strategy, other approaches have been proposed in the literature with different distance metrics. For example, Generalized Likelihood Ratio (GLR) was described in (Willsky1976) and it is obtained as a likelihood ratio between the likelihood of the assumption that both segments belong to the same class and the likelihood of each segment belongs to different classes. In (Siegler1997), a Kullback-Leibler Divergence (KL) was used for acoustic

segmentation and speaker segmentation in broadcast news environments. In (Wu2006) the authors propose a *Minimum Description Length* (MDL) approach that permits multiple breakpoints for any generic data.

The second group is known as *segmentation-by-classification* and consists of classifying consecutive fixed-length audio segments. The segmentation is produced directly by the classifier as a sequence of decisions. This sequence is usually smoothed to improve the segmentation performance. An example of this procedure can be found in (Misra2012) where the author combines different features with a Gaussian Mixture Model (GMM) and a maximum entropy classifier. The final decisions were smoothed with a Hidden Markov Model (HMM) to avoid sudden changes. In (Lu2003), an audio stream is segmented by classifying each window into five broad-classes. The solution combines *Support Vector Machines* (SVM) and evaluates the classification over some new proposed features.

The different strategies outlined in the preceding paragraphs have their advantages and disadvantages. Long-time features (*segment-based* features) are not suitable for training statistical models (Huang2006) in a *segmentation-by-classification* strategy. However, they provide great discriminative power for audio classification in *classification-after-segmentation* systems (Lu2002). On the other hand, short-time features (*frame-based* features) allow statistical models to make decisions over short-duration windows in *segmentation-by-classification* strategies (Foote1997) but they are usually less discriminative for audio classification since they were mainly designed for speech related tasks such as *Automatic Speech Recognition* (ASR) (Huang2006). The most common solution to avoid the shortcomings and enjoy the benefits of each strategy is to create hierarchical systems with multiple steps where each level is designed with class specific features and segmentation systems as in (Gallardo2010) and (Castan2011). Nevertheless, these systems become very specific for the intended task and are quite difficult to adapt for other databases.

2.3.2 Acoustic Concept Recognition

The Acoustic Concept Recognition (ACR) (also known as Acoustic Event Detection in the literature) is the task entailing the identification of any class of sounds which is caused by different sources and their temporal position. The classes are not related to speech but they can be produced by humans and they are meaningful.

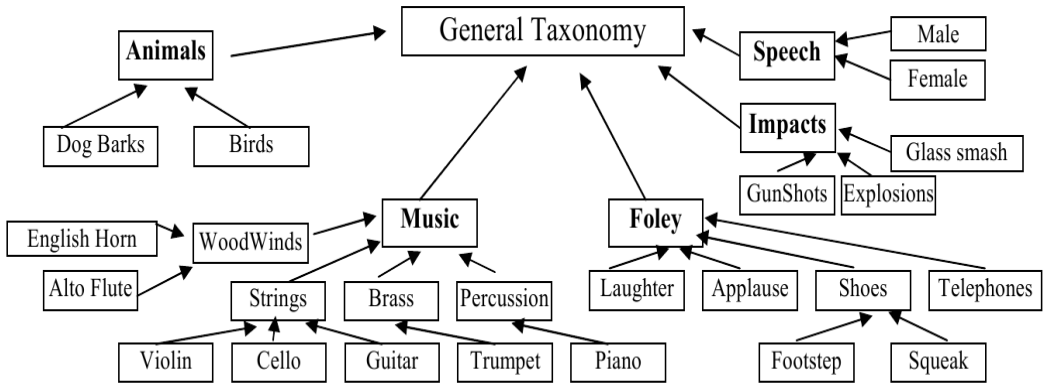


Figure 2.4: Taxonomy proposed by (Casey2002) suitable for MED.

The ACR has been implemented in hospitals (Vacher2003), kitchens or restaurants (Lukowicz2003), public places (Lee2009) and even in bathrooms (Chen2005) to give a few examples. The applications for ACR are very diverse but they include following major areas: audio indexing and retrieval in multimedia documents, meetings description and surveillance. The taxonomy[1] of the sound for each application or area is very different. As we have already shown in the last subsections of this chapter, the MED task presents a very unconstrained taxonomy and the detection of concepts in this area becomes very difficult. For example, Figure 2.4 shows a larger classification scheme including animal sounds, musical instruments, people and sound effects (Foley) proposed by (Casey2002) that could be considered suitable for any multimedia application due to its wide structure. Recently, researchers are making significant efforts to provide support systems to identify multimedia events by acoustic concepts recognition (Pancoast2012a, Castan2013b, Castan2013, Elizalde2013). However, the recognition of the concepts on this wild environment is far from optimal.

A selection of a more constrained taxonomy can be useful to develop technology because the number of the concepts is limited and with reduced variability. Figure 2.5 proposes a taxonomy suitable for acoustic concepts detection in meeting rooms. This taxonomy was described by Temko in (Temko2009a) for the CHIL project (Temko2006). The project deals with the detection of acoustic concepts produced in meeting room environments to describe the human and social activity in the room.

[1]The taxonomy of the sound shows the audio structure by different semantic levels

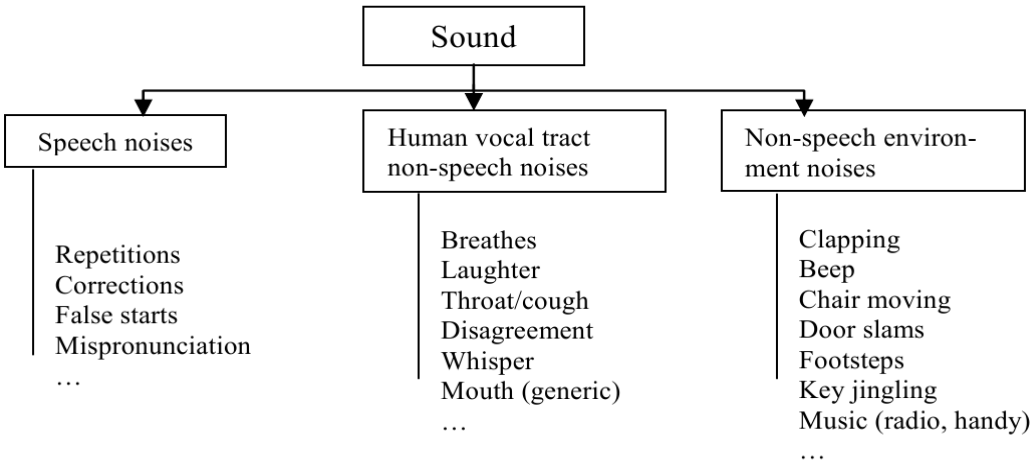


Figure 2.5: Taxonomy proposed by (Temko2009a) for acoustic concepts recognition in meeting rooms.

Under the framework of the CHIL project, the CLEAR 2007 Evaluation campaign (Stiefelhagen2007) was proposed to evaluate systems for the perception, activities and interaction of people in meetings. Two important conclusions were drawn from this evaluation: first, the performance of the classification of single acoustic concept is relatively high. As an example, (Temko2009) and (Zieger2008a) developed an SVM based system and an HMM-based system, respectively, to classify different acoustic sounds (e.g., steps, door slams, or paper noise) in the meeting room environment using the CHIL-2007 database in which the acoustic concepts are isolated and recorded in a controlled environment (Mostefa2008a).

Secondly, the time overlapping of acoustic concepts with speech or other acoustic concepts has not yet been solved. This situation produces the major source of detection errors (70%) and, therefore, it is still under study. The winner of the evaluation (Zhou2008) proposed a system based on Kullback-Leibler distance with an AdaBoost to select a set of discriminative features to identify the segments of speech and the acoustic concepts because they have notable spectral differences. But in spite of this intelligent system, the results are far from satisfactory.

Recently, researches have tried to approach the meeting-room overlapping concept problem. The approaches proposed merge the information from different sources. For example, in (Butko2011c) the information coming from video helps to detect the acoustic concept based on the position of every person in the scene. In (Chakraborty2013),

the overlapping problem is tackled by exploiting the signal diversity that results from the use of multiple array beamformers. Both approaches improve the detection of overlapped acoustic concepts but, on the other hand, the systems are not as general as desired because a clear video information or multiple microphones are required.

The ACR is quite important in the field of applications for monitoring and surveillance. Some of these applications are developed to assist disabled people because image-based systems could not be applied due to privacy reasons. For example, Chen et al. (Chen2005) proposed an automated bathroom monitoring system to detect the activity of elderly people based on acoustic concepts. In surveillance, most of the systems detect abnormal situations by the analysis of visual clues, but some specific situations are easier to detect by audio clues. For example, in (Clavel2005) the authors developed an automatic gun shot detection system. In (Atrey2006), the authors proposed an approach for detecting various normal and excited state human activities with four different audio features and Gaussian Mixtures Models.

2.3.3 Audio Segmentation and Classification Technology

Segmentation and classification systems in the literature are characterized by the set of features and classifiers employed for each approach. This section shows a summary of the most widely used technology for audio segmentation and classification. First, a compilation of the common features is described in section 2.3.3.1. Then, section 2.3.3.2 shows some common approaches to model the statistical behavior of the features for each class.

2.3.3.1 Audio Features Extraction

Here, we summarize the most typical features for speech and audio recognition technologies:

- **Mel-frequency cepstrum coefficients (MFCC):** Representation of the relationship between short-term power spectrum sub-bands equally spaced on the mel scale of frequency which approximates the human auditory system's response more closely than linearly-spaced frequency bands. Perceptual Linear Prediction (PLP) features are another popular acoustic features with a lot of similarities

with MFCC. The main differences lie in the filter-banks and in the application of a linear prediction among others.

- **Frequency Filtering (FF):** These features can be seen as another technique to represent the spectral envelope, such as the MFCC, but with less computational cost and with a clearer idea about the behavior of the frequencies in the final result. It consists of filtering the frequency sequence of filter-bank energies with a filter that equalizes the variance of the cepstral coefficients (Nadeu2001).

- **Zero Crossing Rate (ZCR):** Defined as the number of zero crossing in a frame. In other words, ZCR is the average number of times the signal change its sign within a frame. There are many variations of this feature like the High Zero-Crossing Rate Ratio (HZZCR) which is defined as the ratio of the number of frames whose ZCR is above 1.5 times the average zero-crossing rate.

- **Short-time Energy (STE):** Representation of the amplitude variation over time. This feature also has some variations like the Low Short-Time Energy Ratio (LSTER) that is defined as the ratio of the number of frames whose short-time energy is less than 0.5 times the average STE.

- **Spectral Flux (SF):** This feature provides an idea about the changes that occur in the shape of the spectrum frame by frame. It is typical to compute the variation of the spectral flux (VSF) over consecutive frames.

- **Amplitude Modulation Ratio (AMR):** Relationship between local minima and local maxima in the envelope signal. The envelope can be obtained by filtering the signal with a lowpass filter. The alternation of high energy and low energy segments (vowels and consonants) in a speech signal causes the amplitude modulation ratio to be higher for speech and lower for music signals.

- **Chroma:** Representation of the energy spectrum onto 12 bins representing the 12 distinct semitones of the musical scale. This feature has been widely used in music recognition applications.

The choice of proper features is a crucial step in the design of automatic audio recognition systems and it impacts directly on the robustness of the system. Many types of

audio features has been proposed for different tasks of sound description, like speech recognition (Yapanel2008, Gonzalez2010, Sarikaya2000, Mporas2007), speaker verification (Barras2003), multimedia content analysis (Liu1998), audio segmentation and classification (Mierswa2005, Mckinney2003) and emotion identification (Luengo2010) among others. However, a proper set of features for a system does not necessarily work well in other systems. For example, the *Mel Frequency Cepstrum Coefficients* (MFCC) or *Perceptual Linear Prediction* (PLP) features become the standard front-end in many speech applications but audio segmentation and classification in general is not so clear yet because the authors mix the MFCC with other features very often. Therefore, researchers have not clear arguments in favor of a particular set of features and the final decision about feature selection is mainly based on their prior knowledge.

Features can be grouped in several ways. Some authors group the features by domain: *time-domain* and *frequency-domain* like in (Chu2009). *Time-domain* features are computed directly over the waveform and characterize the temporal behavior of the audio signal. *Frequency-domain* features are computed over the spectrum of a segment and describe the distribution of the signal in frequency bands.

The acoustic features can also be classified into *frame-based* and *segment-based* according to the portion of time analyzed. *Frame-based* features are extracted within short periods of time (between 10 and 30 ms) and are commonly used in speech related tasks where the signal can be considered stationary over that short period. MFCC or PLP are generally used as *frame-based* features as proposed in (Imai1983, Vergin1996, Vergin1999, Wong2001, Hasan2004), and more recently in (Dhanalakshmi2011) where these features are classified with an autoassociative neural network. *Frame-based* features have been also proposed for audio segmentation and classification into broad-classes of broadcast news. Among others, in (Xie2010) two pitch-density-based features are proposed, in (Saunders1996, Li2001, Lu2002) the authors use *Short-Time Energy* (STE) and *Harmonic features* are used in (Nwe2005, Hauptmann2003, Dhara1999). *Frame-based* features can be used directly in the classifier. However, some classes are better described by the statistics computed over consecutive frames (from 0.5 to 5 second long). These characteristics are referred to in the literature as *segment-based* features (Gallardo-Antolin2010, Butko2011). For example, in (Markaki2011) a content-based speech discrimination algorithm is designed to exploit the long-term information

inherent in modulation spectrum. In (Huang2006) the authors propose two segment-based features: the *variance of the spectrum flux* (VSF) and the *variance of the zero crossing rate* (VZCR).

2.3.3.2 Statistical Modeling

This subsection shows a review of the most widely used classification algorithms in this field.

- **Support Vector Machines (SVM):**

Kernel classifiers have become very popular for multimedia applications and, among all kernel-based classifiers, the Support Vector Machine (SVM) is the most common algorithm due to its balanced performance across different tasks and its efficiency in high dimensional problems especially in non-linear classification (Scholkopf2002). The data from the classes are mapped into a higher-dimensional space in a first step. This procedure is known as the kernel trick. In the new high-dimensional space, the data can be classified using linear discriminant functions (also called hyperplane) because the classes are divided by a gap that is defined by the closest training data to the decision surface. The test data are then mapped into the same high-dimensional space and the belonging class is determined by the side of the decision surface where they are located.

Let's assume the typical MED task where two classes, target event (+1) and non-target event (-1), are defined. The decision function for a feature vector $\boldsymbol{x}$ of a test video has the form:

$$f(\boldsymbol{x}) = \sum_i \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) - b, \tag{2.1}$$

where $f(\boldsymbol{x})$ represents the distance (in general sense) of the feature vector from the hyperplane, b is the threshold parameter, $\boldsymbol{x}_i$ are the support vectors, α_i is the support vector weight and y_i is the corresponding label such that $\sum_i \alpha_i y_i = 0$ and $\alpha_i > 0$. The kernel function $K(.,.)$ is the inner product between feature vectors:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x})^T \Phi(\boldsymbol{y}), \tag{2.2}$$

where $\Phi : R^N \mapsto R^M$ represents the mapping from input space into a higher-dimensional space where the classes can be linearly separated by an hyperplane. In the training

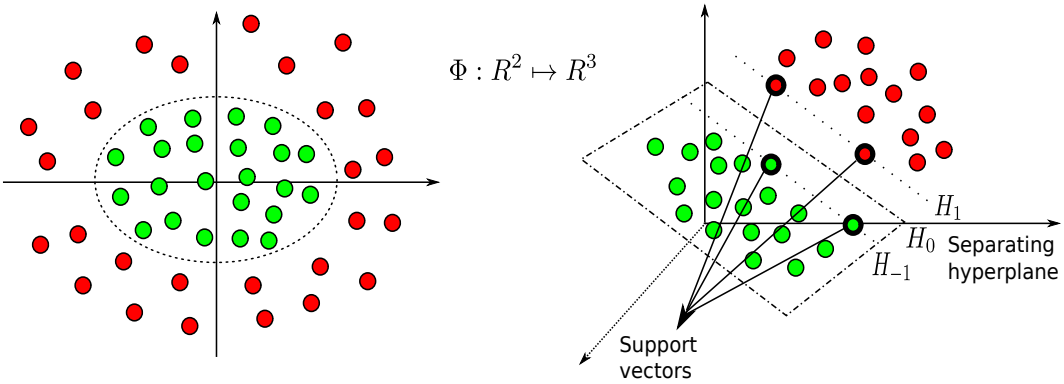


Figure 2.6: Two-class classification example with Support Vector Machine. The original data (2-dim) are mapped into a high-dimensional space (3-dim). The support vectors are highlighted in the figure.

phase, the algorithm searches for the hyperplane that maximizes the margin between target and non-target events in the high dimensionality space. The training events closer to the boundaries are known as *support vectors*.

Choosing a suitable kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ is critical to perform a good classification. The optimal kernel depends on the task and the nature of the features. Possible choices of kernel functions include the polynomial kernel, the Gaussian radial basis function (RBF), the histogram intersection kernel or the multilayer perception kernel. Among them, the polynomial kernels and especially the RBF kernel are widely used in the literature. The polynomial kernel is defined as:

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y} + c)^d, \tag{2.3}$$

where the parameter d is the degree of the polynomial. If $d = 1$ the kernel is known as linear kernel. The Gaussian RBF is defined as:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{y})^2}{2\sigma^2}\right), \tag{2.4}$$

where the parameter σ is the width of the Gaussian function. Gaussian kernel has been used recently in event recognition systems with good performance (Jiang2010, Natarajan2011, Myers2013, Jhuo2013).

The features are usually accumulated into a single vector to be used in the SVM classifier. While this approach seems reasonable, the temporal information is not described and, therefore, other approaches are implemented as a middle step to model

the temporal behavior of the events with graphical models. The graphical model approaches combine the likelihood of the data and the graph theory to find structures in sequential data. A popular graphical model method is the Hidden Markov Model which will be described deeply below.

- **Gaussian Mixture Models (GMM):**

A Gaussian Mixture Model (GMM) is a mixture distribution that represents the probability distribution as a weighted sum of Gaussian component densities. The GMM were proposed by Liporace (Liporace1982) and introduced in speech technologies by Juang (Juang1985) to the problem of speech recognition of isolated digits. It is an efficient way of modelling multimodal distributions which is the case in speech processing and audio segmentation and classification.

For a feature vector $\boldsymbol{x}$ the GMM distribution is defined by

$$P(\boldsymbol{x}|\lambda) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k), \tag{2.5}$$

where K is the number of Gaussian components in the mixture and $\lambda = \{\omega, \mu, \Sigma\}$ are the model parameters collectively represented, ω_k, μ_k and Σ_k are the weight, mean and covariance matrix associated with component k, the weights satisfy the constraints $\omega_k \geq 0$ and $\sum_{k=1}^{K} \omega_k = 1$, and the Gaussian distribution is defined as

$$\mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|} e^{-\frac{1}{2}(\boldsymbol{x}-\mu_k)^T \Sigma_k^{-1}(\boldsymbol{x}-\mu_k)}. \tag{2.6}$$

The most popular method for training GMM parameters is the maximum likelihood (ML) estimation. This estimation technique tries to find the parameters λ of the model that give the maximum log-likelihood [1]. The logarithm of function 2.5 for a dataset $\mathbf{X}$ of N points is given by

$$\ln P(\mathbf{X}|\lambda) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \omega_k \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)\}, \tag{2.7}$$

To obtain the maximum likelihood we derive the log-likelihood function with respect to the parameters but, in the case of the GMM, the derivation is complex. The

[1] The logarithm is used to reduce the numerical dynamic range because the product of very low probabilities can underflow the numerical precision of the computer, and the computation of sums of log probabilities instead of product solves this problem

most commonly used alternative ML approach is the expectation-maximization (EM) algorithm than can followed in (Bishop2006).

In the verification phase for audio segmentation and classification, the log-likelihood of a test utterance for each GMM (one model for each class) is computed. The class whose GMM gives the highest likelihood is considered the correct one. Another approach to verify the classes in the test phase is the log-likelihood ratio (LLR) between a specific class model (called target class) and a universal model (known as *universal background model* (UBM)) which represents the class independent distribution of features. The target class model is usually estimated by adapting the means from the UBM using a maximum a posteriori (MAP) criterion and the LLR is computed as:

$$LLR = \frac{1}{T}\sum_{t=1}^{T} lnP(\boldsymbol{x}_t|\lambda_{target}) - lnP(\boldsymbol{x}_t|\lambda_{UBM}) \tag{2.8}$$

- **Hidden Markov Models (HMM):**

Before defining the Hidden Markov Models (HMM) we have to define the Markov chain because within an HMM lies a hidden Markov process. Therefore, a Markov chain is a sequence of states and their probability depends only on the immediately preceding state. An HMM represents stochastic sequences (like audio concepts, words or segments) as Markov chains. The states of this process are not directly observed (the states are hidden). The sequence of states can only be observed through the stochastic processes defined into each state. Therefore, the true sequence of states is hidden by a first layer of stochastic processes.

An HMM is characterized by:

- the *emission probabilities* which are given by the probability density function (pdf) that characterize each state $p(x|q_i)$ denoted by $b_i(x)$ in the literature. They are usually Gaussians or GMM but it could be any other kind of pdf.

- the *transition probabilities* which are the probability to go from state i to state j ($P(q_j|q_i)$) where the states are defined as $Q = \{q_1, ..., q_k\}$. They are stored in matrices where each term a_{ij} denotes a probability $P(q_j|q_i)$.

- the *initial state distribution* can be seen as a discrete event to model the "start" of a sequence and it is denoted as π_i.

There are multiple types of topologies for an HMM, which include:

- *Ergodic*: the transitions go from any emitting state to any other emitting state.

- *Left-to-right*: the transitions only go from one state to itself or to a unique follower.

In audio segmentation and classification, every class is modeled by an HMM trained following the Baum-Welch algorithm. The initial *emission probabilities* and *transition probabilities* are chosen randomly and the *initial state distributions* are uniformly distributed for each class. After training, we have all the parameters for each class Θ_k where k represents each acoustic classes. For each testing sequence X_T, the likelihood $P(X, \Theta_k)$ is computed for each class, and the testing sequence is classified to the class with the maximum likelihood following the Viterbi algorithm.

- **Joint Factor Analysis (JFA):**

Joint factor analysis (JFA) is a generative model to estimate the class GMM taking into account the different sources of variability. In this model the means of all the components of the GMM are concatenated to build high-dimensional supervectors and modelled as a sum of factors. As in the classical MAP of Reynolds (Reynolds2000), the most commonly used implementation adapts the means of the GMM-UBM to the class while the weights and variances are shared among all the classes. In other words, the means are not fixed and can vary from segment to segment due to several sources that increase the within-class variability (Kenny2007). The JFA for audio segmentation and classification can be written as:

$$\boldsymbol{m} = \boldsymbol{t}^c + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{U}\boldsymbol{x}, \tag{2.9}$$

where $\boldsymbol{m}$ is the supervector of means for class c, t^c is the class-location vector obtained by MAP adaptation from the UBM. The term $\boldsymbol{V}$ is known as the *eigenvoices matrix*, $\boldsymbol{y}$ as class factor vectors and they model the class variability. Usually the term $\boldsymbol{V}\boldsymbol{y}$ is not used in language ID or segmentation approaches and the variability is restricted to be modeled in a channel subspace. $\boldsymbol{U}$ is known as the *within-class variability matrix* that we use to compensate that variability of the channel and x_s is the *channel factor*, a low-dimension hidden variable whose role is to give the channel information of the utterance. While t^c is obtained with all the training data of class c (it is fixed once calculated), the channel factor is utterance-dependent, and is the term that moves the class-specific model for each utterance $\boldsymbol{s}$.

Figure 2.7 shows a simplified vector representation of Factor Analysis in three dimensions without the term $\boldsymbol{V}\boldsymbol{y}$. For simplicity we only show one component of the

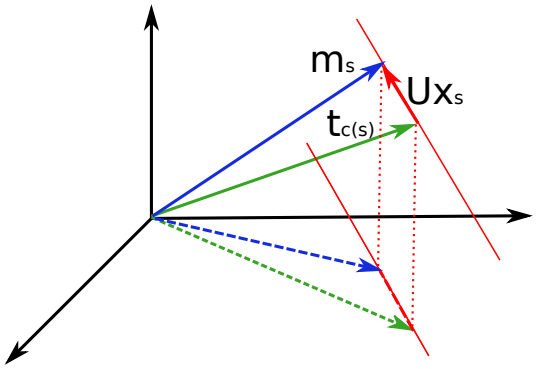


Figure 2.7: Vector representation of the within-class variability compensation

GMM. The non-compensated mean of the class comes from the UBM and is obtained via MAP adaptation to end in the corresponding mean of class c given by $\boldsymbol{t}^c$. This mean is further adapted to compensate for the channel by the within-class variability term $\boldsymbol{U}\boldsymbol{x}_s$ to determine the final vector location $\boldsymbol{m}^s$ for utterance $\boldsymbol{s}$.

The JFA model is also trained via ML, and like the GMM, there is no closed form solution for the derivatives of the log-likelihood function over the parameters. The adopted solution is again the application of the EM algorithm.

Different scoring techniques for JFA have been developed in the literature. In (Glembek2009), the most common techniques are compared in terms of performance and speed. All of them have in common that the final scores obtained for each class are given in terms of the log-likelihood ratio between the log-likelihoods given by the model of each class and the UBM. Also, the log-likelihoods are normalized by he number of frames in the utterance, being an average of the log-likelihoods of all the frames.

- **Total Variability Subspace (i-Vectors):**

In the JFA framework, the class and the channel variability are modeled separately by two different factors. In (Dehak2010), the authors find that the speaker information was not completely removed from the channel subspace in a speaker identification problem with speaker and channel subspace. Therefore, they redesign an FA model with a single factor, including both the speaker and channel information. The training procedure of this new model is the same as for JFA but, in this case, the model is class-independent. The new mean supervector of equation 2.9 becomes

$$\boldsymbol{m}^s = \boldsymbol{m}_0 + \boldsymbol{T}\boldsymbol{w}_s. \tag{2.10}$$

As we can see there is no MAP adaptation to the class and the model is centered at the UBM. The subspace is spanned by matrix T, and the point in the subspace of utterance $\boldsymbol{s}$ is given by the latent variable $\boldsymbol{w}_s$ (we just renamed U and x_s of JFA, which were defined for the channel subspace). The new defined subspace contains all the variability of the signal.

The same EM procedure defined for JFA can be used by removing the dependence on the class. In practice, we pool the files of all the classes and calculate a single JFA model as if all the utterances belonged to the same class.

The i-Vector approach presented is used as a front-end to extract fixed-length low-dimension features (known as i-Vectors in the literature). One i-Vector is obtained per utterance from the class-independent factor $\boldsymbol{w}_s$. From this point, the i-Vectors become our features, and several approaches can be used to classify each vector into classes.

- **Classification Trees:**

A classification tree is a model where the decision structure is a tree with edges and nodes. The intermediate nodes are labeled by a single attribute, and the edges extending from the intermediate node are predicates on that attribute. The leaf nodes are labeled by the predefined classes.

Figure 2.8 shows an example of a classification tree. In this specific example, the tree was trained to classify music frames and speech frames with five different features previously described (VSF, LSTER, HZCRR and Minimum Energy). The intermediate nodes are represented by purple squares and the leaf node is represented by an orange circle. As a general description, each frame is classified by checking the current test and then falling down the appropriate branch until a leaf is reached.

The classification trees have two important advantages: firstly, they can be implemented very efficiently and, secondly, they are very easy to interpret because the tree is comprised of individual classifiers for each dimension of the feature space. This easy interpretation can be seen as a feature space partition recursively as it can be read in (Breiman1984) or (Quinlan1986) where the authors describe the CART and the C4.5 trees respectively.

