UNIVERSITAT POLITÈCNICA
DE CATALUNYA

PhD Thesis

Feature Selection for Multimodal Acoustic Event Detection

Author: Taras Butko
Advisor: Dr. Climent Nadeu

Speech Processing Group
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
Barcelona, June 2011

ACTA DE QUALIFICACIÓ DE LA TESI DOCTORAL

Reunit el tribunal integrat pels sota signants per jutjar la tesi doctoral:

Títol de la tesi: **FEATURE SELECTION FOR MULTIMODAL ACOUSTIC EVENT DETECTION**
Autor de la tesi: **TARAS BUTKO**

Acorda atorgar la qualificació de:

□ No apte
□ Aprovat
□ Notable
□ Excel·lent
□ Excel·lent Cum Laude

Barcelona, …………… de/d’…................…………….. de ....….

El President

....
(nom i cognoms)

El Secretari

....
(nom i cognoms)

El vocal

....
(nom i cognoms)

El vocal

....
(nom i cognoms)

El vocal

....
(nom i cognoms)

Abstract

The detection of the Acoustic Events (AEs) naturally produced in a meeting room may help to describe the human and social activity. The automatic description of interactions between humans and environment can be useful for providing: implicit assistance to the people inside the room, context-aware and content-aware information requiring a minimum of human attention or interruptions, support for high-level analysis of the underlying acoustic scene, etc. Activity detection and description is a key functionality of perceptually aware interfaces working in collaborative human communication environments like meeting-rooms. On the other hand, the recent fast growth of available audio or audiovisual content strongly demands tools for analyzing, indexing, searching and retrieving the available documents. Given an audio document, the first processing step usually is audio segmentation (AS), i.e. the partitioning of the input audio stream into acoustically homogeneous regions which are labelled according to a predefined broad set of classes like speech, music, noise, etc. In fact, AS can be seen as a particular case of acoustic event detection, as it will be done herewith.

Acoustic event detection (AED) is the objective of this thesis work. A variety of features coming not only from audio but also from the video modality is proposed to deal with that detection problem in meeting-room and broadcast news domains. Two basic detection approaches are investigated in this work: a joint segmentation and classification using Hidden Markov Models (HMMs) with Gaussian Mixture Densities (GMMs), and a detection-by-classification approach using discriminative Support Vector Machines (SVMs). For the first case, a fast one-pass-training feature selection algorithm is developed in this thesis to select, for each AE class, the subset of multimodal features that shows the best detection rate.

AED in meeting-room environments aims at processing the signals collected by distant microphones and video cameras in order to obtain the temporal sequence of (possibly overlapped) AEs that have been produced in the room. When applied to interactive seminars with a certain degree of spontaneity, the detection of acoustic events from only the audio modality alone shows a large amount of errors, which is mostly due to the temporal overlaps of sounds. This thesis includes several novelties regarding the task of multimodal AED. Firstly, the use of video features. Since in the video modality the acoustic sources do not overlap (except for occlusions), the proposed features improve AED in such rather spontaneous scenario recordings. Secondly, the inclusion of acoustic localization features, which, in combination with the usual spectro-temporal audio features, yield a further improvement in recognition rate. Thirdly, the comparison of feature-level and decision-level fusion strategies for the combination of audio and video modalities. In the later case, the system output scores are combined using two statistical approaches: weighted arithmetical mean and fuzzy integral. On the other hand, due to the scarcity of annotated multimodal data, and, in particular, of data with temporal sound overlaps, a new multimodal database with a rich variety of meeting-room AEs has been recorded and manually annotated, and it has been made publicly available for research purposes.

For audio segmentation in the broadcast news domain, a hierarchical system architecture is proposed, which properly groups a set of detectors, each one corresponding to one of the acoustic classes of interest. Two different AS systems have been developed for two broadcast news databases: the first one corresponds to audio recordings from the TV debate program Àgora from the Catalan TV3 channel, and the second one includes diverse audio segments from the 3/24 Catalan broadcast news TV channel. The output of the first AS system was used as the first stage in the automatic translation and subtitling applications developed for the Tecnoparla project, a project that developed several speech technologies to extract all possible information from audio. The second AS system, which is a hierarchical HMM-GMM-based detection system with feature selection, obtained competitive results in the *Albayzín-2010* audio segmentation evaluation.

Finally, it is worth mentioning a few side outcomes from this thesis work. The author has been responsible for the organization of the above mentioned *Albayzín-2010* evaluation of audio segmentation systems, taking a leading role in the specification of acoustic classes, databases, metric and evaluation protocol, and also carrying out a posterior analysis of the systems and results submitted by eight research groups from Spanish and Portuguese universities. Moreover, a 2-source HMM-GMM-based acoustic event detection system has been implemented in the UPC's smart-room, which works in real time, for purposes of testing and demonstration.