Instead of classifying the frames into a predefined class, it can be useful to provide class probabilities especially in problems involving noise. In these cases, each leaf node

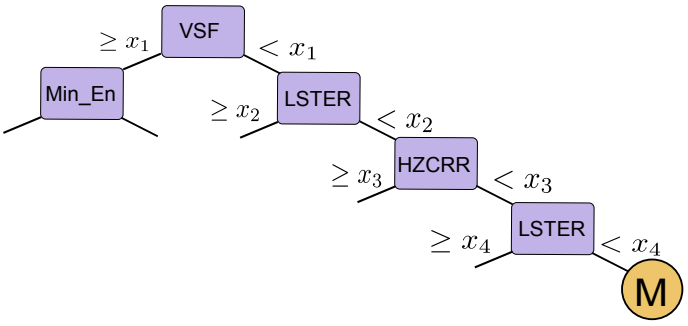


Figure 2.8: Classification tree example.The branches represent different acoustic features and the leave represents an audio class (in this exmaple, the leave is "Music").

has a vector of class probabilities. In (Buntine1992) the authors show how a tree learning algorithm can be derived using Bayesian statistics.

2.4 Chapter Summary

In this chapter we have briefly reviewed the state of the art in MED and the most relevant approaches in audio segmentation and classification. Firstly, the multimedia event detection task is described with the most widely used techniques for multimodal features extraction (image, video and audio), event characterization and fusion techniques. Secondly, the chapter has delved into the audio segmentation and classification and, more precisely, into speech/non-speech techniques and acoustic concepts recognition with the most relevant audio features and statistical models.

Chapter 3

Multimedia Event Detection

Contents

3.1 Chapter Overview

Previous chapters have summarized a set of common solutions for MED: unsupervised approaches where the events are detected by clusters of low-level features, and supervised approaches where the acoustic concepts are used as high-level features to identify the event.

This chapter examines two solutions to model multimedia audio content with a supervised acoustic concept extraction technique. Firstly, we describe baseline systems where the acoustic concepts are evaluated with fixed segmentation (Pancoast2012a) or by HMM-based *acoustic concept recognition* (ACR) (Castan2013). Secondly, we propose a novel approach where the recognition lattices of the ACR are used to extract posterior N-gram counts (Castan2013b). The N-gram counts are used as features in SVM-based classification for MED task. Given the high variability present in user-submitted Internet videos, this approach improves the MED performance based only on ACR because 1-best hard-decisions are less informative. This approach will be compared with ACR and Bag-of-Audio-Word approaches to compare the behavior with a baseline supervised approach and unsupervised approach. Finally, we propose a fusion with spoken concepts since the information captured by the spoken concepts is different to the information captured by the acoustic concepts and the systems can be combined.

The remainder of this chapter is organized as follows: the TRECVid2011 dataset is described in 3.2. Section 3.3 evaluates and describes the acoustic concepts annotations, the audio features and the acoustic concepts classification, detection and recognition. Section 3.4 deals with the two baseline systems for MED: *segmental-GMM* and *ACR*. Section 3.5 shows the lattices N-grams counts approach and its comparison with *segmental-GMM*, *ACR* and *BoAW*. A fusion system with spoken and acoustic concepts is proposed in Section 3.6 where the importance of acoustic concepts for MED is shown. Finally, Section 3.7 summarizes the most relevant aspects of this chapter and briefly points out some conclusions.

3.2 TRECVid2011 Dataset

The Text Retrieval Conferences Video Retrieval Evaluation (TRECVid) is an annual conference sponsored by the National Institute of Standards and Technology (NIST) and the goal of the conference is to encourage research in information retrieval. The

TRECVid2011 (NIST2011) focuses on the problem of Multimedia Event Detection (MED) in website quality videos for hard-to-detect events (e.g., Landing a fish). The evaluation dataset consists of non-professional videos collected from various social networks on the Internet with high variability (each video is recorded with different devices in a different acoustic context) and short duration (a couple of minutes). Fifteen different video event categories can be found in the database with only five of those categories available for testing purposes in this study.

To develop and evaluate our proposed approaches, we use three sets of data: first set (*train-1*) is for training the acoustic concept models, second set (*train-2*) is for training the MED classifiers after extracting acoustic concept indexes on this data and using them as MED features, and the third set (*test*) is for testing the system. These sets are the same sets used in (Pancoast2012) and (Pancoast2012a) to be able to provide fair comparison with previously published works. The videos are provided in MP4 format. We extract the audio components with a sampling rate of 16KHz. There is a total of 2640 videos in the test set and 7881 in the training set. Table 3.1 shows, for each of the five video events, the numbers of positive samples in the test and training sets. Note that the categories group several videos. For example "feeding an animal" includes animals from different species and therefore, different animal sounds, while "attempting a board trick" includes people skateboarding, snowboarding and surfing. The remainder of the videos in the test set are random videos that do not belong to any of the event categories.

Table 3.1: Video event class abbreviations (Abbr.) and full names along with the number of positive samples appearing in the training and test sets

Abbr.	Full Name	# Train	# Test
E001	Attempting a board trick	91	32
E002	Feeding an animal	81	30
E003	Landing a fish	69	26
E004	Wedding ceremony	66	25
E005	Woodworking project	77	25
—	None	7497	2502

Table 3.2: Broad and specific acoustic concepts proposed in (Pancoast2012a). The specific concepts are subgroups of the broad concepts.

5 Broad Concepts	20 Specific Concepts	
	Air traffic	Individual applause
	Birds	Individual yells
Crowds/audience	Crowd applause	Large crowd
Animal sounds	Crowd cheers	Scraping-Sanding
Repetitive sounds	Crowd laughter	Sewing
Machine noise	Crowd yells	Skateboard
Environmental sound	Farm animals	Small party
	Ground traffic	Water running
	Hammer	Water splashing
	Wind	Home appliances

3.3 Acoustic Concepts

This section presents the acoustic concept annotations and the results with respect to the classification and recognition of the concepts. The section can be seen as a preliminary study to show the difficulty of creating a well-trained model for these acoustic concepts due to the high variability of the audio. This section is organized as follows: firstly, two sets of concepts annotations are presented in section 3.3.1. Secondly, we describe the front-end audio features used in this approach in section 3.3.2. Finally, the experiments of classification and segmentation of the acoustic concepts are reported in sections 3.3.3 and 3.3.4 respectively.

3.3.1 Acoustic Concepts Annotations

Because the ultimate goal of the system is to detect multimedia events on the videos using acoustic concept recognition, an initial set of labels of acoustic concepts has been created that will be useful in discriminating the five video event classes presented in Table 3.1 while also being clear and understandable for the annotators.

The acoustic concepts are divided into five broad classes as Table 3.2 shows. These classes can be extended with more specific acoustic concepts. In (Pancoast2012a) the five broad classes were extended to twenty specific classes as can be seen in Table 3.2.

These classes have been extended with Speech and Music classes because most of the videos contain speech or music as the predominant audio. In fact, some of the acoustic concepts are overlapped with speech or music that is barely audible in the background. However, those segments were annotated as that acoustic concept.

3.3.2 Front-End Audio Features

This section is a summary of the front-end audio feature extraction method used in (Castan2011). We extract 16 MFCCs (including C0) computed in a 25ms frame length with a 10ms frame step and their first and second derivatives. Due to the high variability of every acoustic concept, the fact that the segments are overlapped with speech and music, and the different devices used to record the video, a normalization of these features is needed. In an attempt to generalize the features, a cepstral mean normalization is computed over the whole video and the mean and standard deviation are computed over 1-second windows with an overlap of 0.75 seconds. Thus, the system uses 96 features (48 for the mean and 48 for the standard deviation of the $MFCC + \Delta + \Delta\Delta$ features) every 0.25 seconds.

3.3.3 Acoustic Concept Classification Experiments

To model the acoustic concepts we used an HMM/GMM-based system. As described in the previous section, to train and test these models, a subset of the National Institute of Standards and Technology (NIST) is provided for the TRECVid2011 evaluation. This set is comprised of 1536 videos (47 hours approximately) with an average length of 1.8 minutes per file.

The goal of this experiment is to classify a set of segments extracted with the ground truth boundaries in one of the broad classes. The segments are overlapped with speech and music in the background in some cases. However, the classification is done with the five broad classes (without speech and music models) keeping the seven broad classes (with speech and music models) for the recognition task. The segments are extracted from the video database generating 13,520 segments of different durations. Each concept is modeled as one state HMM/GMM with 256 Gaussians. Table 3.3 shows the results of a first approximation experiment using the same subset of data to train and test. As it can be seen, the task is very difficult due to the high within-class variability of each concept. The system classified 71.1% of the segments correctly.

Table 3.3: Confusion matrix of a first approximation classification experiment using the same set of data for training and testing. Each row of the matrix represents the ratio of segments in an actual class and each column represents the ratio of segments in a predicted class.

	CA	AN	RS	MN	ES
CA	**0.77**	0.04	0.04	0.06	0.09
AN	0.08	**0.80**	0.04	0.02	0.06
RS	0.08	0.04	**0.75**	0.05	0.08
MN	0.11	0.04	0.09	**0.61**	0.16
ES	0.12	0.09	0.08	0.07	**0.63**

To test the system, a 4-fold cross-validation was performed using 3 quarters to train the models and 1 quarter to test. Table 3.4 shows the confusion matrix and how the classification rate is reduced compared with Table 3.3 getting 45.9% of the segments correctly classified. It can be seen that "Animal Noise" and "Environmental Sounds" are the concepts with the highest error rate. This can be easily explained by two major aspects: firstly, neither of the two classes have enough data to train the models and secondly, the audio of these concepts presents low energy levels.

Table 3.4: 4 Folds cross-validation confusion matrix for acoustic concept classification. Each row of the matrix represents the ratio of segments in an actual class and each column represents the ratio of segments in a predicted class.

	CA	AN	RS	MN	ES
CA	**0.61**	0.03	0.06	0.12	0.18
AN	0.18	**0.12**	0.20	0.19	0.31
RS	0.11	0.05	**0.45**	0.18	0.21
MN	0.17	0.02	0.16	**0.40**	0.25
ES	0.24	0.07	0.15	0.21	**0.33**

3.3.4 Acoustic Concept Recognition Experiments

In the MED task, a recognition system is needed to be able to detect and classify the acoustic concepts related to the video. Due to the fact that most of the acoustic concepts are overlapped with speech and music, two extra models are required to identify the segments where there is no acoustic concept and to be able to produce a clear

segmentation. In addition, these models can be useful to describe the video in the MED task. Using the same models trained for the classification task, a segmentation is executed over the whole video where the segments were extracted for the classification and detection system.

In this experiment, every concept (speech and music included) is modeled by an HMM/GMM of one state. The main difference is that a segmentation is produced when there are transitions between the models in the Viterbi algorithm. Table 3.5 shows the recognition result per concept independently of the segment duration. As it can be seen, "Crowds" and "Repetitive Sounds" have the better results in comparison with "Animal Noise" or "Environmental Sound" because "Crowd" and "Repetitive Sounds" were trained with more data than "Animal Noise" or "Environmental Sound". The following sections show how the multimedia events related to the acoustic concepts "Animal Noise" or "Environmental Sound" have a bad detection rate due to the fact that the models are not well-trained.

Table 3.5: Segmentation confusion matrix for the 5 broad classes of acoustic concepts plus speech and music models. Each row of the matrix represents the ratio of frames in an actual class and each column represents the ratio of frames in a predicted class.

	CA	AN	RS	MN	ES	SP	MU
CA	**0.41**	0.03	0.03	0.04	0.04	0.20	0.23
AN	0.11	**0.01**	0.01	0.05	0.07	0.52	0.20
RS	0.07	0.02	**0.35**	0.09	0.09	0.16	0.20
MN	0.14	0.10	0.10	**0.26**	0.16	0.07	0.15
ES	0.23	0.02	0.07	0.05	**0.11**	0.12	0.13

3.4 Acoustic concepts as features for MED. Baseline Systems

3.4.1 Methods

The acoustic concepts can determine some properties of the audio to facilitate the detection of a multimedia event easier than other features. For example, a video where someone is fishing is strongly correlated with "environmental sounds" like "water splashing"

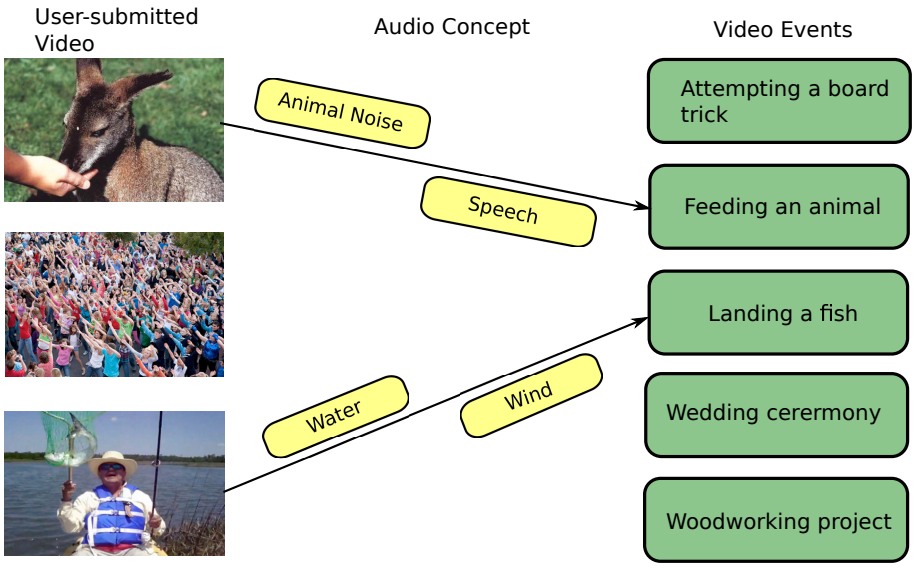


Figure 3.1: Diagram of a supervised MED approach where the acoustic concepts are used as high-level features.

or "wind" as Figure 3.1 shows. This section shows two different approaches using acoustic concepts to detect the multimedia event.

3.4.1.1 Segmental-GMM Approach

The first approach is described in (Pancoast2012a) and it is known as Segmental-GMM. In this approach, each selected concept is trained with a GMM. Then, the audio of a video is divided into fixed-length segments to generate score vectors where each element in the vector corresponds to a likelihood of a GMM concept. These score vectors are known as Segmental-GMM feature vectors. In our experiments the segmental GMM vectors are 7-dimensional (the 5 broad classes, speech and music). We generate segmental GMM vectors for every T second segment within each video. In (Pancoast2012a) different values of T were used getting the best results for $T = 5$ seconds. If $K = \lfloor \frac{M}{T} \rfloor$, a video that is M seconds long will therefore be represented by a 7xK dimensional matrix with each column corresponding to a segmental GMM vector. Therefore, the video is currently represented by a non fixed-length matrix. However, we need to have constant length features that are independent of the video length in order to be used with an SVM classifier to detect the final multimedia event. The solution proposed in (Pancoast2012a) is to represent the video with a *co-occurrence*

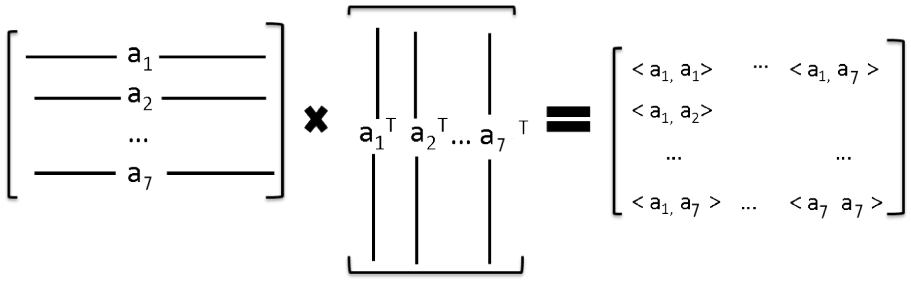


Figure 3.2: Co-occurrence segmental GMM matrix representation (Pancoast2012a). Each row of **A** corresponds to the likelihoods of a given acoustic concept occurring at every T-second fixed-length segment.

matrix where each element represents the probability that a pair of acoustic concepts occur in the video. If **A** is the segmental-GMM matrix, the co-occurrence matrix is computed as $\mathbf{A}\mathbf{A}^T$ as Figure 3.2 shows.

3.4.1.2 Acoustic Concept Recognition Approach

The second approach is described in (Castan2013) and it is known as Acoustic Concept Recognition (ACR). In this approach, each concept is modeled as an HMM/GMM of one state. The main difference with the Segmental-GMM approach is that the length of the segments is not fixed anymore and the segmentation is based on the transitions among the HMM models according to the Viterbi algorithm. The score vector is the accumulated likelihood for each model. Therefore, a video is represented by a 7xK dimensional matrix with each column corresponding to a different-length segments. Then, the elements of the original likelihood matrix are normalized with the sigmoid function:

$$S(x_{i,j}) = \frac{1}{1 + e^{x_{i,j}/a}} \tag{3.1}$$

where $x_{i,j}$ is the likelihood score corresponding to acoustic concept i in segment j. The value a was chosen empirically from the training data. The resulting values are therefore normalized to be greater than 0 and less than 1.

Finally, the matrix is vectorized to generate SVM feature vectors to be able to perform the event detection (one-against-all) for each video event class with an SVM classifier with a linear kernel.

3.4.2 Results

To measure the system performance we use Detection Error Tradeoff (DET) curves (Martin1997) which are commonly used to show the tradeoff between false alarm errors and missed detections. We generated the DET-curves with plotting software available from the NIST website (NIST). From these curves we also extracted the equal error rate (EER) as the point where the probability of false alarm (pFA) is equal to the probability of a miss (pMiss). Since TRECVid MED 2011 simulates a retrieval task from wild videos on the Internet, the assumption is that high miss rates can be tolerated in favor of low false alarm probabilities. A rate of 6% false alarms has been widely used in the literature as a low false alarm probability. Therefore, the benchmark compares the number of misses at a given false alarm rate of 6%. However we also calculate the EER because it provides a clear idea about the performance of the system.

Figure 3.3 shows the DET curves for every acoustic event. The blue curves represent the performance of the Segmental-GMM approach and the red curves represent the performance of the ACR approach. As it can be seen, system performance varies across video events. *"Wedding ceremony"* and *"Woodworking project"* show the best results while *"Feeding an animal"* and *"Landing a fish"* show the worst results. These behaviors are consistent with the previous results presented in section 3.3. It can be seen that the concepts, *"Animal sounds"* and *"Environmental sound"*, have the highest error rate and those concepts are more related to *"Feeding an animal"* and *"Landing a fish"* videos, respectively. On the other hand, the concepts *"Crowds and audience"* and *"Repetitive sounds"* have the best results and they are more related with *"Wedding ceremony"* and *"Woodworking project"* events respectively. Likewise, *"Feeding an animal"* and *"Landing a fish"* videos contain short bursts of sounds overlapping with a widely varying background noise, which make detection much more difficult.

Table 3.6 shows the EER and the benchmark given a false alarm rate of 6% for both approaches. The EER is better using Segmental-GMM for almost all the events except for the "*Wedding ceremony*" event. However, the benchmark is better using ACR with the exception of E002 event where the model is under trained and E005 where the difference between Segmental-GMM and ACR is not very significant as it can be seen on Figure 3.3. Leaving aside these subtle differences, both approaches are very similar so it can be said that there is not a very big improvement using ACR. However, the

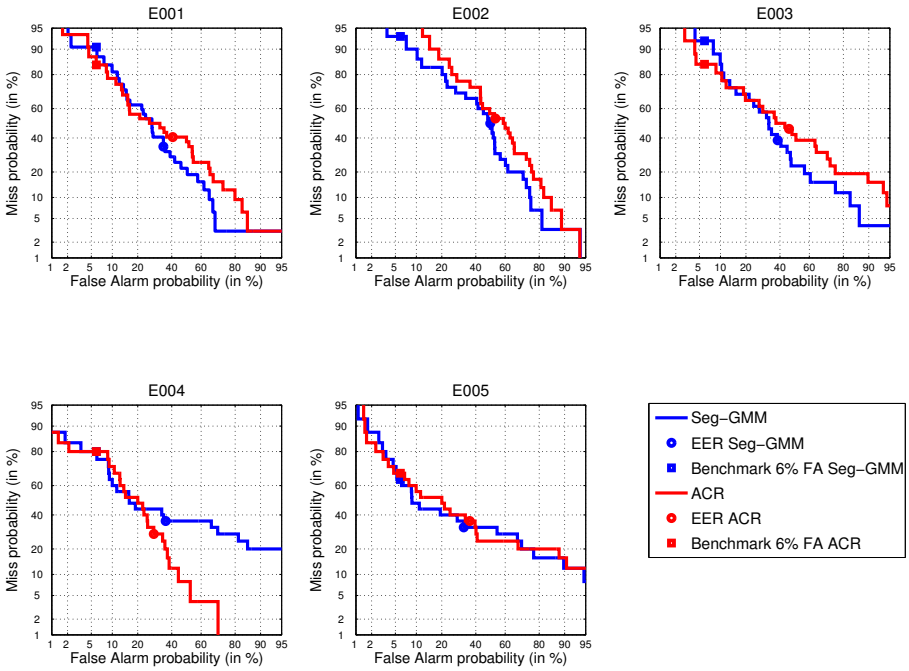


Figure 3.3: DET curves of Segmental-GMM approach versus ACR approach with the 5 broad classes, speech and music. The marks for EER and the benchmark for 6% pFA on the same curves

lattices of the recognition process can be used to provide context information about the video as is shown in the following section.

3.5 Acoustic Concept Lattices as features for MED - Context Information

3.5.1 Method

In this section, we propose an approach to model multimedia audio content with a supervised acoustic concept extraction technique. First, we employ an HMM-based *acoustic concept recognition* (ACR) system (the one described previously in 3.4.1.2) to convert the audio signal into a recognition lattice, which we refer to as *acoustic concept lattice*. Next, we create an acoustic concept index for each file from the ACR lattice by extracting posterior N-gram counts. The main idea is that a sequence of acoustic concepts can be indicative of a specific multimedia event. This approach has been successfully applied to identify different languages (Campbell2007, Richardson2008) or

Table 3.6: EER and benchmark of 6% pFA for segmental-GMM and ACR approaches

	Segm-GMM		**ACR**	
	EER	BM-6%	EER	BM-6%
E001	**0.343**	0.906	0.406	**0.843**
E002	**0.500**	**0.933**	0.533	1.000
E003	**0.384**	0.923	0.461	**0.846**
E004	0.360	0.800	**0.280**	**0.800**
E005	**0.320**	**0.640**	0.360	0.680
Mean	**0.381**	0.840	0.408	**0.833**

different dialects (Akbacak2012) by using phonetic N-gram counts. Finally, the acoustic concept indexes are used as features in an SVM-based classification for multimedia event detection (MED) task.

This approach is different to the previously mentioned supervised techniques like (Pancoast2012a, Jiang2010, Castan2013) in several ways. First, we do not use any fixed segmentation, but instead use recognition to dynamically extract acoustic concept segments. More importantly, in this approach, soft-decisions for the acoustic concept extraction are used as MED features via lattice-based representations to consider alternative recognition hypotheses, creating rich representations to be used for the MED task. Given the amount of variation in audio characteristics of user-submitted Internet videos, this becomes critical since 1-best hard-decisions will very often obtain errors and this will degrade MED performance. And the last difference is that context information is used (via N-gram representations) in our work during MED modeling.

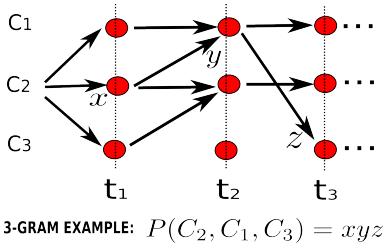


Figure 3.4: An example of 3-gram extraction from a sample acoustic concept recognition (ACR) lattice

The method can be described carefully in this way: lattices represent alternative hypotheses resulting in a richer representation compared with the 1-best recognition

output. These hypotheses can be seen as multiple paths with different likelihood for every node of the path. In our approach, each node represents an acoustic concept. The N-gram counts are the accumulated likelihoods of co-ocurrence concepts as shown in Figure 3.4.

Let us explain the N-grams counts with an easy example. Suppose we have a hypothesized string of concepts, $C = c_1, \cdots, c_n$. As an example, bigrams are created by grouping two tokens at a time to form, $C2 = (c_1c_2), (c_2c_3), \cdots, (c_{n-1}c_n)$. The count function for a given bigram, d_i $(count(di|C2))$ is the number of occurrences of d_i in the sequence C2. To extend counts to a lattice, L, we find the expected count over all possible hypotheses in the sequence:

$$count(d_i|L) = E_C[count(d_i|C)] = \sum_{C \epsilon L} p(C|L) count(d_i|C) \tag{3.2}$$

3.5.2 Results

We evaluate the performance of the lattice-based acoustic concept indexing with the DET curves, the EER and the benchmark given a false alarm rate of 6% as we did in the previous section. The HMM models described in section 3.4.1.2 were used to generate lattices across the train and the test data sets. As an initial experiment, the SVM vectors were produced by stacking 1-grams, 2-grams and 3-grams of the 5 broad acoustic concepts with speech and music, obtaining 520 dimension vectors. To get a sense of how well the lattice-based acoustic concept indexing approach (denoted as A.C. Latt.) performs, we compare the DET curves with ACR and Segmental-GMM approaches (denoted as ACR and seg-GMM in the plots respectively) for the same acoustic concepts. This is shown in Figure 3.5. It can be seen how the DET curve for the ACR-lattice approach is below the segmental-GMM approach curve for most of the events. Table 3.7 shows the improvement of the lattice count approach in terms of EER and a benchmark of 6% pFA. The proposed approach, ACR-lattices, improves the detection in almost all the events. The table shows that ACR-lattices is the best performing approach for a retrieval system because it has the lowest pMiss for the benchmark of 6% pFA. However, E002 and E003 still have the worst behavior due to the reduced amount of data to train the acoustic concept models that appear in these video event categories.