Resum

La detecció d'esdeveniments acústics (*Acoustic Events* (AEs)) que es produeixen naturalment en una sala de reunions pot ajudar a descriure l'activitat humana i social. La descripció automàtica de les interaccions entre els éssers humans i l'entorn pot ser útil per a proporcionar: ajuda implícita a la gent dins de la sala, informació sensible al context i al contingut sense requerir gaire atenció humana ni interrupcions, suport per a l'anàlisi d'alt nivell de l'escena acústica, etc. La detecció i la descripció d'activitat és una funcionalitat clau de les interfícies perceptives que treballen en entorns de comunicació humana com sales de reunions. D'altra banda, el recent creixement ràpid del contingut audiovisual disponible requereix l'existència d'eines per a l'anàlisi, indexació, cerca i recuperació dels documents existents. Donat un document d'àudio, el primer pas de processament acostuma a ser la seva segmentació (*Audio Segmentation* (AS)), és a dir, la partició de la seqüència d'entrada d'àudio en regions acústiques homogènies que s'etiqueten d'acord amb un conjunt predefinit de classes com parla, música, soroll, etc. De fet, l'AS pot ser vist com un cas particular de la detecció d'esdeveniments acústics, i així es fa en aquesta tesi.

La detecció d'esdeveniments acústics (*Acoustic Event Detection* (AED)) és un dels objectius d'aquesta tesi. Es proposa tot una varietat de característiques que provenen no només de l'àudio, sinó també de la modalitat de vídeo, per fer front al problema de la detecció en dominis de sala de reunions i de difusió de notícies. En aquest treball s'investiguen dos enfocaments bàsics de detecció: 1) la realització conjunta de segmentació i classificació utilitzant models de Markov ocults (*Hidden Markov Models* (HMMs)) amb models de barreges de gaussianes (*Gaussian Mixture Models* (GMMs)), i 2) la detecció per classificació utilitzant màquines de vectors suport (*Support Vector Machines* (SVM)) discriminatives. Per al primer cas, en aquesta tesi es desenvolupa un algorisme de selecció de característiques ràpid d'un sol pas per tal de seleccionar, per a cada AE, el subconjunt de característiques multimodals que aconsegueix la millor taxa de detecció.

L'AED en entorns de sales de reunió té com a objectiu processar els senyals recollits per micròfons distants i càmeres de vídeo per tal d'obtenir la seqüència temporal dels (possiblement superposats) esdeveniments acústics que s'han produït a la sala. Quan s'aplica als seminaris interactius amb un cert grau d'espontaneïtat, la detecció d'esdeveniments acústics a partir de només la modalitat d'àudio mostra una gran quantitat d'errors, que és sobretot a causa de la superposició temporal dels sons. Aquesta tesi inclou diverses contribucions pel que fa a la tasca d'AED multimodal. En primer lloc, l'ús de característiques de vídeo. Ja que en la modalitat de vídeo les fonts acústiques no se superposen (exceptuant les oclusions), les característiques proposades milloren la detecció en els enregistraments en escenaris de caire espontani. En segon lloc, la inclusió de característiques de localització acústica, que, en combinació amb les característiques habituals d'àudio espectrotemporals, signifiquen nova millora en la taxa de reconeixement. En tercer lloc, la comparació d'estratègies de fusió a nivell de característiques i a nivell de decisions, per a la utilització combinada de les modalitats d'àudio i vídeo. En el darrer cas, les puntuacions de sortida del sistema es combinen fent ús de dos mètodes estadístics: la mitjana aritmètica ponderada i la integral difusa. D'altra banda, a causa de l'escassetat de dades multimodals anotades, i, en particular, de dades amb superposició temporal de sons, s'ha grabat i anotat manualment una nova base de dades multimodal amb una rica varietat d'AEs de sala de reunions, i s'ha posat a disposició pública per a finalitats d'investigació.

Per a la segmentació d'àudio en el domini de difusió de notícies, es proposa una arquitectura jeràrquica de sistema, que agrupa apropiadament un conjunt de detectors, cada un dels quals correspon a una de les classes acústiques d'interès. S'han desenvolupat dos sistemes diferents de SA per a dues bases de dades de difusió de notícies: la primera correspon a gravacions d'àudio del programa de debat *Àgora* del canal de televisió català TV3, i el segon inclou diversos segments d'àudio del canal de televisió català 3/24 de difusió de notícies. La sortida del primer sistema es va utilitzar com a primera etapa dels sistemes de traducció automàtica i de subtitulat del projecte *Tecnoparla*, un projecte finançat pel govern de la Generalitat en el que es desenvoluparen diverses tecnologies de la parla per extreure tota la informació possible del senyal d'àudio. El segon sistema d'AS, que és un sistema de detecció jeràrquica basat en HMM-GMM amb selecció de característiques, ha obtingut resultats competitius en l'avaluació de segmentació d'àudio *Albayzín-2010*.

Per acabar, val la pena esmentar alguns resultats col·laterals d’aquesta tesi. L’autor ha sigut responsable de l'organització de l'avaluació de sistemes de segmentació d'àudio dins de la campanya *Albayzín-2010* abans esmentada. S'han especificat les classes d’esdeveniments, les bases de dades, la mètrica i els protocols d'avaluació utilitzats, i s'ha realitzat una anàlisi posterior dels sistemes i els resultats presentats perls vuit grups de recerca participants, provinents d'universitats espanyoles i portugueses. A més a més, s'ha implementat en la sala multimodal de la UPC un sistema de detecció d'esdeveniments acústics per a dues fonts simultànies, basat en HMM-GMM, i funcionant en temps real, per finalitats de test i demostració.

Thanks to…

First of all, I would like to thank my advisor, Climent Nadeu. During the whole period of my staying in Barcelona he definitely created a pleasant working atmosphere in the UPC's research group. His encouraging support helped me a lot to achieve new results. I am also grateful to him a lot for making a friendly environment out of working time.

The role of Andrey Temko, the friend from my native land, can not be overestimated too. Basically my phd studies at UPC became possible thanks to him. His help was extremely important at the beginning of my study. He still remains a good friend of mine and we continue to communicate even at a distance.

I would like to mention several persons who made an important contribution to my thesis. First, I would thank Cristian Canton. A major part of this thesis was written due to our close cooperation in video technologies, multimodal database recordings etc. Second, I would thank Carlos Segura, whose knowledge in acoustic localization technologies and smart-room demo applications made a strong contribution in my formation. Jordi Luque was a person that I could always ask for advice and he was always willing to provide me support in very kind manner. I would like to thank Javier Hernando for organizing the CHIL meetings and interesting discussions. I would also thank Didac Pérez and Diego Lendoiro. These two smart system administrators and also very easy going guys did everything to make servers and other hardware working all the time.

I would also mention several students with whom I was happy to cooperate in several projects: Mateu Aguilo, Eros Blanco, Fran Gonzalez and Pedro Vizarreta. I was happy to work in the same group, to attend conferences and simply to spend time with Martin Wolf, Martin Zelenak, Jaime Gallego and Henrik Schulz.

I would like to thank the research groups who was willing to participate in Albayzín-2010 audio segmentation evaluation: *ATVS* (Universidad Autónoma de Madrid), *CEPHIS* (Universitat Autònoma de Barcelona), *GSI* (Instituto de Telecomunicações, Universidade de Coimbra, Portugal), *GTC-VIVOLAB* (Universidad de Zaragoza), *GTH* (Universidad Politécnica de Madrid / Universidad Carlos III de Madrid), *GTM* (Universidade de Vigo), *GTTS* (Universidad del País Basco). Their ideas helped me a lot to broaden my knowledge in the topic of audio segmentation.

I am also grateful for funding support from Catalan autonomous government, Tecnoparla Catalan project, SAPIRE Spanish project (TEC2007-65470), SARAI Spanish project (TEC2010-21040-C02-01), European Metanet4u project, and the continuous help from GPS group.

And of course, I would like to thank my wife Anna whose continuous support and romantic environment helped me to finish this thesis. Thanks to my son Yegor for the bright moments in my life and for letting me work during the most crucial time in my research. I appreciate the support from my mother Elena, father Vladimir, brother Platon and mother-in-law Marina who believed in me every moment...