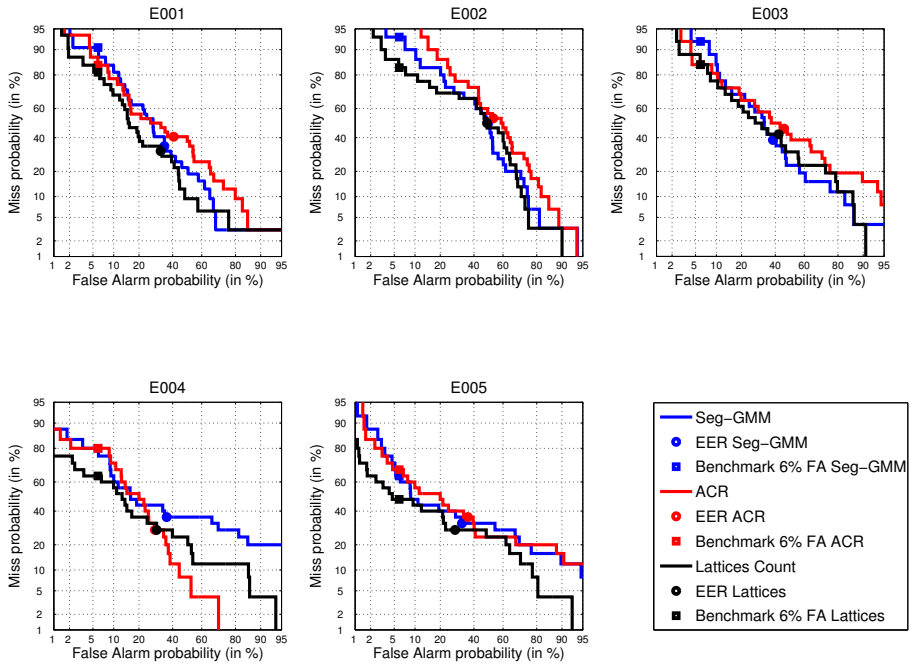


Figure 3.5: DET curves of Segmental-GMM, ACR and Lattice Count approaches. The marks for EER and the benchmark for 6% pFA on the same curves

Table 3.7: EER and benchmark of 6% pFA for segmental-GMM, ACR and Lattice Count approaches

	Segm-GMM		**ACR**		**Latt**	
	EER	BM-6%	EER	BM-6%	EER	BM-6%
E001	0.343	0.906	0.406	0.843	**0.312**	**0.812**
E002	0.500	0.933	0.533	1.000	**0.500**	**0.833**
E003	**0.384**	0.923	0.461	0.846	0.423	**0.846**
E004	0.360	0.800	0.280	0.800	**0.280**	**0.640**
E005	0.320	0.640	0.360	0.680	**0.280**	**0.480**
Mean	0.4354	0.8404	0.408	0.833	**0.359**	**0.722**

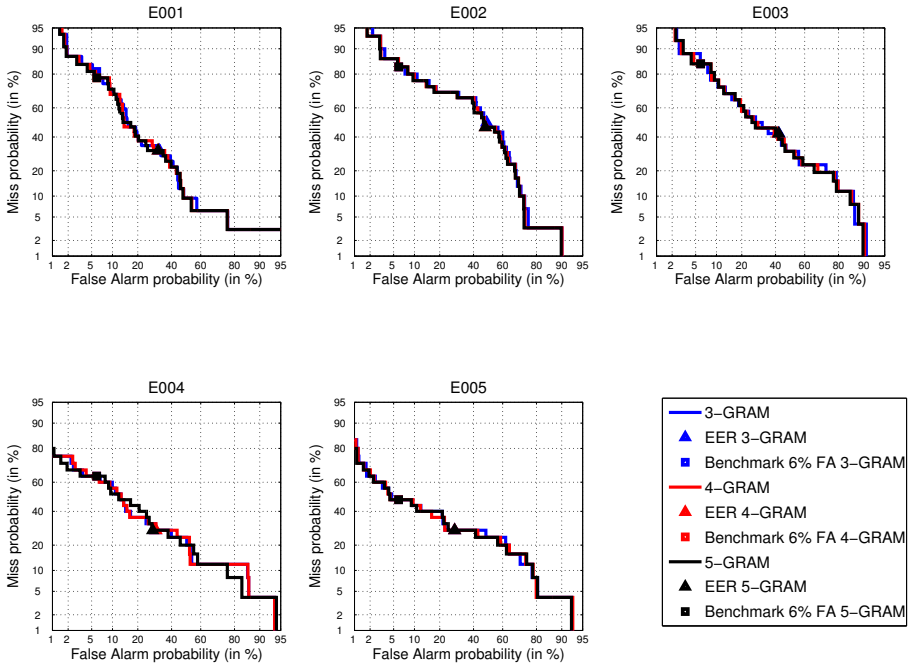


Figure 3.6: DET curves of Segmental-GMM, ACR and Lattice Count approaches. The marks for EER and the benchmark for 6% of pFA on the same curves

The SVM vector dimensions grow exponentially as we increase the order of the N-grams. Fortunately, Figure 3.6 shows the curves increasing the SVM vector with the counts of 4-grams and 5-grams obtaining 3,656 and 25,608 dimensions respectively. It can be seen that the curves are overlapped for all the events and this means that the entropy to detect the events is located in the first order of the N-grams.

3.5.3 Comparison of the Lattice Count approach with other approaches

Next, we compare the proposed ACR-lattices approach with other approaches outside this thesis. We use the 20 acoustic concepts and the segmental-GMM approach used in (Pancoast2012a). We also compare the proposed approach with an unsupervised approach which is a bag-of-audio-words (BoAW) (Pancoast2012). For all of these approaches, the same train and test sets are used. Because the segmental-GMM approach is evaluated with 20 acoustic concepts instead of the broad acoustic concepts used previously, we extend the ACR-lattices approach with the specific acoustic concepts presented in Table 3.2 using 1 state HMM with 256 Gaussians for every acoustic concept as we did for the broad acoustic concepts. The SVM vectors in this experiment

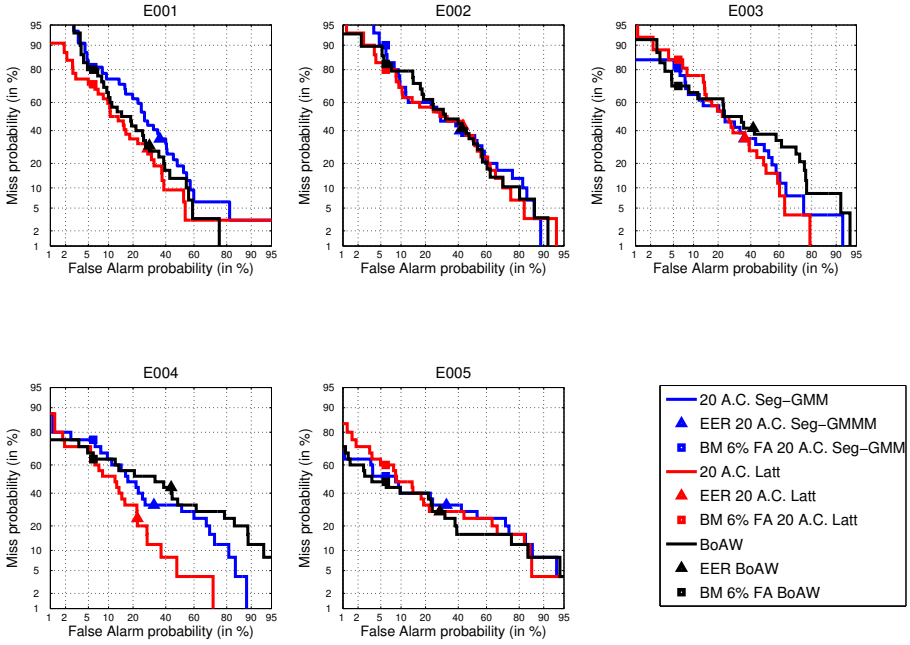


Figure 3.7: DET curves of segmental-GMM with 20 acoustic concepts, Lattice Count with 20 acoustic concepts and BoAW approaches. The marks for EER and the benchmark for 6% of pFA on the same curves

use 3-grams because increasing the order of the N-grams further shows a slight improvement in the detection performance, but vector dimensions increase exponentially. For the BoAW approach, a codebook size of 1000 is used as this was found to yield the best BoAW results in (Pancoast2012).

Figure 3.7 shows the DET curves for these approaches. As it can be seen, for almost all the events, the curve corresponding to ACR-lattices has better behavior than the BoAW and segmental-GMM curves. Also, the performance of the BoAW for events E002 and E003 is very bad even if the BoAW creates unsupervised clusters. This shows the difficulty in detecting that multimedia events using the audio of the video. Table 3.8 summarizes the EER and the pMiss for a benchmark of 6% pFA. Also for these marks, the ACR-lattice approach shows very good behavior.

The next section shows the benefit of this approach as part of a complete audio processing system with spoken and acoustic concepts to detect multimedia events.

Table 3.8: EER and benchmark of 6% pFA for segmental-GMM with 20 acoustic concepts, Lattice Count with 20 acoustic concepts and BoAW approaches

	Segm-GMM		**Latt**		**BoAW**	
	EER	BM-6%	EER	BM-6%	EER	BM-6%
E001	0.343	0.812	**0.281**	**0.718**	0.300	0.800
E002	**0.400**	0.900	0.433	**0.800**	0.413	0.827
E003	0.346	0.807	**0.346**	0.846	0.416	**0.708**
E004	0.320	0.760	**0.240**	**0.640**	0.440	0.640
E005	0.320	0.520	**0.280**	0.600	0.280	**0.480**
Mean	0.345	0.759	**0.316**	0.720	0.369	**0.691**

3.6 Spoken and Acoustic Concept Fusion for MED

Speech content is rich in information since some keywords will provide invaluable clues for detecting certain events. Words like "water", "boat" or "fish" are closely related to the event "Landing a fish". The purpose of this section is to explore using acoustic concepts and spoken concepts extracted via audio segmentation/recognition and speech recognition respectively for Multimedia Event Detection (MED). The fusion of both approaches can improve the detection rate since the information provided by the systems is complementary.

We first present the Automatic Speech Recognition (ASR) system used to extract spoken concepts. The ASR lattices are used to get expected counts (as we did in the previous section with the ACR) since the counts provide a more robust measure of word appearance than 1-best ASR, and use them as features for MED. To model spoken concepts, we consider a linear SVM. Various feature processing techniques are presented to improve the performance of the spoken concepts: word counts weighting and feature dimension reduction by stemming. Finally, we merge the information coming from the spoken concept system with the information coming from the acoustic concept system described in section 3.5 to improve the detection rate. An exhaustive description of the system can be read in (VanHout2013) and an overview of our system is shown in Figure 3.8.

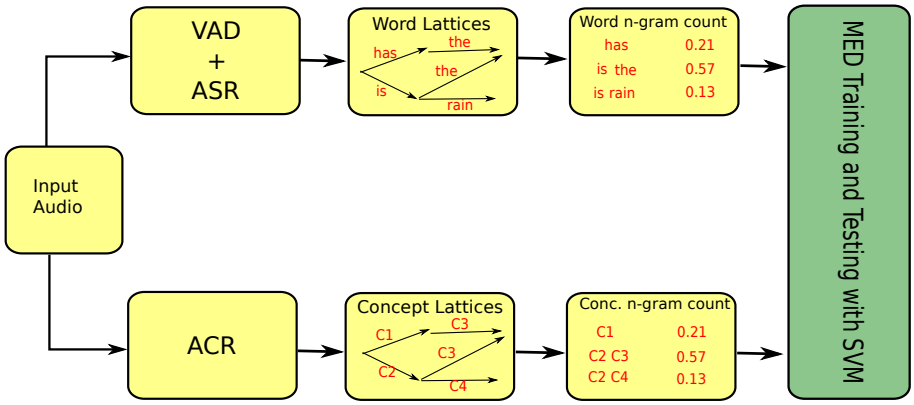


Figure 3.8: Extracting acoustic and spoken concepts as features for MED

3.6.1 Extracting Spoken Concepts

Since audio from multimedia events is so heterogeneous in nature, a good segmentation is essential in order to determine which segments need to be fed to the ASR system. Here, we describe the ASR system that was chosen to generate spoken concepts. For this task, we ran an English ASR system trained on data with channel and speaker characteristics related as much as possible to those of the observed TRECVid MED data. The system used 12 cepstral coefficients, energy, first-, second-, and third-order time-derivatives, and 2×5 voicing features over a 5-frame window.

The 1-best ASR output is the most likely word sequence extracted from the lattice, while expected word counts are computed similarly to the concept N-gram counts with $N = 1$. Because of the relatively low accuracy (28%) of the 1-best ASR due to the noise and overlapped audio, lattice-based counts are expected to be more reliable than 1-best ASR. For each video, these counts are aggregated into a feature vector of dimension 54484, the size of the ASR vocabulary.

Since some words are inherently more frequent than others, their counts can be several orders of magnitude larger than counts of rarer but potentially discriminative words. We tried a weighting scheme by which we boost counts of infrequent words over frequent ones. The weighting approach ($\mathbf{W_{Log}}$) maps raw expected counts $c(w|L)$ to log counts $c_{log}(w|L)$ as follows: $c_{log}(w|L) = log(c(w|L) + f)$ where f is a flooring parameter that was optimized to 10^{-4} limiting the impact of infrequent words.

3.6.2 Results

MED results are shown in Table 3.9 for each of the five events in terms of Average P_{miss} (APM), which measures the area under a Detection-Error Tradeoff curve.

Table 3.9: Average-P_{miss} by event for the proposed MED systems

System	Event					
ASR Lattice SVM	E001	E002	E003	E004	E005	Avg.
no-stem	0.30	0.37	0.50	0.23	0.29	0.34
stem	0.30	0.38	0.45	0.23	0.28	0.33
no-stem W_{Log}	0.26	0.27	0.23	0.20	0.16	0.22
stem W_{Log} (**1**)	0.26	0.27	0.21	0.19	0.14	**0.21**
ACR Lattice SVM	E001	E002	E003	E004	E005	Avg.
5 broad concepts (**2**)	0.22	0.41	0.32	0.25	0.28	0.30
20 specific concepts (**3**)	0.17	0.30	0.23	0.15	0.20	**0.21**
Fusion ASR + ACR	E001	E002	E003	E004	E005	Avg.
(**1**) + (**2**)	0.21	0.23	0.22	0.17	0.15	0.20
(**1**) + (**3**)	0.16	0.23	0.15	0.11	0.13	**0.15**
(**1**) + (**2**) + (**3**)	0.14	0.23	0.16	0.12	0.13	**0.15**

We observe that stemming the words and adding the counts to those sharing the same stem results in a relatively small improvement since the APM decreases by 0.1 when stemming is applied, with and without log-weighting. The MED system based on ACR performed significantly better with 20 events (0.21 APM) than with 5 events (0.30) as we pointed out in the previous section. Note that the level of performance is similar to that of the best ASR-based MED system. This clearly shows the tremendous importance of ACR approaches for MED.

In order to leverage information from both Acoustic and Spoken concepts, we perform score-level fusion of these two or three systems by normalizing their prediction scores to have zero-mean and unit variance and adding them with equal weighting. The best performing combination was obtained by combining the ASR MED system with the 20-concepts ACR MED system. The combined system performed better than both of the original systems for all five events, and provided a relative 28% improvement in APM (from 0.21 to 0.15) in average over all events. This result shows that acoustic and

spoken concepts capture different kinds of information that can easily be combined to build a significantly more robust MED system.

3.7 Chapter Summary

In this chapter, we have presented a robust approach to build an audio-based Multimedia Event Detection system using N-gram counts from the lattices of acoustic concept recognition. This approach has been compared with other supervised and unsupervised approaches showing an improvement in terms of event detection. A similar approach has been used to extract spoken concepts. Both systems have been merged since the information is complementary. After score-level system combination of both MED systems, we leverage the complementarity of both approaches and obtain a 28% relative decrease in APM.

It has become apparent that a good segmentation and a good concept classifier are crucial for MED. Therefore, these systems must be robust enough to compensate the enormous variability that is presented in Internet videos. In the next two chapters, we will present different approaches to compensate this variability at model level, introducing techniques based on Factor Analysis.

Chapter 4

Audio Segmentation-by-Classification

Contents

4.1 Chapter Overview

The previous chapter has shown the importance of a proper segmentation to identify the speech/non-speech segments to provide speech transcription or acoustic concept transcription respectively. An important issue to segment and classify speech/non-speech segments is the high variability that can be found in each class. This chapter will present a novel system to segment and classify audios into broad classes with techniques to compensate the variability. The approach will be tested in a broadcast TV news (BN) domain since this is a challenging scenario because the audio is comprised of several materials with a non-homogeneous style. Therefore, it becomes suitable to propose systems that are able to work with unconstrained databases as MED. Some examples of the audio sequences in BN in different conditions are as follows:

- **News Anchor Speech:** Traditional news anchor reading text in clean conditions.
- **Interviews:** Conversations between two people with spontaneous speech.
- **Debates:** Conversations between two or more people with overlapped speech.
- **Reporter in the Field:** The audio comes from a wide range of noises generally overlapped with speech.
- **Advertising:** Speech with music in the background and a variety of acoustic noise effects (slams, explosions, cars, screams ...).
- **Jingles:** Jingles are commonly used as a short tune to introduce different topics during the news.
- **Broadcasting of sports events:** Speech with a strong background noise, diegetic music and sounds.
- **Telephone Connections:** Commonly used when reporters do not have a camera or microphone.

Recently, an audio segmentation task in the BN domain in the context of the Albayzin 2010 evaluation campaigns was proposed in (Butko2010a). The proposed evaluation task consisted of segmenting a broadcast news audio document into five acoustic

classes: speech (SP), speech with noise (SN), speech with music (SM), music (MU), and others (OT). The main difficulty in this database is the classification among the classes with speech because these classes have some very homogeneous segments (especially, between SP and SN). In this context, we introduce a novel and generic segmentation-by-classification system based on factor analysis (FA) with two clear advantages: (1) the system does not need class-dependent features with a hierarchical structure to classify different classes and (2) the algorithm compensates the within-class variability with high accuracy being able to classify well-defined classes in generic tasks. The FA technique has been successfully applied in speaker ID (Kenny2005, Kenny2006, Kenny2007), speaker diarization (Vaquero2010a, Vaquero2011, Vaquero2013), and language recognition (Brummer2009) with significant improvements with respect to previous approaches. However, the system proposed in this thesis has several differences from those systems. In contrast to a segmentation task, the speaker ID or language recognition has well-delimited segments (usually in separate files) and, therefore, FA is applied over the whole file. Unlike speaker ID, speaker diarization, or language recognition tasks, we can find here the same speaker in two different acoustic classes, for example, the situation where an anchor is inside the studio with clean conditions (SP) and outside the studio with noise in the background (SN). Due to all these factors, we propose an extension of a FA segmentation system proposed in (Castan2013a) and (Castan2013c) with a new and more discriminative scoring using class/non-class parameters and with a set of back-end systems that perform a better segmentation than the traditional FA systems for language recognition or speaker ID.

The remainder of the chapter is organized as follows: the database and metric of Albayzin 2010 evaluation is presented in Section 4.2. Section 4.3 shows the theoretical approach based on FA and the smoothing back-end subsystem. The experiments are presented in Section 4.4. Finally, a chapter summary is presented in Section 4.5.

4.2 Dataset & Metric

4.2.1 Database

The Albayzin campaigns are internationally open evaluations organized by the RTTH[1] every 2 years. A complete description of the Albayzin 2010 audio segmentation and clas-

[1]Spanish Thematic Network on Speech Technologies: http://www.rthabla.es.

sification evaluation can be found in (Butko2011) where the participant's approaches and the results are presented. We describe the database and the metric used in the evaluation in the following subsections.

The database consists of BN audio in Catalan recorded by the TALP[1] Research Center. It includes approximately 87 h of annotated audio divided into 24 files. Five audio classes were defined for the evaluation. The classes are distributed as follows: clean speech, 37%; music, 5%; speech over music, 15%; speech over noise, 40%; others, 3%. The class 'others' is not evaluated in the final test. The database for the evaluation was split into two parts: for training (two thirds of the total amount of data divided into 16 files) and testing (the remaining one third divided into 8 files).

Each segment is labeled with one previously described class. Most of the segments are between 10 and 20 s long. However, there is a considerable amount of long segments (longer than 60 s). More details about the database and the labeling process can be found in (Butko2011).

4.2.2 Metric

The metric that was proposed for the evaluation represents the relative error averaged over all acoustic classes (ACs):

$$\text{Error} = \text{average}_i \left(\frac{\text{dur}(\text{miss}_i) + \text{dur}(\text{fa}_i)}{\text{dur}(\text{ref}_i)} \right), \tag{4.1}$$

where $\text{dur}(\text{miss}_i)$ is the total duration of all deletion errors (misses) for the ith acoustic classes (AC), $\text{dur}(\text{fa}_i)$ is the total duration of all insertion errors (false alarms) for the ith AC, and $\text{dur}(\text{ref}_i)$ is the total duration of all the ith AC instances according to the reference file. The incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. A collar of 1 s is not scored around each reference boundary to avoid the uncertainty about when an AC begins or ends.

Since the distribution of the classes in the database is not uniform, the errors from different classes are weighed differently (depending on the total duration of the class in the database). Therefore, the system has to correctly detect not only the best

[1]The Center for Language and Speech Technologies and Applications (TALP) is a specific interdepartmental research center at the Technical University of Catalonia (UPC).

Table 4.1: The five acoustic classes defined in the Albayzin evaluation for audio segmentation in BN

Class	Description
Music [MU]	Music is understood in a general sense (as jingles or diegetic music)
Speech over music [SM]	Overlapping of speech and music classes or speech with noise in background and music classes. This class can be found mostly in spots
Speech over noise [SN]	Speech not recorded in studio conditions or overlapped with some type of noise. Several simultaneous voices belong to this class and also the telephone connections.
Speech [SP]	Clean speech from a close microphone without any kind of background sound. This class is typically composed of anchor speeches.
Others [OT]	This class refers to any type of audio signal (like silence and noises) that does not correspond to the other four classes

represented classes ('speech' and 'speech over noise,' 77% of total duration) but also minor classes (like 'music,' 5%). This metric is different to other segmentation tasks such as the conventional NIST metric (NIST2009) for speaker diarization, where the score is defined as the ratio of the overall segmentation error time to the sum of the durations of the segments that are assigned to each class in the file. In this work, we will present the final results with both metrics.

4.3 Novel Factor Analysis audio segmentation system

We propose a framework for automatic audio segmentation-by-classification. The system deals with the problem of assigning a class label to each fixed-length window using factor analysis (FA) models. In tasks like speaker verification, speaker diarization or language recognition, the systems have to face several sources of variability such as speaker, channel, and environment. The variability of the same class segments is known as within-class variability. The goal of these systems is to model (in the case

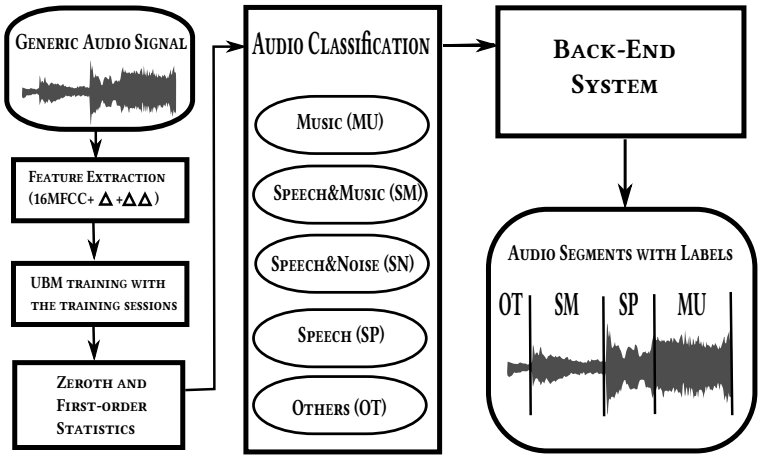


Figure 4.1: Block Diagram of Factor Analysis Segmentation-by-Classification System for Broadcast News Classes

of (Vaquero2011)) or compensate the within-class variability to reduce the mismatch between training and test. As we presented in the first section, there are some differences between those systems and the segmentation-by-classification system proposed in this work. The main difference is that in this task, the classes may contain the same speaker. However, in speaker ID, speaker diarization, or language recognition, the speakers define an independent class. As a result, the within-class variability is more difficult to compensate in our task. Therefore, we introduce a novel approach with class/non-class parameters that compensate the within-class variability more accurately. Figure 4.1 illustrates the proposed framework where each block is described in the following subsections.

4.3.1 Acoustic Feature Extraction

In this work, we extract 16 MFCCs (including the zeroth-order cepstrum) computed in 25-ms frames with a 10-ms frame step and their first and second order time-derivatives. The audio features are packed in windows of 3 s long with 0.1- or 0.5-s window steps depending on the desired computational load and resolution.

4.3.2 Statistics Computation

The fixed-length windows are mapped to sufficient statistics by using a universal background model (UBM) (Reynolds2000) which is a class-independent GMM with C Gaus-

sians estimated with the expectation-maximization (EM) algorithm (Bishop2006) on the training data set. The UBM parameters are the mean vectors, $\boldsymbol{\mu}_k$, and the diagonal covariances matrices, $\boldsymbol{\Sigma}_k$, where k is the Gaussian component index. Let $P_{ksi} = P(k|\boldsymbol{\phi}_{si})$ represent the posterior probability of the kth UBM component, given the feature vector $\boldsymbol{\phi}_{si}$, for a window s with feature vectors indexed $i = 1, 2, \ldots, N_s$, we define the zeroth- and first-order statistics, respectively, as

$$n_{sk} = \sum_{i=1}^{N_s} P_{ksi}, \tag{4.2}$$

$$\boldsymbol{f}_{sk} = \sum_{i=1}^{N_s} P_{ksi} \boldsymbol{\Sigma}_k^{-1/2} (\boldsymbol{\phi}_{si} - \boldsymbol{\mu}_k), \tag{4.3}$$

assuming frame independence (Kenny2010a). These statistics are normalized in mean and standard deviation to the UBM (which defines the center of an affine space).

4.3.3 Theoretical Background

Data from a particular class are modeled by a GMM defined by a set of mean vectors $\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_C$, weights $w_1, w_2, \ldots, w_C$, and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_C$, where C is the number of Gaussians. We can concatenate all GMM mean vectors to one mean supervector $\boldsymbol{m}$ of dimension $CF \times 1$ where F is the feature vector length:

$$\boldsymbol{m} = [\boldsymbol{m}_1^T, \boldsymbol{m}_2^T, \ldots, \boldsymbol{m}_C^T]^T. \tag{4.4}$$

The factor analysis model is the adaptation of the UBM model where the supervector of means is not fixed and it can vary from segment to segment due to several sources that increase the within-class variability (Kenny2007). We assume that these GMMs have segment- and class-dependent means but fixed weights and covariances chosen to be equal to the UBM weights and covariances. Specifically, we use a factor analysis model for the mean vector of the kth component of the GMM for segment s:

$$\boldsymbol{m}_k^s = \boldsymbol{t}_k^{c(s)} + \boldsymbol{U}_k \boldsymbol{x}_s, \tag{4.5}$$

where $c(s)$ denotes the class of segment s. $\boldsymbol{U}_k$ is the factor loading matrix that defines the subspace of the within-class variability and $\boldsymbol{x}_s$ is a vector of L segment-dependent-within-class-variability factors assumed to follow a normal distribution ($N(0, I_L)$). The

class location vector $t_k^{c(s)}$ is obtained by using a single iteration of relevance MAP adaptation from the UBM (Reynolds2000). This adaptation is expressed, in terms of statistics, as

$$\boldsymbol{t}_k^{c(s)} = \frac{\sum_s \boldsymbol{f}_{sk}}{r + \sum_s n_{sk}}, \tag{4.6}$$

where r is the relevance factor.

We stack the component-dependent vectors into supervectors $\boldsymbol{m}_s$ and $\boldsymbol{t}^{c(s)}$ and the component-dependent $\boldsymbol{U}_k$ matrices into a single tall matrix $\boldsymbol{U}$, so Equation 4.5 can be expressed as

$$\boldsymbol{m}^s = \boldsymbol{t}^{c(s)} + \boldsymbol{U}\boldsymbol{x}_s, \tag{4.7}$$

where $\boldsymbol{U}$ is known as the within-class variability matrix that we use to compensate that variability. The columns of the $\boldsymbol{U}$ matrix are the basis spanning the subspace of the within-class variability, and the within-class variability factors are the coordinates defining the position of the supervector in the subspace. The within-class variability factor dimension (L) is smaller than CF, so $\boldsymbol{U}$ has low rank ($CF \times L$ dimensions). Depending on the application, the value of L is between 50 and 200 and CF can be 98,304 if we have 2,048 Gaussians and 48-dim feature vector (with the MFCC-UBM).

4.3.4 Estimation of the Within-Class Variability Matrices

$\boldsymbol{U}$ can be estimated using the EM algorithm, where the $\boldsymbol{x}$ factors of each window are treated as hidden variables. In the E step, the expected value of $\boldsymbol{x}$ (denoted by $\hat{\boldsymbol{x}}$) are estimated for each window, using the current parameters as

$$\hat{\boldsymbol{x}}_s = \left(\boldsymbol{I} + \sum_k n_{sk}\boldsymbol{U}_k^T\boldsymbol{U}_k\right)^{-1}\boldsymbol{U}^T\boldsymbol{f}_s. \tag{4.8}$$

In the M step, we obtain $\boldsymbol{U}$ that maximizes an auxiliary function involving the old and new parameters as

$$\boldsymbol{U}_k = \left[\sum_c \sum_s (\boldsymbol{f}_{sk} - \boldsymbol{t}_k^{c(s)} n_{sk})\hat{\boldsymbol{x}}_s^T\right]\boldsymbol{A}_k^{-1}, \tag{4.9}$$

where

$$\boldsymbol{A}_k = \sum_s \left[\hat{\boldsymbol{x}}_s\hat{\boldsymbol{x}}_s^T\right]^T n_{sk}. \tag{4.10}$$

This chapter does not aim to go into the training process of $\boldsymbol{U}$ in depth; more details and an exhaustive description can be found in (Kenny2007) and in Appendix A of this thesis.

4.3.5 Class model vs alternative model U matrices

The approach proposed in this thesis has several differences with language recognition in the way within-class variability is compensated. Most of the approaches based on FA for language recognition are implemented with a single $\boldsymbol{U}$ matrix because the segments are well-delimited (typically in separated files) and the nature of the within-class variability is similar for all the languages as it can be seen in (Li2012, Castaldo2007, Vogt2008, Kenny2007). In this case, the variability subspace defined $\boldsymbol{U}$ is mostly due to the different speakers of a same language. In (Castan2012), a segmentation system was proposed with five class location vectors (one vector per class) and a single compensation matrix $\boldsymbol{U}$ for all the classes. The paper compared the FA system with the winner of the Albayzin 2010 evaluation, and the conclusion was that the FA system is better as a classification system with oracle segments. On the other hand, the compensation matrix had a bad behavior in a segmentation-by-classification system for the music class due to the different nature of the rest of the classes. In (Castan2011), a hierarchical system was proposed with different features and different techniques at each level depending on the class. Firstly, the system decides between MU, SM, and the rest of the classes by using HMM/GMM and a smoothed combination of MFCC and Chroma as feature vectors. Next, the system classifies SP and SN by using FA and MFCC as acoustic features to improve the performance of the speech classes because the confusion between these two classes is very high. The error rate achieved was lower than the one obtained by the best system presented in the Albayzin 2010 evaluation showing a clear advantage when the classes are homogeneous (like SN and SP), since $\boldsymbol{U}$ models the variability across speakers and phonemes. The background noise is, then, the discriminative information for the classification and segmentation. Nevertheless, hierarchical systems can be very specific for an intended task and are difficult to adapt to other databases with new classes.

Therefore, we propose here a non-hierarchical segmentation-by-classification system with ten class-specific vectors (one class vector and one non-class vector for each class) and five matrices modeling the within-class variability of each pair of class/non-class. Let

$$\begin{aligned} \boldsymbol{T} \quad = \quad & [\boldsymbol{t}^{\mathrm{MU}}, \boldsymbol{t}^{\overline{\mathrm{MU}}}, \boldsymbol{t}^{\mathrm{OT}}, \boldsymbol{t}^{\overline{\mathrm{OT}}}, \\ & \boldsymbol{t}^{\mathrm{SM}}, \boldsymbol{t}^{\overline{\mathrm{SM}}}, \boldsymbol{t}^{\mathrm{SN}}, \boldsymbol{t}^{\overline{\mathrm{SN}}}, \boldsymbol{t}^{\mathrm{SP}}, \boldsymbol{t}^{\overline{\mathrm{SP}}}] \end{aligned} \tag{4.11}$$

and

$$\begin{aligned} \boldsymbol{\Xi} \quad = \quad & [\boldsymbol{U}^{\mathrm{MU}-\overline{\mathrm{MU}}}, \boldsymbol{U}^{\mathrm{OT}-\overline{\mathrm{OT}}}, \\ & \boldsymbol{U}^{\mathrm{SM}-\overline{\mathrm{SM}}}, \boldsymbol{U}^{\mathrm{SN}-\overline{\mathrm{SN}}}, \boldsymbol{U}^{\mathrm{SP}-\overline{\mathrm{SP}}}], \end{aligned} \tag{4.12}$$

where $\boldsymbol{T}$ represents the locations of classes ($\boldsymbol{t}^{C}$) and non-classes ($\boldsymbol{t}^{\overline{C}}$) in the GMM space and Ξ the within-class variability matrices. This approach will be compared to the classic formulation with a single $\boldsymbol{U}$ matrix in 'Experimental results' section for the classification over the oracle segments and the final segmentation system.

4.3.6 Scoring

Here, we study the two most commonly used scoring approaches: integration over the $\boldsymbol{x}$ factors distributions and linear scoring, both of which are summarized in (Glembek2009).

Score 1: The score based on the integration over the $\boldsymbol{x}$ factors distributions is a marginalization using a point estimate of the class $\boldsymbol{m}^{s}$, integrated only over the $\boldsymbol{x}$ factors, when the statistics are centered around the point estimate $\boldsymbol{m}^{s}$ as defined in (Kenny2007).

Score 2: The linear scoring, which is faster than the previous one, is an approximation that makes use of the first-order Taylor expansion (Glembek2009).

In (Kenny2005a, Glembek2009) and (Kenny2007), the score used to detect the speaker is the log-likelihood ratio test (LLR)

$$\mathrm{LLR}_{\mathrm{class}} = \log \frac{P(\chi/\mathrm{class})}{P(\chi/\mathrm{UBM})}, \tag{4.13}$$

where the numerator is the likelihood for the class model and the denominator the likelihood for the UBM. Note that the UBM is used as a general model to describe the alternative hypothesis which is appropriated for speaker identification where the hypothesized speaker is not in the UBM. However, our problem has a small number of classes, and therefore, each class is highly represented by the UBM and may corrupt the test statistics especially with the larger classes.

Here, we propose a compensated log-likelihood ratio test (CLLR) scoring:

$$\mathrm{CLLR}_{\mathrm{class}} = \log \frac{P(\chi/\mathrm{class})}{P(\chi/\overline{\mathrm{class}})}, \tag{4.14}$$

where the alternative hypothesis is the likelihood for the non-class model which is also compensated with the with-in class variability matrix. The CLLR is expected to be

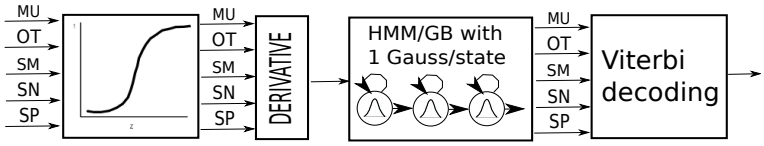


Figure 4.2: Back-end system 2 - derivative HMM/GBE block diagram

more discriminative than the LLR for a segmentation task because the hypothesized class is not present in the denominator and, also, because the non-class model is compensated in the same way as the class model.

4.3.7 Back-end systems

Here, we propose three different back-end systems to combine, smooth, and improve the classification performance of the FA:

1. *Maximum a posteriori (MAP)*: This well-known method has been widely used in the literature (Castan2013a, Castan2013c). To increase the detection performance, we optimize the prior probabilities in a Viterbi algorithm over the training files. Later, these priors are used in the Viterbi over the test files.

2. *Derivative HMM/GBE*: There is an apparent correlation between the likelihood ratios of different classes. For example, if a segment is a jingle, the likelihood ratio for the MU class (music) should be the biggest, but it is very likely that the second one is the SM (speech with music). Furthermore, SN (speech and noise) and SM (speech with music) are more correlated to each other than to the SP class (speech) because both classes have background audio. The classification and, therefore, the segmentation can be improved by combining the outputs of each class-dependent subsystem (Kittler1998).

   Figure 4.2 shows the combination and smoothing back-end system proposed here. In a first step, a calibration of scores is made by a multi-class logistic regression (Brummer2010) estimated using the training partition of the database. In order to benefit from the use of the dynamic behavior of the scores, we compute the first- and second-order time derivatives of the scores. To smooth the decisions after the calibration and the dynamic description, one Gaussian/HMM back-end is used for each class. A left-to-right topology was selected with a full-covariance

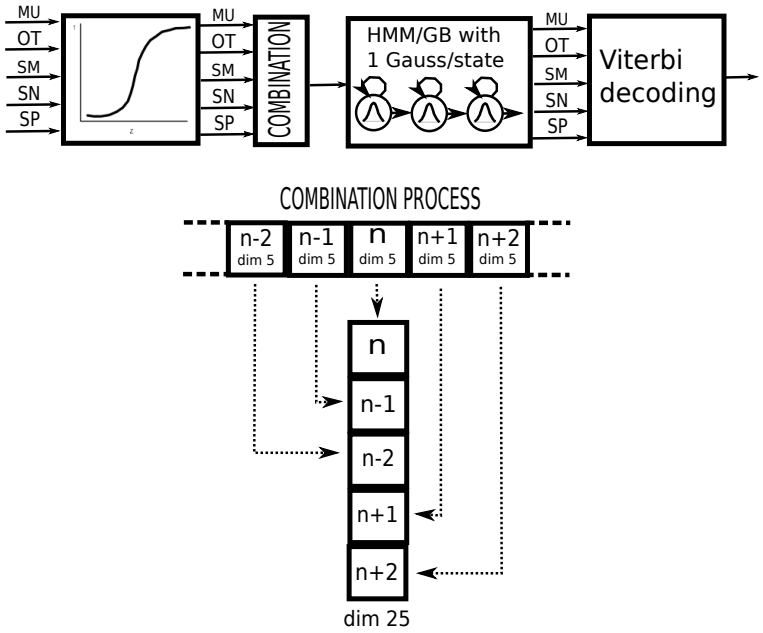


Figure 4.3: Back-end system 3 - stacking HMM/GBE block diagram

Gaussian per state estimated with the scores from the training files. The mean vectors and the covariance matrices are estimated with the samples of the scores based on the class labels with the ML criterion. The number of states for the HMM depends on the desired level of smoothing. The Viterbi algorithm was chosen to determine the maximum likelihood transitions between the classes.

3. *Stacking HMM/GBE*: This back-end system can be considered as a modification of the previous back-end system. The main idea is to provide contextual information through longer term temporal scoring. Instead of the derivation of the scores, this back-end system proposes a stacking of past and future scores with the present score to model the dynamic behavior in a different way. Figure 4.3 shows this combination process where several score frames from the past and several score frames from the future are stacked with the present frame. The experiments are carried out with one, two, and three frames from the past and future and different numbers of states.

In an HMM segmentation system, it is usual to optimize the transition penalties on a development set since this can have a significant impact on performance. However, we

do not optimize any transition penalty because our goal was to create a general approach to segment audio that could be used in other databases with different distributions or with other classes.

4.4 Experimental results

The errors can be produced in two ways: first, a classification error due to a bad labeled frame, and a segmentation error due to a temporal mismatch between the reference boundaries and the hypothesis boundaries. This section shows the experiments for the evaluation data described in Section "Albayzin audio segmentation evaluations and database description" divided into two sets. In the first set, the boundaries between segments are given by the ground truth and the system decides the class of each segment with no segmentation error to evaluate the classification accuracy of the classical FA system versus GMMs. These experiments assess the classification ability of the proposed approach and provide a fair comparison with classical GMMs.

The second set of experiments shows the segmentation and the classification error when the boundaries are not given. A final segmentation-by-classification system based on FA with a class/non-class parameters is proposed. The three back-end systems previously described are tested over this system. The back-end systems show that a combination and smoothing of the scores improve the previous results. Likewise, the systems are compared with the winner system of the 2010 Albayzin evaluation that has a hierarchical structure with specific features for each class.

4.4.1 Classification experiments with oracle segmentation

The classification is made over the segments extracted with the ground truth boundaries to evaluate the classification accuracy over the whole segment. Since the system decides the class that the whole segment belongs to, the smoothing is not needed.

We propose GMM systems as a baseline for classification experiments using the acoustic features described in Section "Acoustic feature extraction". Table 4.2 shows the results for these systems. We have evaluated a different number of Gaussians (from 64 to 2048). The classification is based on the highest accumulated likelihood over the whole segment. As shown in Table 4.2, increasing the number of Gaussians improves the final result. The highest number of Gaussians evaluated was 2048 because the error

Table 4.2: Baseline for classification experiments. Classification error per class and total error for GMM systems with different number of Gaussians over the test files with perfect segmentation

GMM	MU	SP	SM	SN	TOTAL
64G	10.68	45.74	36.68	45.44	34.63
128G	9.81	41.79	32.02	40.75	31.09
256G	10.4	37.6	31.8	37.6	29.3
512G	9.5	35.9	29.3	35.9	27.7
1024G	9.3	34.9	27.0	34.3	26.4
2048G	9.6	33.3	28.0	34.0	**26.2**

for MU and SM classes began to increase although the total result improved slightly compared with the 1024G model.

In the experiments with FA for classification with oracle segmentation, we assess different configurations for the number of $\boldsymbol{x}$ factors and the scoring methods described previously. The UBM used to compute the statistics has a fixed number of 2048 Gaussians to be able to compare the results of the FA systems with the best GMM baseline configuration. Because the boundaries are known, the statistics are calculated over the whole segment without merging underlying partitions. We compute the result using *linear scoring* and the *integration trough the* $\boldsymbol{x}$ *factors distributions scoring* (called as IoChD in this section). The *linear scoring* needs a final calibration because the scoring is scaled by the module of the target model. A Gaussian Back-End (GBE) ((Hubeika2010) (Martinez2011)) provides benefits in two ways: calibration and score combination. The calibration for the *IoChD scoring* does not provide substantial benefits because the score is based on a likelihood ratio over a MAP adaptation using the same UBM and the marginal improvement comes from the combination of scores. The experiments are carried out with a single $\boldsymbol{U}$ matrix to compensate all the within-class variability and different number of $\boldsymbol{x}$ factors (50, 100, 150, 200, 250 and 300) providing the error for each class. Note that the increment of $\boldsymbol{x}$ factors involves an exponential increment of the computational cost.

Table 4.3 compares all the experiments with FA over a perfect segmentation. According to these results, the IoChD scoring is more accurate than the linear scoring for all the configurations and all the classes. Comparing Table 4.2 and Table 4.3, a

Table 4.3: FA systems for classification experiments. Classification error per class and total error for linear and IoChD scoring with perfect segmentation and a single U for all the classes.

	One U for all classes									
	Linear-GBE					IoChnf				
# chnf	**MU**	**SP**	**SM**	**SN**	**TOTAL**	**MU**	**SP**	**SM**	**SN**	**TOTAL**
50	21.6	16.9	23.6	23.4	21.4	10.2	15.9	24.2	21.4	17.9
100	21.8	17.4	21.0	22.9	20.8	9.1	16.0	20.2	20.0	16.3
150	20.8	17.7	20.5	23.5	20.6	9.4	15.5	18.0	18.9	15.4
200	20.7	17.8	20.5	22.4	**20.4**	9.0	15.7	17.3	19.1	15.3
250	20.0	19.2	20.2	23.1	20.6	8.5	16.7	16.0	19.4	**15.1**
300	21.3	19.5	20.5	21.7	20.8	9.8	15.0	18.9	18.9	15.6

significant improvement can be seen using FA versus GMM. Using the best GMM configuration (2048 Gaussians) as reference, the worst FA system improves the total result (18.3% relative error reduction with linear scoring and 50 $\boldsymbol{x}$ factors) and also compared to the best FA configuration (43.1% relative error reduction with IoChD scoring and 250 $\boldsymbol{x}$ factors). Note that the music has been better classified with GMMs than with linear-GBE. However, the rest of the classes presents a high classification error with GMMs (as we knew from the results presented in the Albayzin evaluation).

An important fact about the distribution of the errors is shown in Table 4.4. The table shows the percentages of the segments that have been correctly classified for GMM with 2048 Gaussians, FA with linear-GBE scoring and FA with IoChD scoring both with 100 channel factors. The table is divided into two columns: the first column shows the percentage of the correctly classified segments between 0 and 3 seconds long. It clearly shows that, while the classification is better with FA systems as we shown in Table 4.3, segments shorter than 3 seconds are better classified with GMMs. The second column shows the percentage of the segments longer than 3 second. It can be seen that the best classification system is based on FA with IoChD scoring. By way of conclusion, the FA is a better classifier if the segments are longer than 3 seconds which is a common fact because most of the segments are between 10 and 20 seconds long and a collar of 1 second is not scored around each reference boundary.

Table 4.4: Percentage of correctly classified segments shorter than 3 sec. and longer than 3 sec. for linear-GBE with 100chnf, the IoChD with 100chnf and the GMM-2048G. The total number of segments is 7754

	Seg. $<$ 3 sec	**Seg. $\geq$ 3 sec**
GMM-2048G	**25.4**	56.9
Lin.GBE - 100chnf	19.2	57.3
IoChnf - 100chnf	23.4	**60.8**

4.4.2 Segmentation-by-Classification Experiments

In this subsection, no oracle segment boundaries are considered so the audio stream is segmented by classifying each window into one of the five classes.

Table 4.5 shows the baseline results for this segmentation task. To be able to compare the results with the best baseline classification system of Table 4.2, the baseline segmentation systems in Table 4.5 are based on GMM with 2048 Gaussians. The first row in this table shows the results of a basic GMM-2048G system. The segments in this system are delimited by the transition of the frame-by-frame classification and no smoothing is applied. Note that the degradation of the GMM-2048G (54.6% of total error) comparing to the the GMM-2048G with perfect segmentation in Table 4.2 (26.2% of total error) where the decision of each class were based on the accumulated likelihood of the whole given segment. These results clearly shows that a smoothing stage to avoid sudden changes in the decision sequence is needed in a segmentation task. A widely used technique to smooth the transitions between classes is the left-to-right HMM topologies. Table 4.5 shows different left-to-right HMM configurations where the 2048G are divided by the number of states to maintain the same number of Gaussians in every configuration. The best baseline system for the segmentation task (33.3% of total error) has 32 states with 64G per each state (keeping a total of 2048G). This result proves the dramatic improvement when a temporal smoothing is applied to segmentation-by-classification systems.

Classification experiments in the last subsection indicate that IoChD scoring is more accurate than linear scoring as we stated in Section 4.3.6. For the sake of clarity, results with the linear scoring are not presented in this subsection.

Unlike the oracle segmentation where the $\boldsymbol{x}$ factors were computed for each segment, in this subsection the $\boldsymbol{x}$ factors are computed for each window so an increment in the

Table 4.5: Baseline for segmentation experiments. The table shows an error per class and total error for GMM-HMM systems over the test files without oracle segment boundaries

GMM/HMM LeftToRight	MU	SP	SM	SN	TOTAL
GMM - 2048G	35.5	59.2	65.0	58.6	54.6
2 ST -1024G	29.9	59.2	54.7	56.8	50.2
4 ST - 512G	26.0	49.8	45.9	50.2	43.0
8 ST - 256G	24.3	49.3	41.6	50.1	41.3
16 ST -128G	17.8	40.2	36.0	43.0	34.2
32 ST - 64G	17.3	39.5	33.9	41.5	**33.3**
64 ST - 32G	15.9	41.6	34.2	42.6	33.6

Table 4.6: Error per class and total error for FA segmentation-by-classification systems. The experiments are computed with one single U for all the classes and one U matrix for each class/non-class using IoChD scoring. No score combination or smoothing was carried out.

IoChD SCORING step-0.5 sec. 100chnf					
	MU	SP	SM	SN	TOTAL
One single U	40.3	76.9	60.5	64.3	**60.5**
One U per class	33.3	45.6	36.2	47.4	**40.6**

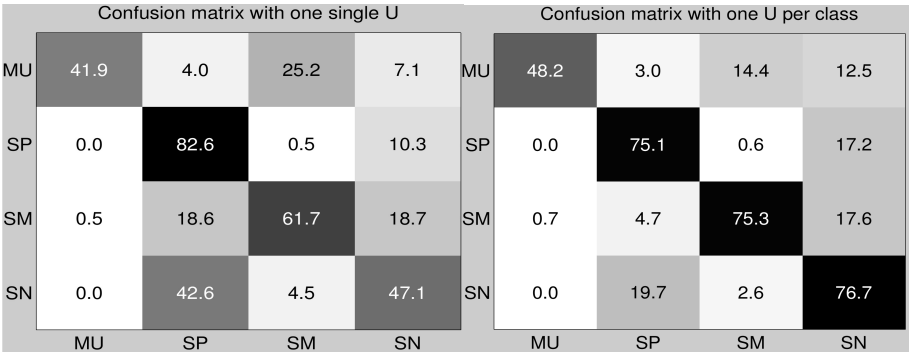


Figure 4.4: Confusion Matrices for the experiments of Table 4.6. Each row of the matrix represents the percentage of frames in an actual class and each column represents the percentage of frames in a predicted class both affected by the collar. One single U for all the classes and one U matrix for each class/non-class are displayed. No score combination or smoothing was carried out.

number of $\boldsymbol{x}$ factors or a reduction of the window-step increases the memory and the time needed to train the models dramatically. As a preliminary experiment, the FA segmentation-by-classification system computes the statistics every 0.5 seconds and 100 $\boldsymbol{x}$ factors. Because the windows (3 seconds long in our experiments) are smaller than the oracle segments, the useful information which describes the class of the window is scarcer. Therefore, a more discriminative scoring is needed and it is provided by the models with one $\boldsymbol{U}$ matrix for each class. Results with a single $\boldsymbol{U}$ matrix for all the classes and one $\boldsymbol{U}$ matrix for each class are presented in Table 4.6. There is a significant improvement in the classes with more data using one $\boldsymbol{U}$ matrix for each class because the CLLR removes the information of the target class in the denominator as we pointed out in Section "Scoring". Figure 4.4 displays the confusion matrices for the experiments of Table 4.6. The percentages have been computed with the frames scored (affected by the collar) divided by all the frames of each class in the reference ($dur(ref_i)$). The table clearly shows less confusion between classes using one $\boldsymbol{U}$ matrix for each class. Although the sensitivity of the SP is lower, the sensitivity of the SN is much higher. Specially, there is a significant reduction in the confusion between SP and SN and a slight reduction in the confusion between MU and SM. The more frequent the class in the data, the more significant the error reduction compared to a single $\boldsymbol{U}$ matrix for all the classes. Accordingly, the total error is reduced around 20%.

Once determined the benefits of the FA system with one $\boldsymbol{U}$ matrix for each class with IoChD scoring, the window-step can be reduced to increase the resolution (0.1 second window-step) at the expense of increasing the computational cost. The number of $\boldsymbol{x}$ factors are not increased because the computation time and the memory grow exponentially. Figure 4.5 shows the scores for each class over a chunk of a test file. The ground truth is plotted in the same figure and it is represented with a square waveform of amplitude 1. The green bars represent the forgiveness collar around each boundary. The color of each score class and the corresponding ground truth is the same. The figure clearly shows that the ratio of the winner class is bigger than zero and corresponds to the ground truth class for most of the frames. The results in Table 4.6 can be compared to the results in Table 4.7 showing a significant error reduction achieved by decreasing the window-step because of the resolution increase.

To avoid sudden changes in the segmentation process, three back-end subsystems are evaluated here. The first back-end system is based on a MAP approach and the two

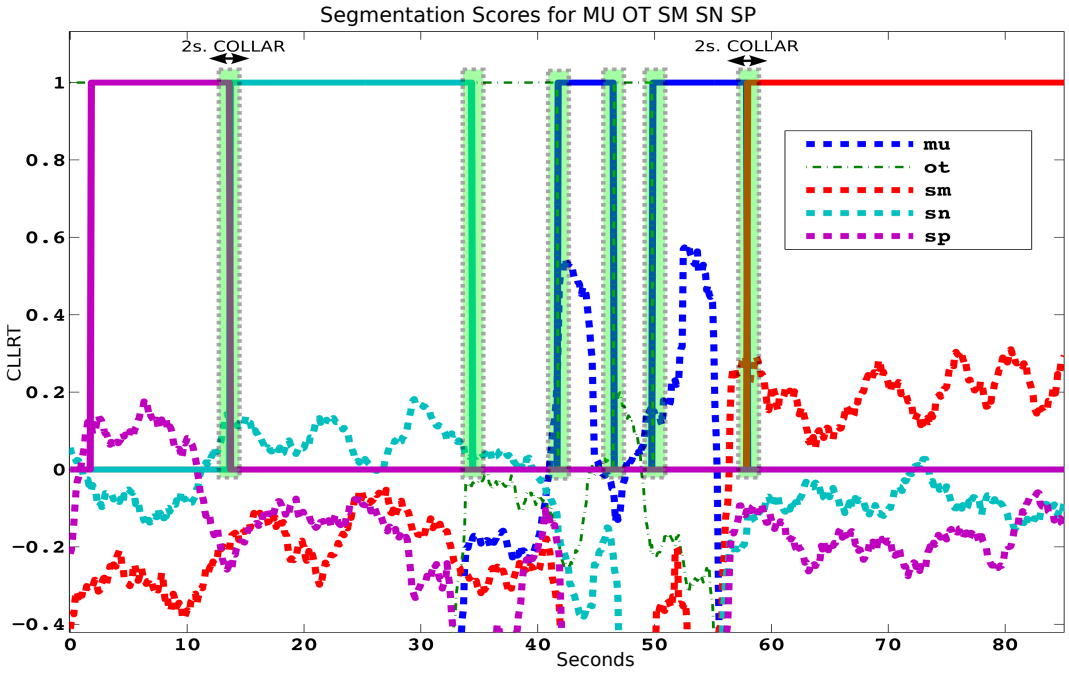


Figure 4.5: Scores and ground truth of each class over a chunk of a test file

Table 4.7: Error per class and total error for FA segmentation-by-classification systems. The experiments are computed with one U matrix for each class/non-class using IoChD scoring. No score combination or smoothing was carried out.

IoChD SCORING step-0.1 sec. 100chnf					
	MU	**SP**	**SM**	**SN**	**TOTAL**
One U per class	27.9	37.9	32.4	40.9	**34.8**

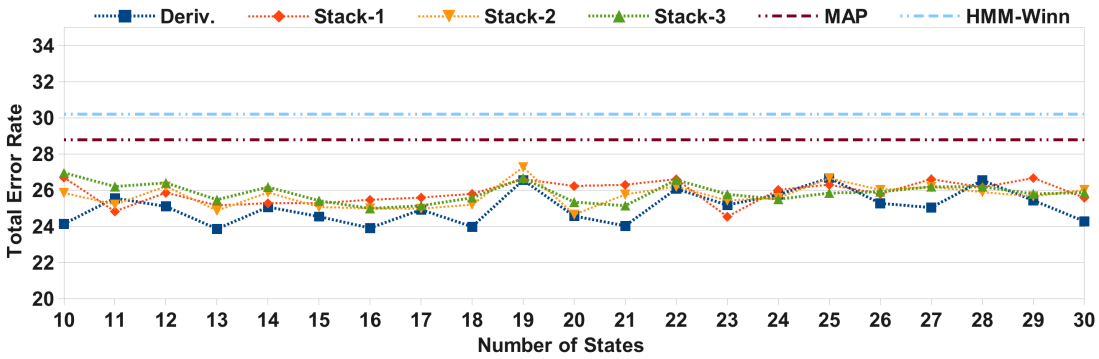


Figure 4.6: HMM/GBE-FA segmentation-by-classification system with different number of states

following systems are very similar but they model the temporal behavior in different ways: on a first step, the scores of both systems are conditioned using a multi-class logistic regression. The dynamic behavior of the scores are extract with the first and second order time derivatives or stacking the past and the futures frames of the scores (we know this systems as *Derivative HMM/GBE* and *Stacking HMM/GBE* respectively). Finally, a left-to-right HMM/GBE with full covariance matrices is used to smooth the scores and improve the results with a scoring combination for both systems. The number of states determines the time spent in an HMM and therefore the minimum length of the segment.

We compare the error of the system proposed in this work with the winner system of the Albayzin-2010 evaluation (Gallardo2010) where 15 MFCCs, frame energy, and their corresponding first and second derivatives are extracted. In addition, the spectral entropy and the Chroma coefficients are calculated. The mean and variance of these features are computed over 1 second intervals creating 122 dimension feature vectors. The segmentation approach chosen is HMM-based. The acoustic modeling is performed using five HMMs with three emitting states and 256 Gaussians per state. Each HMM corresponds to one acoustic class. A hierarchical organization of binary HMM detectors is used. First, audio is segmented into Music/non-Music portions. Second, the non-Music portions are further segmented into Speech-over-music/non-Speech-over-music portions. Finally, the non-Speech-over-music portions are segmented into Speech/Speech over noise.