Contents

List of Acronyms

AC(s)	Acoustic Class(es)
AE(s)	Acoustic Event(s)
AED	Acoustic Event Detection
ANN(s)	Artificial Neural Network(s)
AS	Audio Segmentation
ASL	Acoustic Source Localization
ASR	Automatic Speech Recognition
CASA	Computational Auditory Scene Analysis
BIC	Bayesian Information Criteria
BN	Broadcast News
CHIL	Computer in the Human Interaction Loop
CLEAR	Classification of Event, Activities, and Relationships
DCT	Discrete Cosine Transform
DP	Decision Profile
EM	Expectation-Maximization
FBE	Filter Bank Energies
FF (LFBE)	Frequency Filtered (Log Filter Bank Energies)
FFT	Fast Fourier Transform
FI	Fuzzy Integral
FM(s)	Fuzzy Measure(s)
GMM(s)	Gaussian Mixture Model(s)
GUI	Graphic User Interface
HMM(s)	Hidden Markov Model(s)
ICA	Independent Component Analysis
LDA	Linear Discriminant Analysis
LPC	Linear Prediction Coefficients
MFCC	Mel-Frequency Cepstral Coefficients
MP	Matching Pursuit
NIST	National Institute of Standards and Technology
PCA	Principal Component Analysis
PSVM	Proximal Support Vector Machines
RBF	Radial Basis Function
PF	Particle Filter
SNR	Signal-to-Noise Ratio
SVM(s)	Support Vector Machine(s)
WAM	Weighted Arithmetical Mean

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Thesis Overview and Motivation

Acoustic event detection (AED) aims at determining the identity of sounds and their temporal position in audio signals. The detection of the Acoustic Events (AEs) naturally produced in a meeting room may help to describe the human and social activity that takes place in it. The automatic description of interactions between humans and environment can be useful for providing: implicit assistance to the people inside the room, context-aware and content-aware information requiring a minimum of human attention or interruptions, support for high-level analysis of the underlying acoustic scene, etc. In fact, human activity is reflected in a rich variety of AEs, either produced by the human body or by objects handled by humans. Although speech is usually the most informative AE, other kind of sounds may carry useful cues for scene understanding. For instance, in a meeting/lecture context, we may associate a chair moving or door noise to its start or end, cup clinking to a coffee break, or footsteps to somebody entering or leaving. Furthermore, some of these AEs are tightly coupled with human behaviors or psychological states: paper wrapping may denote tension; laughing, cheerfulness; yawning in the middle of a lecture, boredom; keyboard typing, distraction from the main activity in a meeting; and clapping during a speech, approval. Acoustic Event Detection (AED) is also useful in the broadcast news domain audio segmentation. The recent fast growth of multimedia content available on Internet and other sources strongly demands tools for analyzing, indexing, and searching the available documents in order to offer to users the possibility of selecting the desired content, and to companies the capability of tracking, from contents generated by the users themselves, the people preferences, opinions, etc. Given an audio document, the first processing step usually is audio segmentation (AS), i.e. the partitioning of the input audio stream into acoustically homogeneous regions which are labelled according to a predefined broad set of classes like speech, music, noise, etc. Moreover, AS can contribute to improve the performance and robustness of speech technologies such as speech and speaker recognition, and speech enhancement.

AED is usually addressed from an audio perspective and many reported works are intended for indexing and retrieval of multimedia documents [LZJ02] or to improve robustness of speech recognition [NNM03]. Within the context of ambient intelligence, AED applied to give a contextual description of a meeting scenario was pioneered by [Tem07]. Moreover, AED has been adopted as a relevant technology in several international projects, like CHIL [WS09], and evaluation campaigns [CLE06] [CLE07]. The CLEAR'07 international evaluations in seminar conditions have shown that AED is a challenging problem.

Feature selection plays a central role in the tasks of classification and data mining, since redundant and irrelevant features often degrade the performance of classification algorithms [KJ97]. In the meeting-room AED and the broadcast news AS we face the problem of the large number and variety of features proposed in the literature. In fact, there are many features which exploit acoustic content such as sub-band energies computed in short-time windows, time evolution parameters, modulation spectrum, level of harmonicity, etc. Very often authors do not present strong or clear arguments in favour of a particular feature set they propose, and the final decision about feature subset selection is mainly based on their prior knowledge. The feature selection problem is even more acute when other types of data, like video data, are used besides audio. By using a fast one-pass-training feature selection approach we have selected the subset of multimodal features that shows the best detection rate for each class of AEs, observing an improvement in average accuracy with respect to using the whole set of features.

In this thesis we address the problem of the AED in real-world meeting-room environment from multimodal perspective. To deal with the difficult problem of signal overlaps, we use features coming from the video modality as well as acoustic localization features, which are new for the meeting-room AED task. Since in the video modality the acoustic sources do not overlap (except for occlusions), the proposed features improve AED in spontaneous scenario recordings. Although the considered meeting-room events are not acoustic but audio-visual, in the thesis we refer to acoustic events, because the audio characterization of events provides the main description for them. In other words, the event is considered when it has a specific audio counterpart (sound activity), and video information is only an additional source of information which is used to enhance the audio mono-modal detection.