Figure 4.6 shows the results of the systems described in the previous two paragraphs.

Table 4.8: Results for Albayzin evaluation winner system and Factor Analysis Segmentation system over the test files. The table shows the error per class and the total error with the metric of the evaluation and the NIST metric.

	Error for each class					
	MU	**SP**	**SM**	**SN**	**TOTAL**	**NIST**
HMM-Winn (Gallardo2010)	19.2	39.5	25.0	37.2	30.2	-
HMM-Rep	16.3	40.8	24.0	38.8	30.0	19.3
Worst FA-Segm (25 st.)	19.3	29.5	24.6	33.1	26.6	16.7
Best FA-Segm (13 st.)	18.8	23.7	23.6	29.1	**23.8**	**14.7**

First, two straight lines represent the results for the winner system of the Albayzin-2010 evaluation (Gallardo2010) (30.2% of total error rate) and the FA system with MAP back-end (Castan2013a) (28.8% of total error rate). The behavior of the *Derivative HMM/GBE* and the *Stacking HMM/GBE* back-end systems are plotted in the same figure with a different number of states. The *Stacking HMM/GBE* combines the present frame with one, two and three frames from the past and the future to provide different levels of contextual information. The figure shows a slight improvement in *Derivative HMM/GBE* for almost any number of states. However the result are quite similar for *Derivative HMM/GBE* and *Stacking HMM/GBE*. Both systems for every configuration improve the results of the winner hierarchical-HMM (Gallardo2010) and the MAP-FA system (Castan2013a). Note that the final number of states is not critical because the difference among errors is less than 3%. The best result obtained was an error of 23.8% using 13 states and the worst result was 26.6% with 25 states.

Table 4.8 is divided into two parts: the first part shows the error for each class and the average error for the winner hierarchical-HMM system of the evaluation (HMM-Winn). The last column shows the NIST metric used in the NIST RT Diarization evaluations (NIST2009) to compare the systems with a well-known metric. To be able to compute the NIST error with the hierarchical-HMM system, we replicated the winner system according to (Gallardo2010) (HMM-Rep). The second part of the table shows the FA segmentation-by-classification system (FA-Segm) after the combination and the smoothing with the *Derivative HMM/GBE* back-end subsystem because this subsystem is slightly better than the other back-end subsystems. We choose the best configuration (Best FA-Segm) and the worst configuration (Worst FA-Segm). The hierarchical-HMM

systems perform better than the worst FA system for the MU and SM classes but their behavior is worse for SN and SP. Also, there is not a substantial benefit classifying the MU with the best FA system compared to the hierarchical-HMM system. This is due to the use of specific features to detect the music like the Chroma features. The worst FA systems achieves a relative error reduction of 11.3% respect to the hierarchical-HMM system. Finally, the best FA configuration improves the performance for all the classes and achieves a relative error reduction of 29.2% respect to the hierarchical-HMM system.

4.5 Chapter Summary

This chapter presents a novel system to segment and classify audios into broad classes. The proposed system is based on a factor analysis (FA) approach to compensate the within-class variability with one factor loading matrix per class. Unlike other FA systems (like speaker ID and language recognition), the system proposed in this work does not have well-delimited segments and the nature of the classes can be very different (music, speech, or noise). The relevance of this approach can be summarized in two major aspects: it does not need specific features or hierarchical structure and it performs a very accurate segmentation and classification for all the classes. Therefore, the system is general enough to be used for different tasks and scenarios.

The system has been tested in a broadcast TV news domain to segment and classify into five broad classes. Two sets of experiments have been proposed. The classification experiments with oracle segmentation show a clear improvement compared to the baseline GMM system. A class/non-class FA system is proposed for the segmentation-by-classification experiments in the Section 4.4.2. Different back-end systems have been evaluated in order to exploit the correlation among classes and avoid sudden changes in the decisions. This system is compared to a hierarchical solution with specific features for each level. The results show a significant improvement for all classes, metrics, and configurations achieving a 29.2% relative error reduction with respect to the hierarchical-HMM system for the best configuration.

Chapter 5

Acoustic Concept Detection

Contents

5.1 Chapter Overview

Speech can be considered the most informative part of the audio. However, non-speech sounds can be useful to understand the scene. These sounds are known in the literature as *acoustic concepts* (ACs) or *acoustic events*[1] and can be critical to understand human activities or to describe the scene. For example, human activity produces a variety of sounds coming from the interaction of people with objects that can characterize situations as Chapter 3 has shown. In addition, the detection of these sounds may increase the robustness of the speech recognition systems . For instance, the detection of a specific sound can determine the context of the scene and, therefore, the ASR system can use a specific vocabulary adapted for the situation. Therefore, determining the ACs and their temporal position in the audio signals is under study today. *Acoustic Concept Detection* (ACD) aims at processing a continuous audio stream and determining what concept has been produced and when. Therefore, the system must be able to produce labels to understand the scene behind the concept.

ACD in meeting rooms is a challenging field because the ACs have low SNR and are overlapped with speech or other ACs. The 2007 CLEAR ("Classification of Events, Activities and Relationships") Evaluation (Temko2006b) was performed by the CHIL EU project("Computers in the Human Interaction Loop") on a database recorded in real seminars (five different locations) where the ACs were spontaneously generated. Most of these, are not highlighted and overlapped with other sounds. In this evaluation, the submitted systems showed low accuracies and high error rates. In fact, 5 out of 6 submitted systems showed accuracy below 25% and an error rate above 110% (the winning system (Zhou2008) obtained around 30% accuracy and 99% error rate) where the overlapping segments represent more than 70% of the errors. This problem is related to the "cocktail party" problem where there are two or more sources of speech. However, in our problem, the ACs can be overlapped with speech or with other sounds coming from different sources.

Subsequent investigations have dealt with the overlap problem in different ways. The first attempt was proposed by Temko in his PhD thesis (Temko2009a) at a model level where the author proposed models for isolated sounds and models for overlapped

[1]We will call them Acoustic Concepts (ACs) in this thesis to avoid confusion with Multimedia Events

sounds. Since the meeting rooms in the evaluation are equipped with multiple cameras and multiple microphone arrays, recent approaches propose a fusion information coming from video or multiple audio sources. These systems shows an appreciable improvement in the detection rate of overlapped ACs. In (Butko2010b) the authors propose a multimodal system because some of the ACs have a visual correlate and, therefore, the video modality can be exploited to enhance the detection rate. Furthermore, the authors use multiple microphones to know the position of the AC since some concepts can only occur at particular locations like "door slam". Another popular solution is the separation of overlapped signals with signal processing techniques. In (Chakraborty2013), an approach based on partial signal separation using multiple array beamformers was proposed prior to an HMM-GMM classification system as a solution to attack the problem at signal level.

Since the CLEAR evaluation database was recorded in five different rooms with different furniture, the ACs present some variability that can be compensated. This work studies variability compensation techniques based on factor analysis with one microphone in this meeting room environment. The main goal is to increase the robustness in the classification of the ACs at model level so it does not interfere with multimodal or multichannel techniques that could be applied later. Due to the extremely high error rate shown in the CLEAR evaluation, this chapter proposes a preliminary study where the segmentation is given by the labels to evaluate the classification of the proposed system. Finally, the chapter studies the detection of the ACs in a continuous audio stream.

The sections of this chapter are organized as follows: section 5.2 describes the database and the metric for this task is described in 5.3. The FA framework is described in section 5.4. Section 5.5 shows a comparison of the proposed system with a baseline and, finally, a chapter summary is presented in section 5.7.

5.2 Database

The database consists of multi-sensory audiovisual recordings inside meeting rooms (known as smart rooms) equipped with multiple audio and visual sensors to be able to detect, classify and understand the human activity in the space. The smart rooms are medium-sized conference rooms with supporting computing infrastructure. The

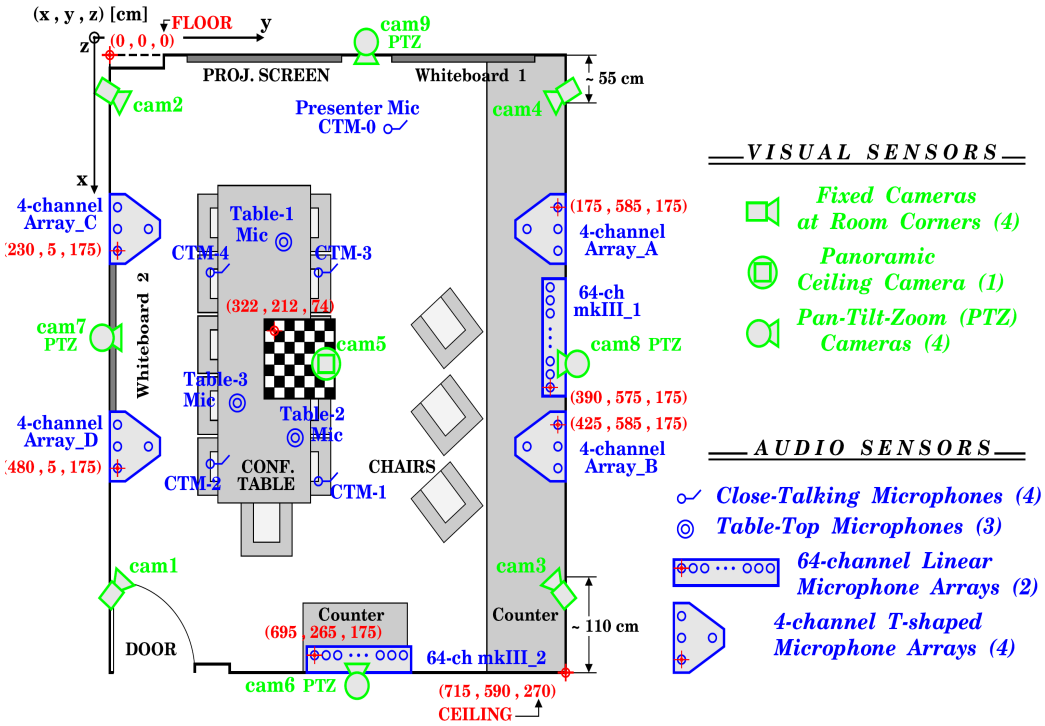


Figure 5.1: Schematic diagram of the IBM smart room described in (Mostefa2008a).

multitude of recording sites provides the desirable variability in the corpus, since the smart rooms obviously differ from each other in their size, layout, acoustic and visual environment. However, all smart rooms have a common hardware setup to produce a homogeneous database across sites, to facilitate technology development. An extensive description of the common software and sensors can be found in (Mostefa2008a) where the authors describe the audiovisual corpus. Figure 5.1 shows an example of one of the five smart rooms used for recording the CLEAR database. All the smart rooms contain a minimum of 88 microphones that capture both close-talk and far-field acoustic data. Since the purpose of this thesis is to study model-based methods to reduce the variability and increase the performance of the classification and segmentation for multimedia documents, we use one microphone located at the center of the meeting rooms (specifically, one of the three table top microphones).

The database used in the CLEAR 2007 evaluation is made up of 25 meetings recorded in five different meeting rooms: AIT (Athens Information Technology), ITC (Instituto Trentino di Cultura), IBM (International Business Machines), UKA (Univer-

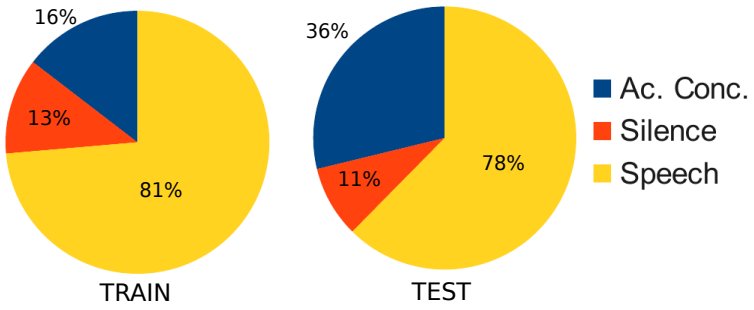


Figure 5.2: Percentage of speech, silence and acoustic concepts for train and test datasets.

sität Karlsruhe), and UPC (Universitat Politècnica de Catalunya). The meetings are divided in 5 lectures (approximately 30 minutes long) and 20 short meetings (5 minutes long). The meetings consist of a presentation to a group of three to five attendees who ask questions during and after the presentation. Not only the main speaker develops activity during the meeting, but also the attendees interact in the scene in terms of entering/leaving the room, opening the door, making noises with objects, speaking among the attendees, etc. Each meeting is comprised of different acoustic scenes: beginning, meeting, coffee break, question/answers, and end.

The development set of the database is made up of one meeting from each location, making a total of 7495 seconds, where 16% of the total time are ACs, 13% silence, and 81% speech. The rest of the meetings (20 5-minute segments) represents the test set, where 36% are ACs, 11% silence, and 78% speech as shown in Figure 5.2. In particular, 64% of the ACs are overlapped with speech and 3% are overlapped with other ACs. These overlapped ACs dramatically increase the difficulty of the task. A database with isolated ACs recorded at UPC (Temko2005) has also been used to get some preliminary results, but these isolated concepts have not been used to train the final system.

The set of ACs is comprised of 12 semantic classes that are: "door knock", "door open/slam", "steps", "chair moving", "spoon/cup jingle", "paper work", "key jingle", "keyboard typing", "phone ring", "applause", "cough", "laugh", "speech", "unknown", and "silence". Table 5.1 shows the ACs in terms of the number of occurrences per concept and the corresponding annotation label. The classes of "speech", "unknown" and "silence" are not evaluated.

Table 5.1: Acoustic concept classes with the corresponding annotation label

Concept name	Label	Train	Test
Knoch in door or table	[kn]	82	152
Door slam	[ds]	73	75
Step	[st]	72	496
Chair moving	[cm]	238	226
Cup jingle	[cl]	28	27
Paper wrapping	[pw]	130	88
Key jingle	[kn]	22	32
Keybord typing	[kt]	72	105
Phone ringing or music	[pr]	21	25
Applause	[ap]	8	13
Cough	[co]	54	36
Laugh	[la]	37	154
Unknown (Unidentified sounds)	[un]	-	-
Speech	[sp]	-	-
Silence	[ ]	-	-

5.3 Metrics

5.3.1 Classification Metric

In the classification experiments, the system has to correctly classify the segment which boundaries are given by the reference labels. The segments where two ACs are overlapped count twice (one for each concept). The error rate for the acoustic concept classification (ACC-ER) can be written as:

$$ACC - ER = \frac{number\ of\ segments\ incorrectly\ classified}{number\ of\ total\ segments} \tag{5.1}$$

5.3.2 Detection Metric

In the detection experiments the boundaries of the acoustic concepts are not given. Two metrics have been defined: an F-score of detection accuracy (ACD-ACC) and an error rate (ACD-ER):

The ACD-ACC measures the detection of all instances of what is considered as a relevant acoustic concept. The ACD-ACC is defined as the harmonic mean between

Precision and Recall. The most relevant aspect of this metric is the detections of the instances and not the temporal resolution. It can be written as:

$$ACD - ACC = \frac{(1+\beta^2) * Precision * Recall}{\beta^2 * Precision + Recall} \tag{5.2}$$

where

$$Precision = \frac{number\ of\ correct\ system\ output}{number\ of\ total\ system\ output} \tag{5.3}$$

$$Recall = \frac{number\ of\ correctly\ detected\ references}{number\ of\ total\ references} \tag{5.4}$$

β is a weighting factor that balances Precision and Recall. In this evaluation $\beta = 1$.

On the other hand, the second metric for detection experiments measures the temporal resolution of the detected acoustic concepts. The ACD-ER scores the ACD as a general audio segmentation task and it is more relevant for content-based audio indexing/retrieval. The ACD-ER can be written as:

$$ACD - ER = \frac{\sum_{seg}\{dur(seg) * (max(N_{ref}, N_{sys}) - N_{correct}(seg)\}}{\sum_{seg}\{dur(seg) * N_{ref}\}} \tag{5.5}$$

where, for each segments seg, the $dur(seg)$ is the duration, N_{ref} is the number of references, N_{sys} is the number of system outputs and $N_{correct}$ is the number of references that correspond to system outputs.

5.4 Factor Analysis Framework

Here, we propose to evaluate the framework based on FA that has been successfully used for segmentation as we were able to see in Chapter 4. In this task, the system has to face several sources of variability as a result of recording the database in five different locations. The main difficulty is that the ACs have low energy and are overlapped with speech or other ACs which makes hard to detect the AC even for humans. In this section we give a brief description of the system for ACD since it is basically the same system used for segmentation.

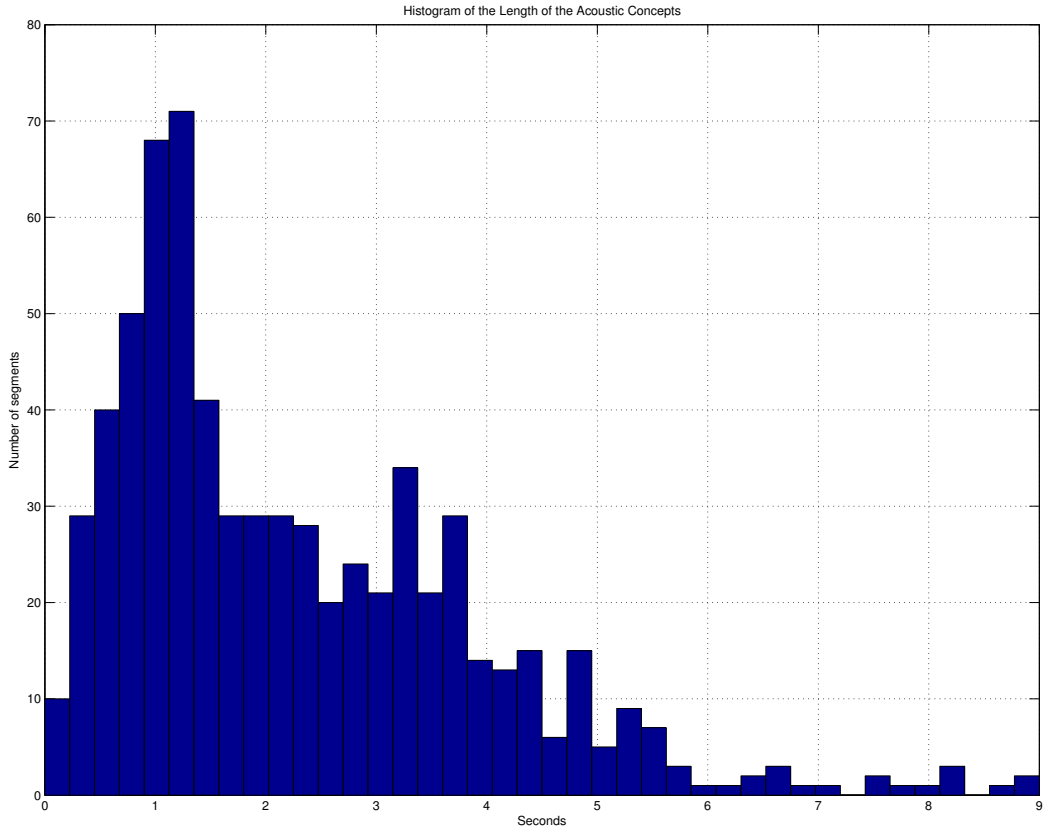


Figure 5.3: Histogram of the length of the ACs

5.4.1 Acoustic features and statistics

We extract 16 MFCCs (including the zeroth order cepstrum) computed in 25 ms frames with a 10 ms frame step, their first and second derivatives. The feature vectors are normalized in mean for each file.

The zeroth and first-order statistics (eq.4.2 and eq.4.3 respectively) of fixed-length windows are computed by using the UBM. The main difference with respect to the segmentation approach is that the fixed-length windows are much shorter because the ACs are very brief. Figure 5.3 shows a histogram of the length of the ACs. We can see that the mode is located around 1 second long but there are a lot of segments even shorter. Therefore a 0.3 second long is the size of the fixed-length windows for this approach instead of 3 seconds used in the segmentation approach.

5.4.2 Models and scoring methods

We use the same FA model as in the last chapter for segmentation:

$$\boldsymbol{m}^s = \boldsymbol{t}^{c(s)} + \boldsymbol{U}\boldsymbol{x}_s, \tag{5.6}$$

where $c(s)$ denotes the class of segment s, $\boldsymbol{t}^{c(s)}$is the class-location vector, $\boldsymbol{U}$ is the *within-class variability matrix* and $\boldsymbol{x}_s$ is a vector of L *segment-dependent-within-class-variability factors* (known as channel factors).

The class/non-class FA models (already described in 4.3.5) are also used as follows:

$$\boldsymbol{T} = [\boldsymbol{t}^{class}, \boldsymbol{t}^{\overline{class}}] \tag{5.7}$$

$$\boldsymbol{U} = \boldsymbol{U}^{class-\overline{class}} \tag{5.8}$$

Finally, we study the detection of the ACs with both scoring methods (the LLR 4.13 and the CLLR 4.14)described in 4.3.6.

5.5 Experimental Results

Three different sets of experiments have been carried out with a clear increase in the difficulty of the task. The first set is comprised of isolated ACs with oracle boundaries that have been generated artificially. The second set verifies the quality of the models to classify the overlapped ACs when the boundaries are given. Unlike previous experiments, the ACs have been generated spontaneously. Finally, the ACs are detected (segmentation and classification) in a continuous audio stream in the third set of experiments. Therefore, the boundaries are not given and difficulty increases dramatically.

5.5.1 Classification of Isolated Acoustic Concepts

The ACs used in these experiments were recorded in the UPC smart room for development. Although the ACs are the same as the CLEAR ACs shown in Table 5.1, these isolated ACs are not used in the posterior experiments because the ACs are not generated in an spontaneous way and they are not overlapped with speech or other ACs. However, this experiment is useful to study the behavior of the proposed system, to set

some parameters and it shows how the errors are distributed with different ACs. The database is divided into three groups: two of them are used to train the model and the third one is used to test.

Table 5.2: Classification error rate for isolated acoustic events with GMM systems

Gaussians	4	8	16	32	64	128	256
ACC-ER %	3.92	3.59	3.26	3.26	3.26	3.26	3.26

A set of GMMs with a different number of Gaussians has been used as baseline system. Table 5.2 shows the error rate given by the metric of eq.(5.1) for each configuration of the GMM. It is apparent from this table that the classification of isolated ACs when they have been artificially produced is an easy task since the error rate is very low for all the configurations. A stable error rate of 3.26% is achieved from 16 Gaussians to 256 Gaussians. It seems reasonable to choose 32 Gaussians as a parameter for the next experiments because the database is not very large and this does not represent a large number of resources.

Therefore, the UBM for the FA approach was also trained with 32 Gaussians over all the train set to be able to compare the results with the baseline. Table 5.3 shows the results of this experiment with FA for 32 Gaussians, 10 channel factors and different values of τ. This parameter is known as *relevance factor* (τ) and it controls the MAP adaptation of the means of the model. If we increase τ to infinite, the MAP will remain in the original UBM. On the other hand, if we decrease τ, the means will be more affected by the new frames. The value of τ must evaluated in this chapter because the ACs present different durations for each class. Therefore, we must determine the value of τ to allow the best MAP adaptation for all the ACs. Table 5.3 compares the scoring methods described in section 4.3.6 for different values of τ. The results obtained in this preliminary analysis show that the CLLR scoring method improves the classification of the artificially generated ACs because the class and non-class data subsets are well delimited and, therefore, the discrimination is maximum. A value of $\tau = 100$ can be defined as a parameter of reference because the ER is minimum for both scoring methods. However, the results are worse than the GMM baseline. A hypothesis of this behavior is the limited number of isolated AC's presented in this database because the

within-class variability matrix needs a representative amount of data to avoid a poor estimation. Therefore, some of the parameters will be analyze again in the experiments with spontaneous generated ACs because the number of occurrences per concept is higher.

Table 5.3: Classification error rate for isolated acoustic events with Factor Analysis with LLR and CLLR scoring methods

τ	30	50	100	250	500
LLR ACC - ER %	6.53	6.53	6.53	7.51	9.15
CLLR ACC - ER %	6.53	6.20	**5.22**	5.55	6.20

Figure 5.4 shows the error for each class with GMM and FA with CLLR scoring for different values of τ. As it can be seen, eight of the ACs are not correctly classified once or more and five of the ACs are correctly classified for all the systems. Furthermore, $\tau = 100$ (yellow bar) shows the best performance of all the FA systems. It is apparent from $\tau = 100$ configuration that only 5 concepts are not correctly classified compared with GMM approach: one [cm], two [co] and three [la]. However, we can conclude that both systems easily classify the isolated concepts because these ACs are artificially generated and there is not overlap with speech or other concepts.

5.5.2 Classification of Spontaneous Acoustic Concepts

Following the procedure presented in the last subsection, this experiment is carried out with oracle segmentation over the CHIL database where the concepts can be overlapped with speech or other concepts dramatically increasing the difficulty. In addition, the audio has been recorded in five different locations which increases the variability of each concept.