For broadcast news audio segmentation the hierarchical AS systems with feature selection has been developed. The hierarchical system architecture is a group of detectors (called modules), where each module is responsible for detection of one acoustic class of interest. As input it uses the output of the preceding module and has 2 outputs: the first corresponds to audio segments detected as the corresponding class of interest, and the other is the rest of the input stream. One of the most important decisions when using this kind of architecture is to put the modules in the best order in terms of information flow, since some modules may benefit greatly from the previous detection of certain classes. For instance, previous detection of the classes that show high confusion with subsequent classes potentially can improve the overall performance. In this type of architecture, it is not necessary to have the same classifier, feature set and/or topology for different detectors. Two different AS systems has been developed using two broadcast news databases: the first one includes audio recordings from the TV debate program Àgora from the Catalan TV3 channel, and the second one includes diverse audio segments from the 3/24 Catalan broadcast news TV channel. The output of the first AS system was used as the first stage in the automatic translation and subtitling application developed for the Tecnoparla project, a project that included several technologies to extract all possible information from audio: speaker diarization, language recognition, speech recognition, speech translation and text-to-speech syntesis. The second HMM-based AS system with feature selection got competitive results in the Albayzín-2010 audio segmentation evaluation among 8 participants from Spanish and Portuguese universities.

In the presented thesis two-source AED and acoustic source localization (ASL) systems running in real-time in the UPC's smart-room have been developed. The detection of AEs is performed using HMM-GMM approach, which allows analyzing the input waveform on a frame-by-frame basis with low latency. The AED and ASL systems are visually monitored by a GUI application which shows the output of AED and ASL technologies in real-time. Using this application, a video recording has been captured that contains the output of the GUI during a session lasting about 2 min, where three people in the room speak, interact with each other or produce one of the 12 isolated as well as overlapped with speech meeting-room AEs.

1.2 Thesis Objectives

The goal of this thesis work is designing efficient algorithms for acoustic event detection in meeting-room environments and audio segmentation in the broadcast news domain. Determining the occurrence of events is fundamental to developing systems that can observe and react to them. The thesis work contemplates several main research directions: multimodal feature extraction and feature selection for AED, research on different detection/segmentation approaches, fusion of both audio and video modalities to detect AEs.

For meeting-room environments, the task of AED is relatively new; however, it has already been evaluated in the framework of two international evaluation campaigns: in CLEAR (Classification of Events, Activities, and Relationships evaluation campaigns) 2006 [CLE06], and in CLEAR 2007 [CLE07]. In the last evaluations, 5 out of 6 submitted systems showed accuracies below 25%, and the best system got 33.6% accuracy. It has been found that the overlapping segments account for more than 70% of errors produced by every submitted system.

Dealing with the overlap problem is the central research objective in our work. Since the human-produced AEs have a visual correlate, video modality can be exploited to enhance the detection rate of certain AEs. A number of features can be extracted from video recordings by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of AEs. From the audio perspective, the video modality has an attractive property: the disturbing acoustic noise usually does not have a correlate in the video signal.

Enhancing of the recognition results using information from multiple microphones is the other research objective in our work. Since the characteristics of the meeting-room are known beforehand, the position (*x, y, z*) of the acoustic source may carry useful information. Indeed, some acoustic events can only occur at particular locations, like door slam and door knock can only appear near the door, or footsteps and chair moving events take place near the floor. Based on this fact, we can define a set of meta-classes that depend on the position where the acoustic event is detected. The proposed meta-classes and their associated spatial features are: “near door” and “far door”, related to the distance of the acoustic source to the door, and “below table”, “on table” and “above table” meta-classes depending on the *z*-coordinate of the detected AE.

The demonstration of obtained results is an important aspect during any research activity. Our objective is online implementation of AED technologies in the smart-room in order to investigate the potentialities of video and audio perception of the computer systems.

Taking into account great variety audio features proposed in the state-of-the-art literature we aim to propose a theoretically-based systematization to the topic of feature extraction for AED, and also to propose fast and effective approach to select individual features or groups of features for different AEs.

Since the task of AED in meeting room environments is relatively new, there is a scarcity of annotated multimodal data, and, in particular, of data with temporal sound overlaps. Recording and manual annotation of multimodal and making it available for research community is also an objective in our research.

Searching for a suitable fusion approach of different modalities is an important aspect in development of multimodal systems. The information fusion can be done on data, feature, and decision levels. Data fusion is rarely found in multi-modal systems because raw data is usually not compatible among modalities. Concatenating feature vectors from different modalities into one super vector is a possible way for combining audio and visual information. An alternative to feature-level fusion is to model each different feature set separately, to design a specialized classifier for this feature set, and combine the classifier output scores. The two above mentioned fusion approaches will be compared in the presented thesis.

In the context of audio segmentation in broadcast news domain our objective is to find a suitable segmentation strategy, and the best feature set for each acoustic class individually. Quantitative comparison and evaluation of competing approaches is very important in nearly every research and engineering problem since it may help to guess which directions are promising and which are not. So the evaluations have to be organized and coordinated, and appropriate metrics and evaluation tools have to be developed.

1.3 Thesis Outline

The thesis is organized as follows. Chapter 2 presents state of the art in the area of multimodal audio recognition, presenting a literature review from the application point of view, and reporting different features extracted from audio and video modalities, feature selection, classification and detection techniques that have been used so far for multimodal acoustic event detection (AED) and audio segmentation (AS) tasks.

Chapter 3 reports the work done in the area of audio-visual feature extraction. The scheme for grouping of the variety of audio features is presented. Then the feature extraction approach to obtain a set of spectro-temporal features and the localization coordinates of the sound source is detailed. Additionally, a number of features extracted from the video signals by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of several AEs is described.

Chapter 4 describes a few AED systems developed in this thesis. The chapter includes the overview of HMM-GMM, SVM detection systems and the fusion techniques of different modalities. Then the results, obtained with the above-mentioned systems and different databases, are reported. The importance of different modalities using fuzzy theory concept is investigated.

Chapter 5 presents the work done in the broadcast news AS task. A hierarchical system is developed that has been applied for two different databases. The suitability of the presented AS approach has been validated in automatic translation and subtitling application as well as in speaker diarization system. The chapter includes the overview, results and main conclusions from the Albayzín-2010 AS evaluation whose organization was carried out by the author.

Chapter 6 reports the work done in the feature selection problem. A fast one-pass-training technique has been proposed that does not require the re-estimation of acoustic models during the evaluation of the candidate feature sets. Three different feature selection approaches have been compared in the framework of multimodal the meeting-room AED and the broadcast news AS.

Online implementation of the 2-source AED and acoustic source localization systems is presented in chapter 7.

Finally, chapter 8 concludes the work. The main achievements are summarised in this chapter. Several promising future directions are highlighted.

Chapter 2. State of the Art

2.1 Chapter Overview

In this chapter the current state of the art in the area of multimodal feature extraction, feature selection, audio recognition and fusion of different information sources is presented.

The remaining sections of this chapter are organized as follows. Section 2.2 presents literature review of the multimodal acoustic event detection task. Related topics in multimodal audio recognition from the application point of view are discussed in Section 2.3. Sections 2.4, 2.5, 2.6, 2.7 discuss the currently used multimodal features, feature selection, detection and fusion approaches, respectively.

2.2 Acoustic event detection

The notions of acoustic event, acoustic class, environmental sound [MMS10], sounding object [RF03] are used interchangeably in the literature. Their definition can be attributed to Vanderveer [Van79] as "any possible audible acoustic event which is caused by motions in the ordinary human environment; they have real events as their sources; they are usually more "complex" than laboratory sinusoids; they are meaningful, in the sense that they specify events in the environment (…)". Acoustic Event Detection (AED) is usually addressed from an audio perspective and most of the existing contributions are intended for indexing and retrieval of multimedia documents [LZJ02] or to improve robustness of speech recognition [NNM03]. Detection of acoustic events has been carried out in several environments like living environments [DM11], hospitals [VIB03], kitchen rooms [SLP03], bathrooms [JJK05], public places, and in the broadcast news domain. For instance, in [CLH06] ten key audio effects are taken into consideration: applause, car-racing, cheer, car-crash, explosion, gun-shot, helicopter, laughter, plane, and siren. Moncrieff et al. proposed to detect violent events in feature films by analyzing environmental sounds such as gunfire, engines, and explosions [MDV01].

Within the context of ambient intelligence, AED applied to give a contextual description of a meeting scenario was pioneered by [Tem07]. Moreover, AED has been adopted as a semantically relevant technology in CHIL European project [WS09] and several evaluation campaigns [CLE06], [CLE07]. They were an international effort to evaluate systems designed to recognize events, activities, and their relationships in interactive scenarios like lectures or meetings. In the framework of the CHIL project it has been decided that for the chosen meeting-room environment it is reasonable to have an acoustic sound taxonomy for general sound description and a semantic sound taxonomy for a specific task. The proposed acoustic scheme is shown in Figure 2.2.1. Actually, almost any type of sounds can be referred to one of the proposed groups according to its acoustical property. On the contrary, the semantic scheme that is presented in Figure 2.2.2 is very specific to the CHIL meeting-room scenario. Additionally, with two sound taxonomies (acoustic and semantic) it is possible to cope with situations when the produced event does not match any semantic label but can be identified acoustically.