Table 5.4 compares the classification of the spontaneously generated oracle segments with LLR and CLLR with different values of τ. The table provides some interesting conclusions. Firstly, the error rate increases dramatically since most of the ACs are masked by speech. The ACs also have low energy because they have been generated spontaneously and they may seem background noises. Secondly, the results are consistent with Table 5.3 because the best result is achieved with $\tau = 100$. However, the LLR

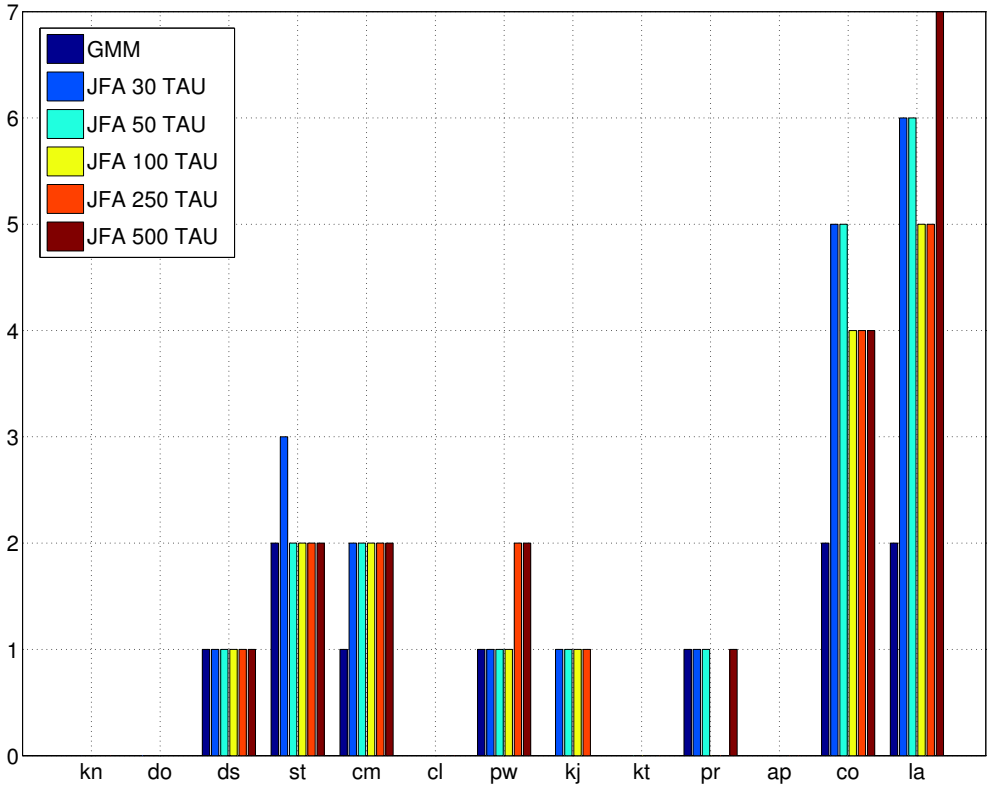
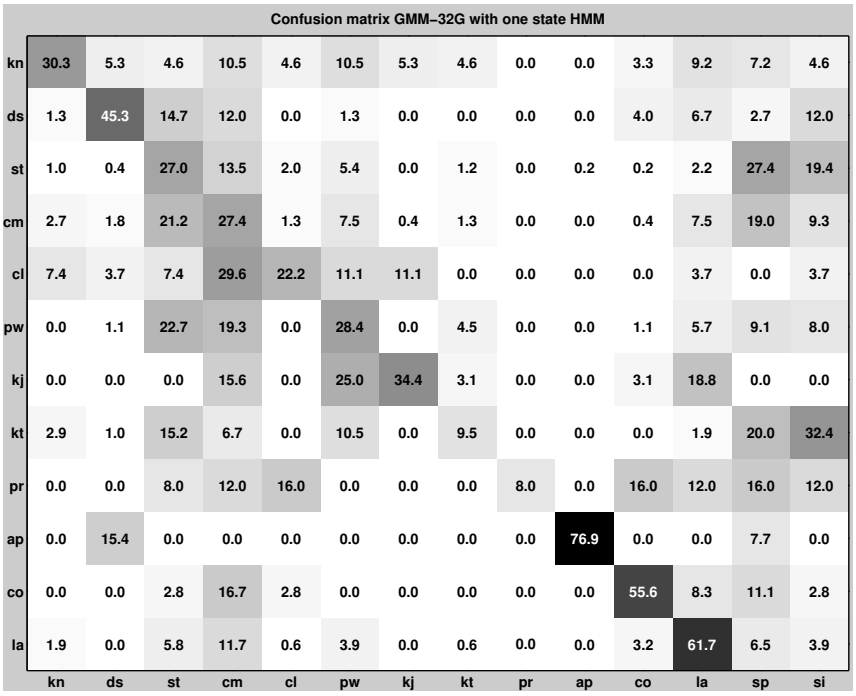


Figure 5.4: Number of errors for each isolated acoustic concept artificially generated

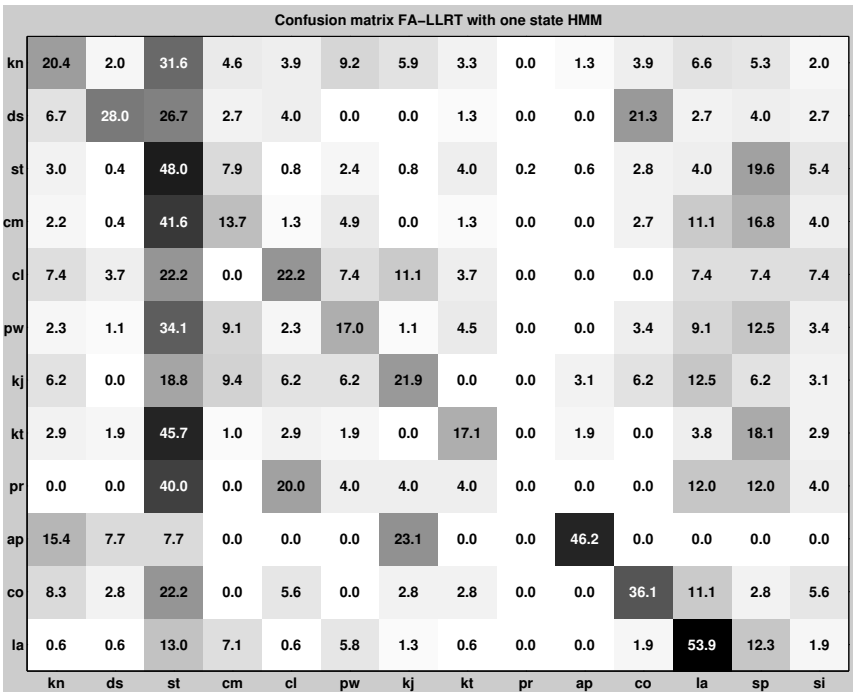
scoring method shows better performance than CLLR. This behavior may occur due to the non-class model because the information of the class model and the non-class model is very similar and the variability is not properly compensated.

The following experiments are computed with $\tau = 100$ because, as it have been shown in the previous experiments, it is the best value for the MAP adaptation of the means of the UBM. Table 5.5 presents the classification error rate for LLR and CLLR with different number of channel factors. As mentioned previously, the within-class variability factor dimension must be smaller than the dimension of the mean vector and, therefore, 10 channel factors were chosen as a previous parameter. The table shows the behavior of our approach for different values of channel factors. The error rate decreases for both metrics with less channel factors showing that there is no need to have a high number of dimensions in the subspace to compensate the variability of the oracle segments.

Table 5.6 compares a baseline based on GMM, with FA system. The parameters



(a) GMM-32G / HMM-1st



(b) FA-LLR-3Chnf / HMM-1st

Figure 5.5: Confusion matrices for (a) GMM-32G / HMM-1st and (b) FA-LLR-3Chnf / HMM-1st. Each row of the matrix represents the percentage of ACs in an actual class and each column represents the percentage of ACs in a predicted class.

Table 5.4: Classification error rate for spontaneously generated ACs with Factor Analysis with LLR and CLLR scoring methods. Most of the ACs are overlapped with speech.

τ	30	50	100	250	500
LLR ACC-ER %	69.41	69.34	**69.13**	69.55	70.25
CLLR ACC-ER %	72.14	72.00	72.28	73.33	74.45
GMM 32G - ER %			68.22		

Table 5.5: Classification error rate for spontaneously generated ACs with Factor Analysis with LLR and CLLR and different number of channel factors.

Num. ChnF	25	10	3
LLR ACC-ER %	73.05	69.13	**67.52**
CLLR ACC-ER %	75.57	72.28	69.62
GMM 32G - ER %		68.22	

for this experiment are the values that were predefined in the previous subsection: 32 Gaussian for GMMs and UBM, $\tau = 100$ for the MAP adaptation and 3 channel factors to model the variability. The first two rows show the error rate for a GMM-32G and the same GMM inside a one-state HMM where the transition probabilities have been estimated with the training labels, slightly improving the results compared to the GMM system. The remaining rows show the error rate for the FA systems with two different scores: the FA-CLLR approaches use eq.4.14 whilst the FA-LLR approaches use eq.4.13. The results clearly show that the FA-LLR approaches are more discriminative than the FA-CLLR since the model and the anti-model share common information due to concepts overlapped with speech. Therefore, the best result is given for the FA-LLR system that slightly improves the final result with the transition probabilities compared to the HMM-GMM.

Finally, Figure 5.5 shows the confusion matrices with the classification percentage in each concept combination for the best baseline system and the best FA system: GMM-32G with one state HMM approach and the FA-LLR with one state HMM approach. Some conclusions can be drawn from these figures. First, the GMM system tends to classify the ACs as "speech" [sp] or "silence" [si] more easily than the FA which shows

Table 5.6: Classification error rate for CHIL acoustic concepts with oracle segmentation

System	Error Rate %
GMM-32G	68.22
GMM-32G / one state HMM	68.15
FA-CLLR-3 Chnf	69.62
FA-CLLR-3 Chnf / one state HMM	69.27
FA-LLR-3 Chnf	67.52
FA-LLR-3 Chnf / one state HMM	**67.17**

that the FA system compensates the variability due to the speech in the overlapped ACs. In addition, the FA system classifies the concept better with more occurrences "Steps" [st]. On the other hand, the concepts "Applause" [ap], "Cough" [co], "Door Slam" [ds] and "Laugh" [la] have been much better classified with GMM. However, a final count shows that the GMM and the FA have been correctly classified 455 and 469 ACs respectively from a total of 1429 ACs.

5.5.3 Detection of Spontaneous Acoustic Concepts

These experiments aim at determining the class of the ACs and their temporal position in a continuous audio signal of a meeting. Therefore the boundaries are not given as in the previous experiments and the results are evaluated with different metrics to measure the detection of the ACs and the temporal resolution.

Table 5.7: Detection of ACs with HMM/GMM systems with a different number of Gaussians and different number of states

N Gauss	8 G		16 G		32 G		64 G		128 G	
N States	ACC	ER	ACC	ER	ACC	ER	ACC	ER	ACC	ER
1	13.0	158.6	14.6	153.2	14.5	154.5	15.1	145.8	15.0	146.3
2	16.9	159.3	18.4	145.2	19.5	141.2	20.1	133.2	20.5	135.5
3	19.6	142.3	20.9	132.4	22.0	135.3	23.1	125.5	24.3	126.9
4	23.8	130.9	25.7	118.2	28.5	110.3	28.2	113.0	30.1	105.6
5	24.9	122.4	27.1	113.6	28.6	106.8	28.6	111.4	29.6	102.1
6	27.6	117.1	29.5	107.7	29.8	107.0	30.2	110.9	28.6	102.0
7	28.4	112.3	28.3	108.3	27.5	110.6	27.7	103.6	26.6	101.1
8	29.1	110.3	30.6	101.7	**31.5**	**99.4**	**30.3**	**97.8**	27.2	100.4

As we have done in the previous experiments, we use an HMM/GMM as a baseline. Table 5.7 presents the accuracy (ACC) and the error rate (ER) described in section 5.3.2. The table scans a different number of states of a left-to-right structure of HMM and a different number of Gaussians per state. The number of states determines the time spent in an HMM and therefore the minimum length of the segment. For example, the HMM with 8 states produces a minimum segment of 80 ms with features extracted every 10 ms as frame step. Almost all the configurations present results with accuracies below 30% and error rates above 100% which shows the difficulty of the task. Moreover, only two configurations (8st-32G and 8st-64G) exceed the standard behavior and both are close to the winning system of the CLEAR evaluation (Zhou2008).

Table 5.8 compares the CLLR and the LLR scoring methods with a FA system with 3 channel factors, a UBM of 32 Gaussians and a $\tau = 100$ since this system achieves the best result for the classification task. The table shows the accuracy and the error rate for different configurations. First, the results are computed without any back-end system to show the performance when the system classifies frame by frame to detect ACs. Then, we use the derivative HMM/GBE back-end system described in section 4.3.7 to combine and smooth the classification performance of the FA system. The table shows the results with the back-end systems with a set of configurations. We scan different values of Gaussians and number of states. A clear improvement can be seen with a high number of states for both scoring methods compared to the system without back-end. The best performances are achieved with CLLR with and without back-end system. The best back-end system configurations is with 8 states and 2 or 4 Gaussians per state. However, any configuration shows accuracies and error rates below the best HMM/GMM system. Therefore, the FA systems shows a clear limitation when the occurrences have low energy and they are overlapped with speech or noises with high energy.

5.6 Limitations of the FA Approach

As we have stated above, this task presents some challenging facts. The ACs can be overlapped with speech or others ACs, the SNR can be very low depending on the position of the microphone, and most of the ACs are very short. The length of the ACs is a critical factor to train FA models properly since the number of parameters can be

Table 5.8: Detection of ACs with FA system with both scoring methods. A back-end system based on GMM/HMM is used with a different number of Gaussians and different number of states

CLLR with UBM=32G CHNF=3 $\tau = 100$										
Results without back-end system: ACC = 12.7% ER = 119.0%										
N Gauss	**1 G**		**2 G**		**4 G**		**8 G**		**16 G**	
N States	**ACC**	**ER**	**ACC**	**ER**	**ACC**	**ER**	**ACC**	**ER**	**ACC**	**ER**
1	4.5	230.1	8.6	186.3	9.7	186.3	8.3	199.4	8.6	190.4
2	7.7	193.7	12.3	164.5	11.9	174.1	10.5	170	12.5	169.7
3	6.8	201.2	12.7	164.5	12.5	156.9	13.0	153.1	13.6	142.5
4	9.2	152.9	13.9	164.3	13.8	148.8	15.9	147.3	15.6	139.8
5	10.3	172.6	14.0	153.9	16.8	144.3	17.6	136.8	17.6	132.7
6	10.9	170.3	16.3	148.9	18.5	132.2	18.2	128.2	18.1	118.8
7	11.7	156.6	18.7	143.9	16.1	132.5	18.1	120.8	18.7	117.2
8	12.6	160.6	**19.3**	**139.1**	18.2	127.8	**16.9**	**119.4**	18.6	119.8
LLR with UBM=32G CHNF=3 $\tau = 100$										
Results without back-end system: ACC = 3.4% ER = 102.0%										
N Gauss	**1 G**		**2 G**		**4 G**		**8 G**		**16 G**	
N States	**ACC**	**ER**	**ACC**	**ER**	**ACC**	**ER**	**ACC**	**ER**	**ACC**	**ER**
1	4.3	232.4	9.0	189.5	9.1	194.2	8.3	198.4	9.6	182.7
2	6.7	199.3	10.8	174.6	12.3	177.5	10.6	189.2	12.0	174.1
3	6.9	196.7	12.6	173.5	11.8	163.9	13.0	162.8	13.6	145.0
4	9.9	159.5	13.9	173.1	13.8	146.0	15.1	143.6	16.2	131.4
5	9.9	169.5	13.4	161.8	15.5	147.0	16.8	136.9	18.0	125.6
6	10.6	171.6	17.4	152.5	16.9	143.3	17.7	136.4	18.8	127.1
7	11.1	159.6	17.4	151.8	17.6	131.7	17.7	131.1	17.2	120.0
8	11.4	159.8	18.0	151.3	17.0	133.6	16.7	124.7	16.2	122.6

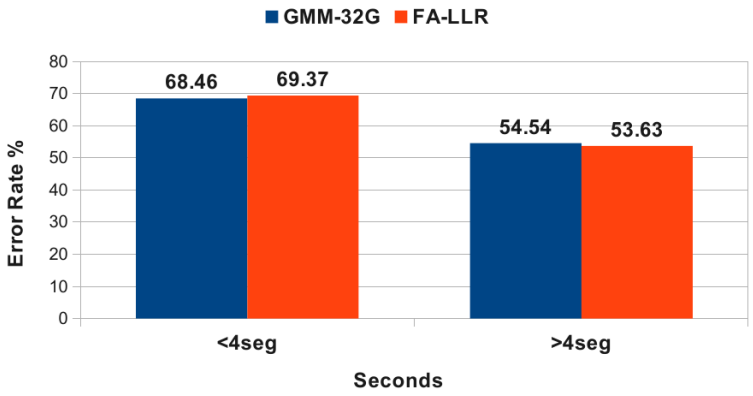


Figure 5.6: Error rate for short segments (from 0 to 4 sec.) and long segments (over 4 sec.)

high. In addition, the larger the number of utterances, the better characterization of the variability.

This section shows the limitations of the FA approach with short segments and how they affect the performance of the system. The best configuration of the FA approach to classify segments has been used to study how the short segments affect the classification accuracy. This configuration is comprised of a 32G-UBM, a $\tau = 100$, 3 channel factors and an HMM of one state as can be seen in table 5.6. Figure 5.6 compares the error rate for short segments (from 0 to 4 seconds) and long segments (over 4 seconds) and it clearly shows that the error rate decreases almost a 15% for long segments. We compare the FA approach with the best GMM baseline configuration with 32 Gaussians and one HMM state. The results are quite similar but it can be seen how the GMM approach classifies the short segments slightly better than the FA approach while the long segments are classified slightly better with the FA approach.

5.7 Chapter Summary

This chapter has presented the classification and the detection of ACs that may happen in a meeting room using the CLEAR evaluation database. Since the database is comprised of tracks recorded in five different locations and the concepts can be overlapped with speech, the ACs present a variability that can be compensated with FA techniques. Three sets of experiments have been carried out in this work. The first

one evaluates the FA system over isolated ACs. This isolated AC database has been used as a development to choose the *relevance factor* (τ) of the MAP adaptation for the FA systems. This experiment also shows that the classification of the isolated AC is not a challenging problem since the error rate is very low because the AC were generated artificially with very high SNR. The second set of experiments evaluates the FA system with spontaneous generated ACs that can be overlapped with speech or other ACs. The proposed system slightly improves the results of a baseline system based on GMM/HMM. The confusion matrices of both systems suggest that the FA system compensates the variability due to the speech in the overlapped ACs. The third set of experiments studies the FA system to detect the ACs when the boundaries are not given. In this case, the FA system shows a limitation as a segmentation-by-classification system because the length of the fixed-length windows is very short (due to the length of the ACs) and may cause the corruption of the models. The baseline also shows very low performance with error rates close to 100% and accuracies below 30% in most of the configurations. Therefore, there is still considerable room for improvement since the classification error is very high. Therefore, further work needs to be done to improve the classification of overlapping sounds with low SNR and the application of FA techniques to the detection problem.

Chapter 6

Conclusions and Future Work

Contents

6.1 Conclusions

In this chapter, we summarize the main conclusions about the work conducted by the author in the area of audio segmentation and acoustic concepts detection as useful resources for the detection of multimedia events. The work presented in this thesis can be categorized into three parts: 1) multimedia event detection, 2) segmentation-by-classification, and 3) acoustic concepts detection.

Multimedia event detection uses different technologies based on video and audio processing to identify events. The audio is processed in different ways such as automatic speech recognition, speaker diarization or unsupervised approaches among others. We have focused our efforts on the part of the audio related to non-speech. For this purpose, we have developed different supervised techniques to extract the information coming from characteristic sounds that we refer to as Acoustic Concepts. Acoustic Concepts are usually overlapped with speech and, therefore, a previous segmentation is needed to delimit the part of the audio with speech, music or other content. However, the audio coming from the videos in the Internet presents a lot of variability because each video is recorded with different cameras in different conditions.

To compensate the variability of the audio, we propose a supervised segmentation-by-classification system based on Factor Analysis techniques to provide segments of speech/non-speech. The system is general enough to be used in different contexts since it does not use specific features. In fact, the system has been tested in a broadcast TV news domain because the audio is comprised of several materials with non-homogeneous style. The system identifies segments of music, speech, speech with noise and speech with music and it shows a better performance than the baseline systems.

Finally, the same approach proposed for segmentation is evaluated to detect acoustic concepts. The system is tested in a meeting room domain because there is an important number of informative sounds and the acoustic concepts have low energy and are heavily overlapped with speech.

The next subsections summarize and report the specific conclusions of each part.

6.1.1 Multimedia Event Detection

The multimedia event detection part shows a comparison between different approaches to detect multimedia events using a set of videos provided in the TRECVid2011 evalua-

tion. These approaches are based on the analysis of the audio of the videos by detecting acoustic concepts. These supervised approaches permit semantic search by the users and help to improve the detection accuracy of video analysis systems.

The acoustic concepts are studied separately by classification, detection and recognition systems. These preliminary experiments show the difficulty of modeling these concepts due to the unconstrained material that can be found in the Internet. The concepts were extracted and annotated from an independent set of data of the TRECVid2011 evaluation. The collection of annotations has two levels of labels: a set of five broad acoustic concepts and a set of twenty specific acoustic concepts. Two extra acoustic concepts (music and speech) are used to produce a realistic segmentation because these extra concepts are present in most of the videos.

This thesis proposes two approaches based on the recognition of the acoustic concepts. Firstly, an *Acoustic Concept Recognition* (ACR) approach is proposed as a natural evolution of a *Segmental-GMM* approach proposed in (Pancoast2012a). The *Segmental-GMM* system creates a feature vector with the likelihood of the acoustic concepts from a GMM model for every acoustic concept extracted every five seconds. On the other hand, the ACR system creates an HMM-GMM model for every acoustic concept to be able to get a segmentation based on the transitions between the models. The systems are compared and the results show similar performance for both approaches.

A second and novel approach is proposed based on the likelihood of the sequences of the audio concepts. This approach is known as *Lattice Count* approach and it has been used before in language recognition tasks. This approach improve the detection of all the events compared to *Segmental-GMM* and *Acoustic Concept Recognition* approaches as it can be seen in Table 3.7. We also compare *Segmental-GMM with twenty acoustic concepts* and *Bag-Of-Audio-Words* approaches described in (Pancoast2012a) and (Pancoast2012) with the *Lattice Count with twenty acoustic concepts* approach. As shown in Table 3.8 and in Figure 3.7, our approach has good behavior for all the events even if the BoAW approach uses one thousand unsupervised clusters.

Finally, a spoken and acoustic concept fusion shows the importance of the acoustic concepts to detect multimedia events. Table 3.9 presents the results of the spoken approach and the acoustic concept approach and both systems show similar performance. Furthermore, both approaches are complementary since the fusion improves the total

result reducing the error by around 5%. The MED task clearly shows the need to improve the segmentation and the classification for unconstrained environments. In these environments the with-in class variability is very high and therefore, a set of systems to compensate the variability that can be found in multimedia documents has been proposed for segmentation and ACR.

6.1.2 Segmentation-by-Classification Approach

The second part of this thesis presents a novel system to segment and classify audios coming from broadcast TV news into five broad classes.

The proposed system is based on a Factor Analysis (FA) approach to compensate the within-class variability with one factor loading matrix per class. This approach is significant in two major aspects: it does not need specific features or a hierarchical structure and it performs a very accurate classification for all classes. Therefore, the system is general enough to be used for different tasks and scenarios. The classification experiments with oracle segmentation (Section 4.4.1) show a clear improvement compared to the baseline HMM/GMM system.

In addition, combining and smoothing back-end is proposed for the segmentation-by-classification experiments (Section 4.4.2) in order to exploit the correlation among classes and avoid sudden changes in the decisions. The final system is compared to a hierarchical solution with specific features for each level. The results show a significant improvement for all classes, metrics and configurations achieving a 29.2% error reduction for the best configuration.

6.1.3 Acoustic Concept Recognition

Finally, the last part of this thesis studies the same approach applied for segmentation in the detection of the acoustic concepts that can happen in a meeting room. The first set of experiments presents a comparison between the GMM/HMM approach and FA approach for isolated acoustic concept that have been artificially generated. Both systems perform an excellent classification showing that the classification of isolated and artificially generated concepts is not a challenging problem because the ACs present very high SNR.

However, acoustic concepts are spontaneously generated in most of the multimedia documents. Furthermore, most of the acoustic concepts that may be useful to describe

activities have low SNR and are overlapped with speech. To study this scenario, the CLEAR evaluation database was used. The database was recorded in five different rooms with different furniture so the ACs present some variability that can be compensated. In a first approach to the problem, we classify the ACs when the boundaries are given to reduce the difficulty of the problem. These experiments show that the FA approach classifies slightly better than a baseline approach based on GMM/HMM.

Finally, the detection experiments aims at determining the class of the ACs and their temporal position. These experiments show a very low performance for every approach denoting the difficulty of the task. In this case, the FA approach presents a limitation because the length of the ACs is very short. Therefore, the estimation of the with-class matrix is inaccurate and the models are corrupted.

6.2 Future Work

Some parts of this thesis have shown a series of limitations that should be studied and dealt with in future works. This section proposes future research lines to study and solve the drawbacks found in this work.

- As we have shown in the last chapter, the ACR with FA has several limitations since the ACs are very short, with low SNR and most of the time they are overlapped with speech or other AC. A possible direction of further research is the study of approaches to train the within-class variability matrix with a small quantity of data or corrupted data. Part of this research should focus on fast adaptation algorithms and part on a set of new artificially generated data to be able to train the models accurately.

- Because the same ACs (i.e. a chair movement) is produced with different sources (different chairs with different materials), a possible research line is the cross-site acoustic concept detection using different datasets from different places to train and test. This scenario would be more realistic and therefore more suitable for being proposed in the MED task.

- The segmentation approach proposed in this thesis was cited as a future work point in (Butko2011c) and a clear improvement has been shown with respect to previous solutions. Following the same line, a total variability approach can

be used to identify the different classes in a total variability space. The audio segmentation can be addressed with the same back-ends proposed in this thesis but using ivectors as features.

- New research lines about MED tend to use unsupervised approaches based on feature clustering. However, the unsupervised approaches are still far away from the supervised approaches. Moreover, the use of acoustic concepts allows a semantic search that can be done with unsupervised approaches. Further research should focus on the fusion of both techniques because the philosophy of both solutions is complementary. In addition, solutions based on artificial neural networks (ANN) could improved the performance of these systems since these algorithms have shown an excellent behavior for classification tasks.

Publications

Journal Article

- D. Castán, A. Ortega, A. Miguel and E. Lleida **Audio Segmentation-by-Classification Approach Based on Factor Analysis in Broadcast News Domain.** EURASIP Journal on Audio, Speech, and Music Processing, June 2014

International Conference Papers

- D. Castán and M. Akbacak **Indexing Multimedia Documents with Acoustic Concept Recognition Lattices.** 14th Annual Conference of the International Speech Communication Association, Interspeech-2013, Lyon, France, August 2013
- D. Castán, A. Ortega, A. Miguel and E. Lleida **Broadcast News Segmentation with Factor Analysis System.** Workshop on Speech, Language and Audio in Multimedia, SLAM-2013, Marseille, France, August 2013
- D. Castán and M. Akbacak **Segmental-GMM Approach based on Acoustic Concept Segmentation.** Workshop on Speech, Language and Audio in Multimedia, SLAM-2013, Marseille, France, August 2013
- D. Castán, A. Ortega, J. Villalba, A. Miguel and E. Lleida **Segmentation-by-Classification System based on Factor Analysis.** IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-2013, Vancouver, Canada, May 2013

- J. van Hout, M. Akbacak, D. Castán, E. Yeh and M. Sanchez **Extracting Spoken and Acoustic Concepts for Multimedia Event Detection.** IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-2013, Vancouver, Canada, May 2013

- D. Castán, C. Vaquero, A. Ortega, D. Martínez, J. Villalba, A. Miguel and E. Lleida **Hierarchical Audio Segmentation with HMM and Factor Analysis in Broadcast News Domain.** 12th Annual Conference of the International Speech Communication Association, Interspeech-2011, Florence, Italy, September 2011

National Conference Papers

- D. Castán, A. Ortega, A. Miguel and E. Lleida **A preliminary study of Acoustic Events Classification With Factor Analysis in Meeting Rooms.** Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH-2014, Las Palmas de Gran Canaria, Spain, November, 2014

- D. Castán, A. Ortega, A. Miguel and E. Lleida **Factor Analysis Segmentation and Classification in Broadcast News Domain.** Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH-2012, Madrid, Spain, November, 2012

- D. Castán, A. Ortega, C. Vaquero, A. Miguel and E. Lleida **Speech/Music classification by using the C4.5 decision tree algorithm.** II Iberian SLTech Workshop, FALA-2010, Vigo, Spain, November, 2010

Evaluation Papers

- D. Castán, M. Rodríguez, A. Ortega, C. Orrite and E. Lleida **ViVoLab and CVLab - MediaEval 2014: Violent Scenes Detection Affect Task.** MediaEval 2014 Workshop, Barcelona, Spain, October, 2014

- D. Castán, A. Ortega, C. Vaquero, A. Miguel and E. Lleida **ViVoLab-UZ Audio Segmentation System for Albayzín Evaluation 2010.** II Iberian SLTech Workshop, FALA-2010, Vigo, Spain, November, 2010

Appendix A

Factor Analysis Training Process

A.1 EM Algorithm

We briefly explain the EM algorithm because it is the algorithm that we have used in all the models to train the parameters. It is a maximum likelihood (ML) optimization technique, in which the log-likelihood of the data given a set of model parameters is increased in each iteration. For the details see (Bishop2006).

The EM algorithm is an algorithm to maximize the likelihood of probabilistic models having latent variables. Consider a dataset of observed variables by $\mathbf{X}$, and all of the hidden variables by $\mathbf{Z}$. Our goal is to maximize likelihood of the data given the parameters $\mathbf{\Theta}$. The likelihood function is given by

$$p(X|\Theta) = \sum_{Z} p(X, Z|\Theta). \tag{A.1}$$

We assume that the latent variables are discrete, but the derivation is identical if $\mathbf{Z}$ is continuous, or combination of discrete and continuous, with summation replaced by integration as appropriate.