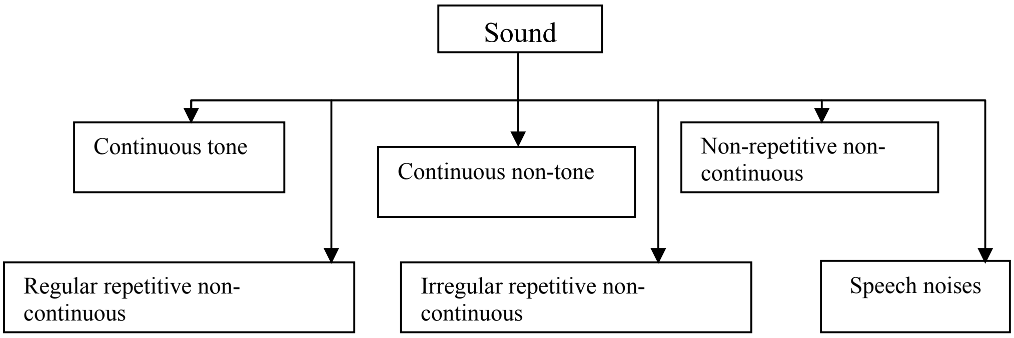

Sound
Continuous tone
Continuous non-tone
Non-repetitive non-continuous
Regular repetitive non-continuous
Irregular repetitive non-continuous
Speech noises

Figure 2.2.1. CHIL acoustic sound taxonomy

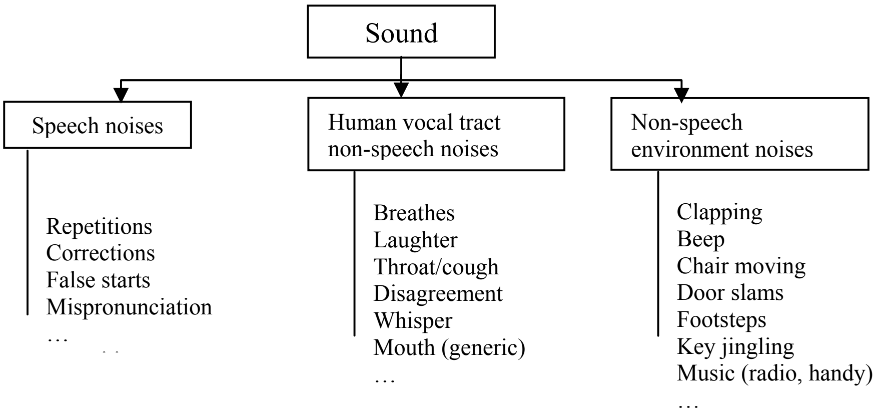


Figure 2.2.2. CHIL meeting-room semantic sound taxonomy

Although much progress has been achieved in the framework of CHIL, there is still a need of a theoretically-based systematization of feature extraction for AED, multi-microphone processing to exploit the spatial diversity of sounds, integration of audio and video information, especially for overlapping sounds, detection of multiple simultaneous acoustic sources, etc. The problem of acoustic overlaps is closely related to the ‘‘cocktail party” problem [WB06]. In that problem, one usually tries to separate one speech source from other; however, in AED we would like to separate acoustic events from speech. Temporal overlaps of several speakers have been considered in the NIST RT-09 [RT09] evaluation campaign, where the involved tasks (e.g. speaker diarization) have been evaluated on overlapped speaker segments as well. In fact, the overlap problem has recently gained a strong interest in speech processing. For instance, in [WBW05], the authors propose several different features and investigate their effectiveness for detection of overlaps of two and more speakers. Also, some improvement in detection of speech overlaps for speaker diarization is shown in [ZSH10].

The idea of using more than one modality arises from two main observations: 1) when one or the other modality is not available the system will still be able to return an AE estimation and 2) when both modalities are available, the diversity and complementarity of the information, should couple with an improvement on the general performance of the system. Most existing researches on multimodal event detection focus on a specific domain, such as the analysis in the sports video domain, as high-level semantic events normally rely much on the domain knowledge. In particular, authors in [LXY06] show the multimodal-based approach that can generate reliable annotation for basketball video which cannot be successfully achieved using a single modality. In [RGA00] the authors developed effective techniques to detect excited announcers' speech and baseball hits from noisy audio signals, and fused them to extract events of exciting segments in baseball programs. In [ZLC07] authors propose an effective fusion scheme of audio and visual modalities for highlight detection in broadcast soccer videos. The extracted annotations are used to build applications for selective browsing of sports videos. Such summarization techniques enable content-based indexing of multimedia documents for efficient storage and retrieval. Since the detecting semantic events is a challenging multimedia understanding problem, one need to depend on multiple modalities to interpret the semantics reliably.