Let us express now the log-likelihood function of A.1 as

$$\ln p(X|\Theta) = \ln \sum_{Z} p(X, Z|\Theta) = \mathcal{L}(q, \Theta) + KL(q||p), \tag{A.2}$$

where we have defined the *lower bound* of the log-likelihood function as

$$\mathcal{L}(q, \Theta) = \sum_{Z} q(Z) \ln \frac{p(X, Z|\Theta)}{q(Z)}, \tag{A.3}$$

and

$$KL(q||p) = -\sum_{Z} q(Z) \ln \frac{p(Z|X,\Theta)}{q(Z)}, \tag{A.4}$$

is the Kullback-Leibler (KL) divergence between $q(Z)$ and the true posterior distribution, where $q(Z)$ is a distribution over the latent variables. We can observe that for any choice of $q(z)$, equation A.2 holds. Recall that $KL(q||p) \geq 0$, with equality if, and only if, $q(Z) = p(Z|X,\Theta)$, and thus it follows that $\mathcal{L}(q,\Theta) \leq \ln p(X|\Theta)$, so $\mathcal{L}(q,\Theta)$ is a lower bound of $\ln p(X|\Theta)$. The same conclusion can be reached by making use of *Jensen's inequality*,

$$\begin{aligned} \ln p(X|\Theta) &= \ln \sum_{Z} p(X,Z|\Theta) = \ln \sum_{Z} q(Z) \frac{p(X,Z|\Theta)}{q(Z)} \\ &\geq \sum_{Z} q(Z) \ln \frac{p(X,Z|\Theta)}{q(Z)} = \mathcal{L}(q,\Theta), \end{aligned} \tag{A.5}$$

with equality only when $q(Z) = p(Z|X,\Theta)$ and the KL divergence previously defined goes to zero.

The main assumption of the EM algorithm is that the direct optimization of $p(X|\Theta)$ is difficult, but the optimization of the complete-data likelihood function $p(X,Z|\Theta)$ is easier. The EM is a two-stage iterative process that maximizes the data likelihood as follows:

- E step: in this stage we start from the previous values for Θ, Θ^{old}, and the lower bound $\mathcal{L}(q,\Theta)$ is maximized with respect to function q, holding Θ^{old} fixed. Given that $\ln p(X|\Theta)$ does not depend on q, the maximum of $\mathcal{L}(q,\Theta)$ occurs when $q = p$ and then $KL(q||p) = 0$. At this point the lower bound is equal to the log-likelihood.

- M step: in this step the distribution $q(Z)$ is held fixed and the lower bound $\mathcal{L}$ is maximized with respect to Θ, to obtain a new estimate of Θ, Θ^{new}. The maximization causes the lower bound to increase, which necessarily makes the log-likelihood of the incomplete dataset, $p(X|\Theta)$, to increase. The reason is that the KL will also increase, because q is held fixed, but now it will not equal the new posterior distribution $p(Z|X,\Theta^{new})$, and the KL will not be zero. The increase in the log-likelihood is greater than the increase in the lower bound.

As a summary, in the E step we calculate the posterior distribution $q(Z) = p(Z|X, \Theta)$, and in the M-step, it can be seen if we substitute this amount into A.3,

$$\begin{aligned} \mathcal{L}(q, \Theta) = \sum_{Z} p(Z|X, \Theta^{old}) \ln p(X, Z|\Theta) - \sum_{Z} p(Z|X, \Theta^{old}) \ln p(Z|X, \Theta^{old}) \\ = \mathcal{Q}(\Theta, \Theta^{old}) + const, \end{aligned} \tag{A.6}$$

that the maximization of the lower bound is equivalent to the maximization of the expectation of the complete data log-likelihood with respect to the posterior probability of Z given X and Θ^{old}, $\mathcal{Q}$, because the constant part of A.6 is independent of Θ. Note that if the joint distribution $p(X, Z|\Theta)$ is a member of the exponential family, or product of such members, the logarithm will cancel the exponential and lead to an M step typically much simpler than the maximization of the incomplete data log-likelihood $p(X|\Theta)$.

A.2 EM for JFA

The JFA model is also trained via ML and there is no closed form solution for the derivatives of the log-likelihood function over the parameters. The adopted solution is again the application of the EM algorithm.

The first assumption made in JFA is that our D-dimension observations, $\mathbf{O}$ follow a GMM distribution with K components, with weights $\omega_1, ..., \omega_K$, means $\mu_1, ..., \mu_K$, and covariance matrices $\Sigma_1, ..., \Sigma_K$. The means are concatenated to obtain a single supervector, $\mu = [\mu_1^T, ..., \mu_K^T]^T$, that has dimension KD. This supervector is not fixed but can vary from utterance to utterance.

The log-likelihood function of this model is given by equation 2.7, but note that we have a different models of our means, and what we obtain with equation 2.7 is the log-likelihood of the data given the hidden variable X. By setting the q function in equation A.2 to the true posterior distribution of the hidden variables, in this case to the true frame alignment that generated each frame or the true responsibilities of each Gaussian, the KL divergence vanishes, and we can express the conditional log-likelihood

of one utterance as per equation A.3,

$$\begin{aligned}
\ln p(O|X) &= \sum_{t=1}^{T} log \sum_{k=1}^{K} \omega_k p(o_t|k, X) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} p(k|o_t, X) \ln p(o_t, k|X) - \sum_{t=1}^{T}\sum_{k=1}^{K} p(k|o_t, X) \ln p(k|o_t, X) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} p(k|o_t) \ln p(o_t|k, X) - \sum_{t=1}^{T}\sum_{k=1}^{K} p(k|o_t) \ln \frac{p(k|o_t)}{p(k)} \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \ln p(o_t|k, X) + \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \ln \frac{\gamma_k(t)}{\omega_k} \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_k(t) \ln p(o_t|k, X) + const_1,
\end{aligned} \tag{A.7}$$

where the constant reflects the terms independent of O. However, it is important to note that we do not really take the true responsibilities, but keep fixed the alignments given by the UBM, and they are independent of X. This makes the KL divergence be greater than 0 and the calculated $\ln p(O|X)$ will therefore be an approximation to the true $\ln p(O|X)$. In fact it will be a lower bound because the KL divergence is always non-negative. Nevertheless, if alignments are accurate enough, as it is normally the case, the approximation will be good.

To apply the EM algorithm we first identify our objective Q function from A.6, with $Z = X$, $X = O$, and $\Theta_l = \{\omega_l, t_l, \Sigma_l, U_l\}$. Note that we make explicit the dependence on l, to train different parameters for each class. However, in practice, the means are obtained with relevance MAP as stated previously and are not adapted. Also the weights and the covariances are normally kept fixed and equal to the UBM weights and covariances, and in case they are adapted they are initialized to the UBM values. Normally, a single U is trained for all the classes (of course not in the proposed solution in this thesis). But let's start with the general case, where all the hyperparameters are adapted and there is a different U for each class, and we will comment the particularities

later. The Q_l objective function of our EM algorithm for classs l is

$$
\begin{aligned}
Q_l(\Theta,\Theta^{old}) &= \sum_Z p(Z|X,\Theta^{old}) \ln p(X,Z|\Theta) = \sum_{s\in l} \int q_s(x) \ln p(O_s,x) dx \\
&= \sum_{s\in l} \int q_s(x) \ln\{p(O_s|x)p(x)\} dx = \sum_{s\in l} \int q_s(x)\{\ln p(O_s|x) + \ln p(x)\} dx \\
&= \sum_{s\in l} \int q_s(x)\{\sum_{t=1}^{T_s}\sum_{k=1}^{K} \gamma_k(t) \ln \mathcal{N}(o_{st}; t_{lk} + U_k x, \Sigma_k) + \ln \mathcal{N}(x;0,I)\} dx \\
&= \sum_{s\in l}\sum_{k=1}^{K} N_k(s) \ln |\Sigma_{lk}|^{-\frac{1}{2}} - \frac{1}{2} tr(\sum_{k=1}^{K} S_k(s)\Sigma_{lk}^{-1}) \\
&+ \frac{1}{2}\sum_{k=1}^{K} F_k(s)^T \Sigma_{lk}^{-1} t_{lk} + \frac{1}{2} tr(E_X[x(s)] \sum_{k=1}^{K} F_k(s)^T \Sigma_{lk}^{-1} U_{lk}) \\
&+ \frac{1}{2}\sum_{k=1}^{K} t_{lk}\Sigma_{lk}^{-1} F_k(s) + \frac{1}{2} tr(E_X[x(s)] \sum_{k=1}^{K} U_{lk}\Sigma_{lk}^{-1} F_k(s)) \\
&- \frac{1}{2}\sum_{k=1}^{K} N_k(s) t_{lk}^T \Sigma_{lk}^{-1} t_{lk} - \frac{1}{2} tr(E_X[x(s)] \sum_{k=1}^{K} N_k(s) t_{lk}^T \Sigma_{lk}^{-1} U_{lk}) \\
&- \frac{1}{2} tr(E_X[x(s)]^T \sum_{k=1}^{K} N_k(s) U_{lk}^T \Sigma_{lk}^{-1} t_{lk}) - \frac{1}{2} tr(E_X[x(s)x(s)^T] \sum_{k=1}^{K} N_k(s) U_{lk}^T \Sigma_{lk}^{-1} U_{lk}) \\
&- \frac{1}{2} tr(E_X[x(s)x(s)^T]) + const_2
\end{aligned}
\tag{A.8}
$$

with s being the index of the file and T_s the number of frames in file s. In the E step we calculate the posterior distribution of the hidden variable given the observation, that

is, $p(X|O)$

$$\begin{aligned}
\ln p(X|O) = \ln p(O,X) - \ln p(O) &= \ln p(O|X)p(X) + const_3 \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_k(t)\ln p(O|k,X) + \ln p(X) + const_3 \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_k(t)\ln \mathcal{N}(o_{st}; t_{lk} + U_k x, \Sigma_k) + \ln \mathcal{N}(x;0,I) + const_3 \\
&= \sum_{k=1}^{K}\frac{1}{2}F_k(s)^T\Sigma_{lk}^{-1}U_{lk}x(s) + \sum_{k=1}^{K}\frac{1}{2}x(s)^T U_{lk}^T\Sigma_{lk}^{-1}F_k(s) \\
&\quad - \sum_{k=1}^{K}\frac{1}{2}N_k(s)x(s)^T U_{lk}^T\Sigma_{lk}^{-1}t_{lk} - \sum_{k=1}^{K}\frac{1}{2}N_k(s)t_{lk}^T\Sigma_{lk}^{-1}U_{lk}x(s) \\
&\quad - \sum_{k=1}^{K}\frac{1}{2}(U_{lk}x(s))^T\Sigma_{lk}^{-1}U_{lk}x(s) - \frac{1}{2}x(s)^T x(s) + const_4 \\
&= \sum_{k=1}^{K}\{x(s)^T U_{lk}^T\Sigma_{lk}^{-1}F_k(s) - N_k(s)x(s)^T U_{lk}^T\Sigma_{lk}^{-1}t_{lk} \\
&\quad - \frac{1}{2}(U_{lk}x(s))^T\Sigma_{lk}^{-1}U_{lk}x(s)\} - \frac{1}{2}x(s)^T x(s) + const_4 \\
= x(s)^T\{\sum_{k=1}^{K}U_{lk}^T\Sigma_{lk}^{-1}(F_k(s) - N_k(s)t_{lk})\} &- \frac{1}{2}x(s)^T\{\sum_{k=1}^{K}(N_k(s)U_{lk}^T\Sigma_{lk}^{-1}U_{lk}) + I\}x(s) + const_4,
\end{aligned} \tag{A.9}$$

where we can identify a Gaussian with

$$p(X|O) \sim \mathcal{N}(E_X[x], L^{-1}), \tag{A.10}$$

being for utterance $\boldsymbol{s}$

$$E_X[x] = L^{-1}\sum_{k=1}^{K}\{U_{lk}^T\Sigma_{lk}^{-1}(F_k(s) - N_k(s)t_{lk})\}, \tag{A.11}$$

with the symbol E_X referring to the expectation taken with respect to x, and

$$L = \sum_{k=1}^{K}(N_k(s)U_{lk}^T\Sigma_{lk}^{-1}U_{lk}) + I, \tag{A.12}$$

and we define

$$E_X[xx^T] = L^{-1} + E_X[x]E_X[x]^T. \tag{A.13}$$

In the M step the Q function is derived with respect to the parameters Θ_l and made equal to zero. Then we can obtain an expression for each of the parameters. It is important to keep in mind that the frame alignments are kept constant during the whole process. Also, the weights will not be adapted. We can adapt the means t_{lk} as

$$\frac{\partial}{\partial t_{lk}} Q(t_{lk}, t_{lk}^{old}) = \sum_{s \in l} \Sigma_{lk}^{-1} F_k(s) - N_k(s) \Sigma_{lk}^{-1} t_{lk} - N_k(s) \Sigma_{lk}^{-1} U_{lk} E_X[x(s)] = 0 \quad \text{(A.14)}$$

$$t_{lk} = \frac{1}{J_l} \sum_{s=1}^{J_l} (\frac{F_k(s)}{N_k(s)} - U_{lk} E_X[x(s)]), \quad \text{(A.15)}$$

being J_l the number of utterances of class l. However the means are rarely adapted. The update of the subspace matrix U_{lk} is

$$\frac{\partial}{\partial U_{lk}} Q(U_{lk}, U_{lk}^{old}) = \sum_{s \in l} \Sigma_{lk}^{-1} F_k(s) E_X[x(s)]^T - N_k(s) \Sigma_{lk}^{-1} t_{lk} E_X[x(s)]^T - N_k(s) U_{lk} E_X[x(s) x(s)^T]^T = 0 \quad \text{(A.16)}$$

$$U_{lk} = C_{lk} A_{lk}^{-1}, \quad \text{(A.17)}$$

where

$$C_{lk} = \sum_{s \in l} (F_k(s) - N_k(s) t_{lk}) E_X[x(s)]^T \quad \text{(A.18)}$$

$$A_{lk} = \sum_{s \in l} E_X[x(s) x(s)^T] N_k(s). \quad \text{(A.19)}$$

An alternative commonly used is to estimate a unique U_k for all classes. The update formula is a weighted average of the U_{lk} of the individual classes

$$U_k = (\sum_l C_{lk})(\sum_l A_{lk})^{-1}. \quad \text{(A.20)}$$

Finally, we obtain the following update formula for Σ_{lk}, which in practice is seldom

adapted

$$\begin{aligned}
&\frac{\partial}{\partial t_{lk}} Q(\Sigma_{lk}, \Sigma_{lk}^{old}) \qquad \text{(A.21)}\\
&= -\frac{1}{2}\sum_{s\in l} N_k(s)\Sigma_{lk}^{-1} + \sum_{s\in l}\{\Sigma_{lk}^{-1} S_k(s)\Sigma_{lk}^{-1} - \frac{1}{2} I \circ (\Sigma_{lk}^{-1} S_k(s)\Sigma_{lk}^{-1})\}\\
&\quad -\frac{1}{2}\sum_{s\in l}\Sigma_{lk}^{-1} F_k(s) t_{lk}^t \Sigma_{lk}^{-1} - \frac{1}{2}\sum_{s\in l}\Sigma_{lk}^{-1} F_k(s) E_X[x(s)]^T U_{lk}^T \Sigma_{lk}^{-1}\\
&\quad -\frac{1}{2}\sum_{s\in l}\Sigma_{lk}^{-1} t_{lk} F_k(s)^t \Sigma_{lk}^{-1} - \frac{1}{2}\sum_{s\in l}\Sigma_{lk}^{-1} U_{lk} E_X[x(s)] F_k(s)^T \Sigma_{lk}^{-1}\\
&\quad +\frac{1}{2}\sum_{s\in l} N_k(s)\Sigma_{lk}^{-1} t_{lk} t_{lk}^t \Sigma_{lk}^{-1} + \frac{1}{2}\sum_{s\in l} N_k(s)\Sigma_{lk}^{-1} t_{lk} E_X[x(s)]^T U_{lk}^T \Sigma_{lk}^{-1}\\
&+\frac{1}{2}\sum_{s\in l} N_k(s)\Sigma_{lk}^{-1} U_{lk} E_X[x(s)] t_{lk}^T \Sigma_{lk}^{-1} + \frac{1}{2}\sum_{s\in l} N_k(s)\Sigma_{lk}^{-1} U_{lk} E_X[x(s)x(s)^T] U_{lk}^T \Sigma_{lk}^{-1} = 0,
\end{aligned}$$

where the symbol $\circ$ means the *Hadamard* or entry-wise product, and we have applied the following identities

$$\frac{\partial}{\partial C}\ln|C| = (C^{-1})^T \qquad \text{(A.21)}$$

$$\frac{\partial}{\partial C} tr(C^{-1}E[xx^T]) = -2C^{-1}E[xx^T]C^{-1} + I \circ (C^{-1}E[xx^T]C^{-1}) \qquad \text{(A.22)}$$

$$\frac{\partial}{\partial C} tr(AC^{-1}B) = -(C^{-1}BAC^{-1})^T = -C^{-T}A^T B^T C^{-T}. \qquad \text{(A.23)}$$

Then, the update formula for Σ_{lk} is

$$\begin{aligned}
\Sigma_{lk} = \frac{1}{J_l}\sum_{s\in l} &\frac{S_k(s) - F_k(s)(t_{lk}^T + E_X[x(s)]^T U_{lk}^T) - (t_{lk} + U_{lk}E_X[x(s)])F_k(s)^T}{N_k(s)}\\
&+t_{lk}(t_{lk}^T + E_X[x(s)]^T U_{lk}^T) + U_{lk}(E_X[x(s)]t_{lk}^T + E_X[x(s)x(s)^T]U_{lk}^T),
\end{aligned} \qquad \text{(A.24)}$$

and we have assumed that the off-diagonal terms of $S_k(s)\Sigma_{lk}^{-1}$ are much smaller than the diagonal terms, and the approximation

$$I \circ (\Sigma_{lk}^{-1} S_k(s)\Sigma_{lk}^{-1}) \approx \Sigma_{lk}^{-1} S_k(s)\Sigma_{lk}^{-1} \qquad \text{(A.25)}$$

is applied. In practice, we use the UBM covariance matrices, Σ_k, for all the classes.

A.3 EM with Minimum Divergence for JFA

EM with minimum divergence (MD) is a second approach to implement the EM algorithm. It is derived from expressing equation A.3 as

$$\begin{aligned}\mathcal{L}(q,\Theta) = \sum_Z q(Z)\ln\frac{p(X,Z|\Theta)}{q(Z)} &= \sum_Z q(Z)\ln\frac{p(X|Z,\Theta_1)p(Z|\Theta_2)}{q(Z)} \\ &= \sum_Z q(Z)\ln p(X|Z,\Theta_1) - \sum_Z q(Z)\ln\frac{q(Z)}{p(Z|\Theta_2)} \\ &= E_{q(Z)}[\ln p(X|Z,\Theta_1)] - D_{KL}(q(Z)||p(Z|\Theta_2)) \\ = E_{p(Z|X,\Theta^{old})}[\ln p(X|Z,\Theta_1)] - D_{KL}(p(Z|X,\Theta^{old})||p(Z|\Theta_1)) \\ = E_{p(Z|X,\Theta^{old})}[\ln p(X|Z,\Theta_2)] - D_{KL}(p(Z|X,\Theta^{old})||p(Z|\Theta_2)),\end{aligned} \tag{A.26}$$

where Θ_1 and Θ_2 are two disjoint subset of parameters, and the model is said to be *overparametrized*, because $\Theta_1 \equiv \Theta 2$, and $\mathcal{L}(\Theta_1) = \mathcal{L}(\Theta_2)$, that is, there exists redundant parametrizations which are equivalent. In general, two models are equivalent if

$$p(X|\Theta_1) = p(X|\Theta_2). \tag{A.27}$$

In our case, being $X = O$ and $Z = X$, $\Theta_{l1} = \{t_{l1}, \Sigma_{l1}, U_{l1}, \Pi_1\}$, where $\Pi_1 = \{\mu_{X_1} = 0, \Sigma_{X_1} = I\}$ are the hyperparameters of the prior distribution $p(X|\Pi_1)$, which are kept fixed with mean equal to zero and covariance matrix equal to identity, and $\Theta_{l2} = \{t_{l2}, \Sigma_{l2}, U_{l2}, \Pi_2\}$, where $\Pi_2 = \{\mu_{X_2}, \Sigma_{X_2}\}$ is also updated. Hence we have a transformation of the distribution $p(X_2|\Theta_2)$ to obtain $p(X_1|\Theta_1)$, and the variables X_1 with standard normal distribution and X_2 are related as

$$X_1 = \phi(X_2) = P^{-1}(X_2 - \mu_{X_2}), \tag{A.28}$$

or equivalently

$$X_2 = \phi^{-1}(X_1) = P(X_1 + \mu_{X_2}), \tag{A.29}$$

where $\Sigma_{X_2} = PP^T$, that is, X_2 is normalized by its mean and of covariance's square root to obtain X_1, and according to the fundamental theorem of Calculus and to the chain rule, their probability density functions are related as

$$f_{X_1|\Pi_1}(x_1) = f_{X_2|\Pi_2}(\phi^{-1}(x_1))\frac{\partial\phi^{-1}(x_1)}{\partial x_1} = p_{X_2}(\phi^{-1}(x_1|\Pi_2))P. \tag{A.30}$$

This is one of the two sufficients conditions for equivalence. The other is

$$p(O|X_1, U_1) = p(O|\phi^{-1}(X_1), U_2), \tag{A.31}$$

which follows from expanding equation A.27.

This type of EM has 2 alternated M steps:

- Log-likelihood maximization of $E_{p(X|O,\Theta_l^{old})}[\ln p(O|X, \Theta_{l1})]$. In this case, the update formulas of the parameters are the same as in equations A.15 A.17, A.20, and A.24. The reason why they are the same is that the only difference between $E_{p(X|O,\Theta_l^{old})}[\ln p(O|X, \Theta_{l1})]$ and equation A.8, which is the one that we maximized before, is the term $\sum_X q(X) \ln p(X)$, which is present in the second but not in the first. However, when deriving these functions with regard to the parameters, this term does not affect, because it does not depend on the parameters Θ_{l1}, and the update equations are the same.

- Minimization of $D_{KL}(p(X|O, \Theta_l^{old})||p(X|\Theta_{l2}))$. This step can be divided in another two:

  - The minimization itself is carried out with respect to $\Pi_2 = \{\mu_{X_2}, \Sigma_{X_2}\}$, because we let the prior distribution $p(X)$ to be a non-standard Gaussian. That is $p(X) \sim \mathcal{N}(\mu_X, \Sigma_X)$. And $p(X|O, \Theta_l^{old})$ was defined in A.10. Then, the KL divergence between two Gaussians is defined as

$$\begin{aligned} & D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) \\ = & \sum_{s=1}^{J_l} \frac{1}{2} \left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) \right. \\ & \left. -D - \ln\left(\frac{\det \Sigma_0}{\det \Sigma_1}\right)\right), \end{aligned} \tag{A.32}$$

    with D the feature dimensionality. Substituiting $\mathcal{N}_0$ by $p(X|O, \Theta_l^{old})$, and $\mathcal{N}_1$ by $p(X|\Pi_2)$, we reach the following expression for the divergence

$$\begin{aligned} & D_{KL}(p(X|O, \Theta_l^{old})||p(X|\Theta_{l2})) \\ = & \sum_{s=1}^{J_l} \frac{1}{2} \left(\mathrm{tr}(\Sigma_X^{-1}L^{-1}) + (\mu_X - E_X[x(s)])^T \Sigma_X X^{-1}(\mu_X - E_X[x(s)]) \right. \\ & \left. -D - \ln\left(\frac{\det L^{-1}}{\det \Sigma_X}\right)\right). \end{aligned} \tag{A.33}$$

Deriving with respect to μ_X and making the derivative equal to zero, we obtain an expression for μ_X

$$\frac{\partial}{\partial \mu_X} D_{KL} = J_l 2\mu_x^T \Sigma_X^{-1} - 2\Sigma_X^{-1} E_X[x(s)] = 0 \tag{A.34}$$

$$\mu_X = \frac{1}{J_l} \sum_{s=1}^{J_l} E_X[x(s)] \tag{A.35}$$

Deriving with respect to Σ_X and making the derivative equal to zero, we obtain an expression for Σ_X

$$\begin{aligned} \frac{\partial}{\partial \Sigma_X} D_{KL} &= -2\Sigma_X^{-1} L^{-1} \Sigma_X^{-1} + I \circ (\Sigma_X^{-1} L^{-1} \Sigma_X^{-1}) \\ &-\Sigma_X^{-1}(\mu_X - E_X[x(s)])(\mu_X - E_X[x(s)])^T \Sigma_X^{-1} + \Sigma_X^{-1} \\ \approx -\Sigma_X^{-1} L^{-1} \Sigma_X^{-1} &- \Sigma_X^{-1}(\mu_X - E_X[x(s)])(\mu_X - E_X[x(s)])^T \Sigma_X^{-1} + \Sigma_X^{-1} = 0 \end{aligned} \tag{A.36}$$

where it is assumed that the off-diagonal terms of $\Sigma_X^{-1} L^{-1}$ are much smaller than the diagonal terms and the approximation

$$I \circ (\Sigma_X^{-1} L^{-1} \Sigma_X^{-1}) \approx \Sigma_X^{-1} L^{-1} \Sigma_X^{-1} \tag{A.37}$$

is applied. Finally

$$\Sigma_X = \sum_{s=1}^{J_l} L^{-1} + (\mu_X - E_X[x(s)])(\mu_X - E_X[x(s)])^T \tag{A.38}$$

– At this point we have to obtain the equivalent model, because in the next iteration we will work again over Θ_{l1}. Therefore, the models with and without standard normal prior must be equivalent, being $p(X_1) \sim \mathcal{N}(x_1; 0, I)$ and $p(X_2) \sim \mathcal{N}(x_2; \mu_{X_2}, \Sigma_{X_2})$, so

$$\begin{aligned} t_{l1} + U_{l1} x_1 = t_{l2} + U_{l2} x_2 = t_{l2} + U_{l2} \phi^{-1}(x_1) \\ = t_{l2} + U_{l2}(P(x_1 + \mu_{X_2})) = (U_{l2}\mu_{X_2} + t_{l2}) + U_{l2} P x_1 \end{aligned} \tag{A.39}$$

were in this case $m_2 = m^{old}$ and $U_2 = U^{old}$ are the hyperparameters from previous iteration, and the update equations are

$$U_l = U_{l1} = U_{l2} P \tag{A.40}$$

$$t_l = t_{l1} = t_{l2} + U_{l2}\mu_{X_2}, \tag{A.41}$$

but in our experiments the means t_l are not updated.

Broadly speaking, MD minimizes the divergence between the posterior distribution of the hidden variables and its prior distribution. In this minimization, we let the prior to be non-standard Gaussian, even when our assumption at the beginning was that it was normal. Once calculated the new prior, we transform the hyperparameters of the model for the equivalent model with standard prior, and alternate between the ML and MD step succesively. In a general case, we could have itegrated from the beginning the mean and covariance of the prior also in the ML step. Each MD assures that the log-likelihood of the current iteration will increase unless we are in a maximum. However, it does not give higher global log-likelihood than EM only with the ML step, but the convergence is faster. In addition, some authors assure that EM without MD is more vulnerable to getting stuck in saddle-points and MD helps to avoid this.

References

[Akbacak2012] Murat Akbacak, D Vergyri, Andreas Stolcke, and Nicolas Scheffer. Effective Arabic dialect classification using diverse phonotactic models. *Interspeech*, pages 2–5, 2012. 50

[Atrey2006] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli. Audio Based Event Detection for Multimedia Surveillance. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, pages V–813–V–816, 2006. 28

[Aucouturier2007] Jean-Julien Aucouturier, Boris Defreville, and François Pachet. The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, 122(2):881–91, August 2007. 19

[Baillie2003] Mark Baillie and Joemon M Jose. Audio-based Event Detection for Sports Video. In *Proc. International Conferences on image and video retrieval*, pages 61–65, 2003. 17

[Ballan2010] Lamberto Ballan, Marco Bertini, Alberto Bimbo, Lorenzo Seidenari, and Giuseppe Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, November 2010. 14

[Barras2003] C. Barras and J.-L. Gauvain. Feature and score normalization for speaker verification of cellular data. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, pages II–49–52, 2003. 30

[Beal2003] MJ Beal, Nebojsa Jojic, and Hagai Attias. A graphical model for audiovisual object tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003. 20

[Bellman1957] Richard Ernest Bellman. Dynamic programming. *Princeton University Press*, 1957. 20

[Bishop2006] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. 2006. 34, 65, 109

[Breiman1984] L. Breiman, J. Friedman, C.J. Stone, and Olshen R.A. *Classification and Regression Trees*. Chapman & Hall/CRC, 1984. 37

[Brummer2009] N. Brummer, Albert Strasheim, Valiantsina Hubeika, P. Mat\vejka, L. Burget, and O. Glembek. Discriminative acoustic language recognition via channel-compensated GMM statistics. In *Proc Interspeech*, pages 2187–2190, 2009. 61

[Brummer2010] N. Brummer. *Measuring, refining and calibrating speaker and language information extracted from speech*. PhD thesis, 2010. 69

[Buntine1992] W Buntine. Learning classification trees. *Statistics and computing*, (January 1991), 1992. 38

[Butko2010a] Taras Butko, C Nadeu Camprubí, and Henrik Schulz. Albayzin-2010 audio segmentation evaluation: evaluation setup and results. In *II Iberian SLTech*, pages 305–308, 2010. 60

[Butko2010b] Taras Butko and C Nadeu Camprubí. Detection of overlapped acoustic events using fusion of audio and video modalities. In *Proc. FALA*, pages 165–168, 2010. 83

[Butko2011] Taras Butko and C. Nadeu. Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. *EURASIP Journal on Audio Speech and Music Processing*, 2011(1):1, 2011. 30, 62

[Butko2011c] Taras Butko and C. Nadeu. *Feature Selection for Multimodal Acoustic Event Detection*. PhD thesis, 2011. 23, 27, 105

[Byun2012] B. Byun, I. Kim, S.M. Siniscalchi, and Lee C.H. Consumer-level multimedia event detection through unsupervised audio signal modeling. In *Interspeech*, 2012. 21

[Campbell2007] W. M. Campbell, F. Richardson, and D. a. Reynolds. Language Recognition with Word Lattices and Support Vector Machines. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 1(2):IV–989–IV–992, 2007. 49

[Casey2002] MA Casey. Sound Classification and Similarity. In *Introduction to MPEG-7: Multimedia Content Description Language, J. Wiley*. 2002. xiii, 26

[Castaldo2007] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair. Compensation of Nuisance Factors for Speaker and Language Recognition. *IEEE Trans Audio Speech Lang Process*, 15(7):1969–1978, September 2007. 67

[Castan2011] Diego Castan, C Vaquero, Alfonso Ortega, David Martínez, and Eduardo Lleida. Hierarchical Audio Segmentation with HMM and Factor Analysis in Broadcast News Domain. In *Proc. Interspeech*, 2011. 25, 43, 67

[Castan2012] Diego Castan, Alfonso Ortega, and Eduardo Lleida. Factor Analysis Segmentation and Classification in Broadcast News Domain. In *Proc. III Iberian SLTech*, 2012. 67

[Castan2013] Diego Castan and Murat Akbacak. Segmental-GMM Approach based on Acoustic Concept Segmentation. In *SLAM Workshop*, 2013. 26, 40, 47, 50

[Castan2013a] Diego Castan, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Broadcast News Segmentation with Factor Analysis System. In *SLAM Workshop*, pages 1–6, 2013. 61, 69, 79

[Castan2013b] Diego Castan and Murat Akbacak. Indexing Multimedia Documents with Acoustic Concept Recognition Lattices. In *Interspeech*, pages 3–7, 2013. 26, 40

[Castan2013c] Diego Castan, Alfonso Ortega, Jesus Villalba, Antonio Miguel, and Eduardo Lleida. Segmentation-by-Classification system based on Factor Analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013. 61, 69

[Chakraborty2013] Rupayan Chakraborty. *Acoustic Event Detection and Localization using Distributed Microphone Arrays*. PhD thesis, 2013. 27, 83

[Chang1996] YL Chang and W Zeng. Integrated image and speech analysis for content-based video indexing. In *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, 1996. 22

[Chaudhuri2012] S. Chaudhuri, R. Singh, and R. Raj. Exploiting Temporal Sequence Structure for Semantic Analysis of Multimedia. In *Interspeech*, 2012. 21

[Chen1997] Scott Shaobing Chen and P S Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA Broadcast News Workshop*, 1998. 24

[Chen2005] Jianfeng Chen, Jianmin Zhang, Alvin Harvey Kam, and Louis Shue. An Automatic Acoustic Bathroom Monitoring System. *2005 IEEE International Symposium on Circuits and Systems*, pages 1750–1753, 2005. 26, 28

[Chu2009] Selina Chu, Shrikanth Narayanan, and CCJ Kuo. Environmental sound recognition with timefrequency audio features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6):1142–1158, 2009. 30

[Clavel2005] C Clavel, T Ehrette, and G Richard. Events detection for an audio-based surveillance system. *IEEE International Conference on Multimedia and Expo*, pages 3–6, 2005. 28

[Dalal] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:886–893. 16

[Dehak2010] N. Dehak, P. Kenny, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, (99):1, 2010. 36

[Dhanalakshmi2011] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam. Classification of audio signals using AANN and GMM. *Applied Soft Computing*, 11(1):716–723, January 2011. 30

[Dhara1999] S Dharanipragada and M Franz. Story segmentation and topic detection in the broadcast news domain. *DARPA Broadcast News Workshop*, pages 1–4, 1999. 30

[Elizalde2013] Benjamin Elizalde, Mirco Ravanelli, and Gerald Friedland. Audio Concept Ranking for Video Event Detection on User-Generated Content. In *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia*, pages 9–14, 2013. 26

[Eronen2006] a.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):321–329, January 2006. 17

[Foote1997] J. Foote. A similarity measure for automatic audio classification. *American Association for Artificial Intelligence: Intelligence Integration and Use of Text, Image, Video, and Audio Corpora*, 1997. 25

[Gallardo-Antolin2010] A Gallardo-Antolín and JM Montero. Histogram equalization-based features for speech, music, and song discrimination. *Signal Processing Letters*, 17(7):659–662, 2010. 30

[Gallardo2010] A. Gallardo and R. San Segundo. UPM-UC3M system for music and speech segmentation. In *II Iberian SLTech*, pages 421–424, 2010. 25, 78, 79

[Glembek2009] O. Glembek, Lukas Burget, N. Dehak, N. Brummer, and P. Kenny. Comparison of scoring methods used in speaker recognition with Joint Factor Analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4057–4060, April 2009. 36, 68

[Gonzalez2010] JA González, AM Peinado, and AM Gómez. A feature compensation approach using VQ-based MMSE estimation for robust speech recognition. *lorien.die.upm.es*, (1):111–114, 2010. 30

[Gorelick2007] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–53, December 2007. 13

[Haitsma2002] Jaap Haitsma and T Kalker. A highly robust audio fingerprinting system. *ISMIR*, 2002. 21

[Hasan2004] MR Hasan, Mustafa Jamil, and MGRMS Rahman. Speaker identification using Mel frequency cepstral coefficients. In *International Conference on Computer and Electrical Engineering*, number December, pages 28–30, 2004. 30

[Hauptmann2003] A Hauptmann, RV Baron, and M Chen. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proc. TRECVID*, 2003. 30

[Huang2006] Rongqing Huang and J Hansen. Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora. *IEEE Trans Audio Speech Lang Process*, 14(3):907–919, 2006. 25, 31

[Hubeika2010] Valiantsina Hubeika and Albert Strasheim. Data selection and calibration issues in automatic language recognition investigation with BUT-AGNITIO NIST LRE 2009 system. In *Odyssey*, number July, pages 215–221, 2010. 72

[Imai1983] S Imai. Cepstral analysis synthesis on the mel frequency scale. In *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, pages 93–96, 1983. 30

[Inoue2011] Nakamasa Inoue, Yusuke Kamishima, and Shunsuke Sato. TokyoTech + Canon at TRECVID 2011. *NIST TRECVid Workshop*, 2011. 20

[Jhuo2013] I-Hong Jhuo, Guangnan Ye, Shenghua Gao, Dong Liu, YG Jiang, D. T. Lee, and Shih-Fu Chang. Discovering joint audiovisual codewords for video event detection. *Machine Vision and Applications*, 25(1):33–47, October 2013. xiii, 13, 17, 20, 32

[Jiang2007] YG Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07*, pages 494–501, 2007. 18

[Jiang2009] Wei Jiang, Courtenay Cotton, Shih-Fu Chang, D Ellis, and Alexander Loui. Short-term audio-visual atoms for generic video concept classification. *Proceedings of the seventeen ACM international conference on Multimedia - MM '09*, page 5, 2009. 20

[Jiang2010] YG Jiang, Xiaohong Zeng, Guangnan Ye, and D Ellis. Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching. *TRECVID*, 2010. 20, 21, 32, 50

[Jiang2010a] YG Jiang, Xiaohong Zeng, Guangnan Ye, and D Ellis. Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching. *NIST TRECVid Workshop*, 2010. 14, 16, 17

[Jiang2011] YG Jiang, G Ye, and SF Chang. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 2011. 13

[Jiang2011a] W Jiang and AC Loui. Audio-visual grouplet: temporal audio-visual interactions for general video concept classification. *Proceedings of the 19th ACM international conference on Multimedia*, 2011. 20

[Jiang2011b] YG Jiang, Guangnan Ye, and SF Chang. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 2011. 20

[Jiang2012] Lu Jiang, Alexander G. Hauptmann, and Guang Xiang. Leveraging high-level and low-level features for multimedia event detection. *Proceedings of the 20th ACM international conference on Multimedia - MM '12*, page 449, 2012. 14, 16

[Jiang2012a] YG Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, November 2012. 12, 14

[Jiang2012b] YG Jiang. SUPER: Towards Real-time Event Recognition in Internet Videos Categories and Subject Descriptors. In *ACM International Conference Multimedia Retrieval*, 2012. 14

[Juang1985] Biing-Hwang Juang and Lawrence R. Rabiner. Mixture autoregressive hidden Markov models for speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6):1404–1413, 1985. 33

[Kenny2005] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 13(3):345–354, May 2005. 61

[Kenny2005a] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Factor analysis simplified. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 637–640, 2005. 68

[Kenny2006] P. Kenny. Joint factor analysis of speaker and session variability: theory and algorithms. *Online: http://www. crim. ca/perso/patrick. kenny*, pages 1–17, 2006. 61

[Kenny2007] P. Kenny, G. Boulianne, Pierre Ouellet, and P. Dumouchel. Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Trans Audio Speech Lang*, 15(4):1435–1447, May 2007. 35, 61, 65, 66, 67, 68

[Kenny2010a] P. Kenny, Douglas Reynolds, and Fabio Castaldo. Diarization of Telephone Conversations Using Factor Analysis. *IEEE Journal of Selected Topics in Signal Processing*, 4(6):1059–1070, December 2010. 65

[Kittler1998] J. Kittler. Combining classifiers: A theoretical framework. *Pattern Anal. Appl.*, 1(1):18–27, March 1998. 69

[Kotti2008] Margarita Kotti, Emmanouil Benetos, and Constantine Kotropoulos. Computationally Efficient and Robust BIC-Based Speaker Segmentation. *IEEE Trans Audio Speech Lang Process*, 16(5):920–933, July 2008. 24

[Laptev2005] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64:107–123, 2005. 16

[Lavner2009] Yizhar Lavner and Dima Ruinskiy. A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation. *EURASIP Journal on Audio Speech and Music Processing*, 2009:1–15, 2009. 17, 24

[Lee2009] Keansub Lee. *Analysis of environmental sounds.* PhD thesis, Columbia University, 2009. 26

[Lee2010] Keansub Lee and D Ellis. Audio-based semantic concept classification for consumer video. *IEEE Transactions Audio, Speech, and Language Processing*, 18(6):1406–1416, 2010. 19

[Lew2006] MS Lew, N Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006. 2

[Li2001] Dongge Li, IK Sethi, Nevenka Dimitrova, and T McGee. Classification of general audio data for content-based retrieval. *Pattern recognition letters*, 22, 2001. 30

[Li2012] Haizhou Li, Bin Ma, and KA Lee. Spoken Language Recognition: from Fundamentals to Practice. *Proceedings of IEEE*, 2013. 21, 67

[Liporace1982] Louis a Liporace. Maximum Likelihood Estimation for Multivariate Observations of Markov Sources. *IEEE Trans Information Theory*, I, 1982. 33

[Liu1998] Z Liu, Y Wang, and Tsuhan Chen. Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI signal processing systems*, 79:61–79, 1998. 17, 30

[Liu2010] Yang Liu, Wan-Lei Zhao, Chong-Wah Ngo, Chang-Sheng Xu, and Han-Qing Lu. Coherent bag-of audio words model for efficient large-scale video copy detection. In *Proceedings of the ACM International Conference on Image and Video Retrieval - CIVR '10*, page 89, New York, New York, USA, 2010. ACM Press. 18

[Lowe2004] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004. 16

[Lu2002] Lie Lu, HJ Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516, 2002. 25, 30

[Lu2003] Lie Lu, Hong-Jiang Zhang, and Stan Z. Li. Content-based audio classification and segmentation by using support vector machines. *Multimedia Syst.*, 8(6):482–492, April 2003. 25

[Lu2008] Lie Lu and Alan Hanjalic. Audio Keywords Discovery for Text-Like Audio Content Analysis and Retrieval. *IEEE Transactions on Multimedia*, 10(1):74–85, January 2008. 19

[Luengo2010] Iker Luengo, Eva Navas, and I Hernáez. Feature analysis and evaluation for automatic emotion identification in speech. *Multimedia, IEEE Transactions . . .*, 12(6):490–501, 2010. 30

[Lukowicz2003] Paul Lukowicz, Niroshan Perera, and T Starner. Soundbutton: Design of a low power wearable audio classification system. In *IEEE Int. Symposium on Wearable Computers*, number October, pages 12–17, 2003. 26

[Marin-Jimenez2013] M. J. Marín-Jiménez, R. Muñoz Salinas, E. Yeguas-Bolivar, and N. Pérez de la Blanca. Human interaction categorization by using audio-visual cues. *Machine Vision and Applications*, 25(1):71–84, June 2013. 16

[Markaki2011] Maria Markaki and Yannis Stylianou. Discrimination of speech from nonspeeech in broadcast news based on modulation frequency features. *Speech Communication*, 53(5):726–735, May 2011. 30

[Marszalek2009] M Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Computer Vision and Pattern Recognition*, number i, 2009. 13

[Martin1997] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. The DET Curve in Assessment of Detection Task Performance. *Proc. Eurospeech '97*, pages 1895–1898, 1997. 48

[Martinez2011] David Martínez, Antonio Miguel, Alfonso Ortega, and Eduardo Lleida. I3A Language Recognition System for Albayzin 2010 LRE. In *Proc. Interspeech*, Florence, Italy, 2011. 72

[Mckinney2003] MF McKinney and Jeroen Breebaart. Features for audio and music classification. *ISMIR*, 4, 2003. 17, 30

[Mertens2011] Robert Mertens, Howard Lei, Luke Gottlieb, Gerald Friedland, and Ajay Divakaran. Acoustic super models for large scale video event detection. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events - J-MRE '11*, page 19, New York, New York, USA, 2011. ACM Press. 18, 21

[Mierswa2005] Ingo Mierswa and Katharina Morik. Automatic Feature Extraction for Classifying Audio Data. *Machine Learning*, 58(2-3):127–149, February 2005. 30

[Mikolajczyk2005] K. Mikolajczyk, T. Tuytelaars, C. Schmid, a. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1-2):43–72, October 2005. 16

[Misra2012] A Misra. Speech/Nonspeech Segmentation in Web Videos. In *Proc. Interspeech*, 2012. 25

[Mostefa2008a] Djamel Mostefa, Nicolas Moreau, Khalid Choukri, Gerasimos Potamianos, Stephen M. Chu, Ambrish Tyagi, Josep R. Casas, Jordi Turmo, Luca Cristoforetti, Francesco Tobia, Aristodemos Pnevmatikakis, Vassilis Mylonakis, Fotios Talantzis, Susanne Burger, Rainer Stiefelhagen, Keni Bernardin, and Cedrick Rochet. The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3-4):389–407, January 2008. xv, 27, 84

[Mporas2007] Iosif Mporas and Todor Ganchev. Comparison of Speech Features on the Speech Recognition Task. *Journal of Computer Science*, 3(8):608–616, 2007. 17, 30

[Myers2013] Gregory K. Myers, Ramesh Nallapati, Julien Hout, Stephanie Pancoast, Ramakant Nevatia, Chen Sun, Amirhossein Habibian, Dennis C. Koelma, Koen E. a. Sande, Arnold W. M. Smeulders, and Cees G. M. Snoek. Evaluating multimedia features and fusion for example-based event detection. *Machine Vision and Applications*, 25(1):17–32, July 2013. 17, 19, 20, 32

[NIST] NIST. NIST DETware V.2. 48

[NIST2009] NIST. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, 2009. 63, 79

[NIST2011] NIST. Evaluation, TRECVID multimedia event detection 2011, 2011. xiii, 12, 13, 41

[Nadeu2001] C. Nadeu, Dušan Macho, and Javier Hernando. Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Communication*, 34:93–114, 2001. 29

[Natarajan2011] P. Natarajan and Vasant Manohar. BBN VISER TRECVID 2011 multimedia event detection system. *NIST TRECVid Workshop*, 2011. 20, 22, 32

[Natarajan2012] P. Natarajan, S. Vitaladevuni, S. Tsakalidis, and R. Prasad. Multimodal feature fusion for robust event detection in web videos. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1305, June 2012. 14, 16, 18

[Nguyen2011] N Nguyen, M Haque, Cheol-hong Kim, and JM Kim. Audio segmentation and classification using a temporally weighted fuzzy C-means algorithm. *Advances in Neural Networks*, pages 447–456, 2011. 24

[Nwe2005] T.L. Nwe and H. Li. Broadcast news segmentation by audio type analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages ii–1065, 2005. 30

[Oh2013] Sangmin Oh, Scott McCloskey, Ilseo Kim, Arash Vahdat, Kevin J. Cannons, Hossein Hajimirsadeghi, Greg Mori, a. G. Amitha Perera, Megha Pandey, and Jason J. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Machine Vision and Applications*, 25(1):49–69, July 2013. 17, 22

[Ojala2002] Timo Ojala. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–35, 2002. 16

[Oliva2001] Aude Oliva and A Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 15

[Pancoast2012] Stephanie Pancoast and Murat Akbacak. Bag-of-Audio-Words Approach for Multimedia Event Classification. *Interspeech*, pages 1–4, 2012. 18, 21, 41, 53, 54, 103

[Pancoast2012a] Stephanie Pancoast, Murat Akbacak, and Michelle Hewlett Sanchez. Supervised Acoustic Concept Extraction for Multimedia Event Detection. In *ACM Multimedia Workshop*, 2012. xiv, xvii, 21, 26, 40, 41, 42, 46, 47, 50, 53, 103

[Pancoast2013] Stephanie Pancoast and Murat Akbacak. N-gram extension for bag-of-audio-words. In *ICASSP*, pages 778–782, 2013. 21

[Papadopoulos2012] Symeon Papadopoulos, Emmanouil Schinas, Vasileios Mezaris, Raphael Troncy, and Ioannis Kompatsiaris. Social Event Detection at MediaEval 2012: Challenges, datasets, and evaluation, 2012. 13

[Peeters2004] G. Peeters. {A large set of audio features for sound description (similarity and classification) in the CUIDADO project}. *CUIDADO IST Project Report*, pages 1–25, 2004. 17

[Philbin2008] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. 18

[Quinlan1986] JR Quinlan. Induction of decision trees. *Machine learning*, pages 81–106, 1986. 37

[Reed2009] J Reed and CH Lee. On the importance of modeling temporal information in music tag annotation. In *ICASSP*, pages 1873–1876, 2009. 21

[Reuter2013] Timo Reuter, Symeon Papadopoulos, Vasileios Mezaris, Cimiano Philipp, Christopher De Vries, and Shlomo Geva. Social Event Detection at MediaEval 2013 : Challenges, Dataset, and Evaluation, 2013. 13

[Reynolds2000] Douglas Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Process.*, 10(1-3):19–41, January 2000. 35, 64, 66

[Reynolds2005] D.a. Reynolds and P. Torres-Carrasquillo. Approaches and Applications of Audio Diarization. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 5:953–956, 2005. 5, 23

[Richardson2008] F. S. Richardson and W. M. Campbell. Language recognition with discriminative keyword selection. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4145–4148, March 2008. 49

[Rodriguez2008a] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. Ieee, June 2008. 13

[Sarikaya2000] R. Sarikaya and J Hansen. High resolution speech feature parametrization for monophone-based stressed speech recognition. *IEEE Signal Processing Letters*, 7(7):182–185, July 2000. 30

[Saunders1996] J Saunders. Real-time discrimination of broadcast speech/music. In *Acoustics, Speech, and Signal Processing*, pages 993–996, 1996. 30

[Scholkopf2002] B. Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002. 31

[Schuldt2004] C Schuldt, I Laptev, and B Caputo. Recognizing human actions: a local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 3–7, 2004. 13

[Scovanner2007] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. *Proceedings of the 15th international conference on Multimedia - MULTIMEDIA '07*, (c):357, 2007. 16

[Siegler1997] MA Siegler, Uday Jain, Bhiksha Raj, and RM Stern. Automatic segmentation, classification and clustering of broadcast news audio. *Proc. DARPA Broadcast News Workshop*, pages 4–6, 1997. 24

[Sivic2003] Josef Sivic and Andrew Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, number Iccv, pages 1470–1477 vol.2. Ieee, 2003. 18

[Smaeton2006] Alan F Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVid. In *MIR'06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press. 13

[Smeulders2000] AWM Smeulders and Marcel Worring. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000. 3

[Smith2003] JR Smith, Milind Naphade, and A Natsev. Multimedia semantic indexing using model vectors. *Multimedia and Expo, 2003.*, pages 0–3, 2003. 14

[Snoek2007] Cees G. M. Snoek and Marcel Worring. Concept-Based Video Retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2007. 16

[Stein2008] Barry E Stein and Terrence R Stanford. Multisensory integration: current issues from the perspective of the single neuron. *Nature reviews. Neuroscience*, 9(4):255–66, April 2008. 3

[Stiefelhagen2007] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, R Travis Rose, Martial Michel, and John Garofolo. The CLEAR 2007 Evaluation. *Evaluation*, 2007. 27

[Tamrakar2012] Amir Tamrakar, Saad Ali, Qian Yu, and Jingen Liu. Evaluation of low-level features and their combinations for complex event detection in open source videos. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3681–3688, 2012. 14

[Temko2005] A. Temko, D. Macho, C. Nadeu, and C. Segura. UPC-TALP Database of Isolated Acoustic Events. In *Internal UPC report*, 2005. 85

[Temko2006] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. In *Jornadas de tecnologia del habla*, volume 65, pages 5–11. Citeseer, 2006. 26

[Temko2006b] A. Temko, Robert Malkin, Christian Zieger, and Dusan Macho. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. In *IV Jornadas en Tecnología del Habla*, pages 1–6, 2006. 82

[Temko2009] A. Temko and C. Nadeu. Acoustic event detection in meeting-room environments. *Pattern Recognition Letters*, 30(14):1281–1288, 2009. 27

[Temko2009a] A. Temko. *Acoustic event detection and classification*. PhD thesis, 2009. xiii, 23, 26, 27, 82

[Tsao2010] Yu Tsao, Hanwu Sun, Haizhou Li, and Chin-Hui Lee. An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4422–4425, 2010. 21

[Vacher2003] Michel Vacher, Dan Istrate, and Laurent Besacier. Smart audio sensor for telemedicine. In *Proc. Smart Object Conference*, pages 3–6, 2003. 26

[VanGemert2010] Jan C van Gemert, Cor J Veenman, Arnold W M Smeulders, and Jan-Mark Geusebroek. Visual word ambiguity. *IEEE transactions on pattern analysis and machine intelligence*, 32(7):1271–83, July 2010. 18

[VanHout2013] Julien van Hout, Murat Akbacak, Diego Castan, Eric Yeh, and Michelle Sanchez. Extracting Spoken and Acoustic Concepts For Multimedia Event Detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2–6, 2013. 19, 22, 55

[Vaquero-AvilesCasco2011] Carlos Vaquero-Avilés Casco. *Robust Diarization for Speaker Characterization*. PhD thesis, 2011. 23

[Vaquero2010a] C Vaquero, Alfonso Ortega, Jesús Villalba, Antonio Miguel, and Eduardo Lleida. Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification. In *Proc Interspeech 2010*, volume 2010, pages 2310–2313, 2010. 61

[Vaquero2011] C Vaquero, Alfonso Ortega, and Eduardo Lleida. Intra-session variability compensation and a hypothesis generation and selection strategy for speaker segmentation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3–6, 2011. 61, 64

[Vaquero2013] Carlos Vaquero, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Quality Assessment of Speaker Diarization for Speaker Characterization. *IEEE Trans. on Acoustics, Speech and Language Processing*, 2013. 61

[Vergin1996] R. Vergin, D. O'Shaughnessy, and V. Gupta. Compensated mel frequency cepstrum coefficients. In *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 323–326, 1996. 30

[Vergin1999] R. Vergin. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on Speech and Audio Processing*, 7(5):525–532, 1999. 30

[Vogt2008] Robbie Vogt and Sridha Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, 22(1):17–38, January 2008. 67

[Wang2008] Feng Wang, Yu-gang Jiang, and Chong-wah Ngo. Motion Relativity and Visual Relatedness. In *Proc. of ACM international conference on multimedia*, 2008. 16

[Willsky1976] AS Willsky and HL Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *Automatic Control, IEEE Transactions on*, (February):108–112, 1976. 24

[Wolpert1997] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997. 19

[Wong2001] E. Wong and S. Sridharan. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In *International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 95–98, 2001. 30

[Wu2006] Chung-hsien Wu and Chia-hsin Hsieh. Multiple change-point audio segmentation and classification using an MDL-based Gaussian model. *IEEE Trans Audio Speech Lang Process*, 14(2):647–657, March 2006. 25

[Wu2006a] Chung-hsien Wu and YH Chiu. Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs. *IEEE Trans Audio Speech Lang Process*, 14(1):266–276, 2006. 24

[Xie2010] Lei Xie, Zhong-Hua Fu, Wei Feng, and Yong Luo. Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news. *Multimedia Systems*, 17(2):101–112, September 2011. 30

[Yapanel2008] U Yapanel and J Hansen. A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition. *Speech Communication*, 50(2):142–152, February 2008. 30

[Yuan2009] Junsong Yuan. Discriminative subvolume search for efficient action detection. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2442–2449, June 2009. 13

[Zheng2001] Fang Zheng, Guoliang Zhang, and Zhanjiang Song. Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, 2001. 17

[Zhou2008] X Zhou, Xiaodan Zhuang, and Ming Liu. HMM-based acoustic event detection with AdaBoost feature selection. In *CLEAR 2007*, 2008. 27, 82, 96

[Zhuang2012] X. Zhuang, S. Tsakalidis, S. Wu, P. Natarajan, and R. Prasad. Compact Audio Representation for Event Detection in Consumer Media. In *Interspeech*, 2012. 21

[Zieger2008a] Christian Zieger and Maurizio Omologo. Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm. *INTERSPEECH*, 2008. 27