Audio segmentation (AS) in the broadcast news domain can be considered as a specific application of AED. The research works on audio and multimedia content segmentation published so far address the problem in different contexts. The first prominent works are dated from 1996, the time when the speech recognition community moved from the newspaper (Wall Street Journal) era towards the broadcast news (BN) challenge [Pal03]. In the BN domain the speech data exhibited considerable diversity, ranging from clean studio to really noisy speech interspersed with music, commercials, sports etc. This time the decision was made to disregard the challenge of transcribing speech in sports material and commercials. The work from [Sau96] and then from [SS97] are the earliest works that tackled the problem of speech/music discrimination from radio stations. The authors found the first applications of AS in automatic program monitoring of FM stations, and in improvement of performance of ASR technologies, respectively. Both works showed relatively low segmentation error rates (around 2-5%).

Within the next years the research interest was oriented towards the recognition of a broader set of acoustic classes, like in [ZK99] or [LZJ02] where, in addition to speech and music classes, the environment sounds were also taken into consideration. A wider diversity of music genres was considered in [MKP00]. Conventional approaches for speech/music discrimination can provide reasonable performance with regular music signals, but often perform poorly with singing segments. This challenging problem was considered in [CG01]. The authors in [SPP99] tried to categorize the audio into mixed class types such as music with speech, speech with background noise, etc. The reported classification accuracy was over 80%. A similar problem was tackled by [BFM02] and [AMB03], dealing with the overlapped segments that naturally appear in the real-world multimedia domain and cause high error rates. The interest in mixed sound detection in the recent years [IMK08] [DPR09] [LV10] shows it is still a challenging problem.

In the BN domain, where speech is typically interspersed with music, background noise, and other specific acoustic events, audio and multimedia segmentation is primarily required for indexing, subtitling and retrieval. However, speech technologies that work on such type of data can also benefit from the AS output in terms of overall performance. In particular, the acoustic models used in automatic speech recognition or speaker diarization can be trained for specific acoustic conditions, such as clean studio vs. noisy outdoor speech, or high quality wide bandwidth studio vs. low quality narrow-band telephone speech. Also, AS may improve the efficiency of low bit-rate audio coders, as it allows that traditionally separated speech and music codec designs can be merged in a universal coding scheme which keeps the reproduction quality of both speech and music [EGR07].

Multimedia information indexing and retrieval research is about developing algorithms, interfaces, and tools allowing people to search and find content in all possible forms. Current commercial search methods mostly rely on metadata as captions or keywords. On the web this metadata is usually extracted and extrapolated through the text surrounding the media, assuming a direct semantic connection between the two. However, in many cases this information is not sufficient, complete, or exact; in some cases this information is not even available. Content–based methods are designed to search through the semantic information intrinsically carried by the media themselves. One of the main challenges in content-based multimedia retrieval still remains the bridging of the semantic gap referring to the difference of abstraction which subsists between the extracted low level features and the high level features requested by humans' natural queries. Tools for efficient storage, retrieval, transmission, editing, and analysis of multimedia content are absolutely essential for the utilization of raw content. Several applications can benefit from semantic analysis from the multimedia content. Filtering of multimedia content can enable automatic rating of Internet sites and restrict access to violent content. Semantic understanding could mean better and natural interfaces in human computer interaction. Very low bit-rate video coding, summarization, and transcoding are among the several applications that could benefit from semantic multimedia analysis [NH02].

2.3 Applications of multimodal audio recognition

Many audio recognition applications considered video modality as a valuable additional source of information. Here we present the most prominent ones.

2.3.1 Audio-visual speech recognition

Multimodal speech recognition is an extension of the traditional audio speech recognition task. A comprehensive survey on joint processing of audio speech and visual speech can be found in [Che01]. The main motivation of the multimodal speech processing is the fact that human speech is bimodal in nature: audio and visual. While the audio speech signal refers to the acoustic waveform produced by the speaker, the visual speech signal refers to the movements of the lips, tongue, and other facial muscles of the speaker. Such bimodality has two aspects, the production and the perception. Speech is produced by the vibration of the vocal cord and the configuration of the vocal tract that is composed of articulatory organs. Using these articulatory organs, together with the muscles that generate facial expressions, a speaker produces speech. Since some of these articulators are visible, there is an inherent relationship between the acoustic and visible speech.

In audio speech, the basic unit is called a phoneme. In the visual domain, the basic unit of mouth movements is called a viseme, which forms the smallest visibly distinguishable unit of visual speech. There are many acoustic sounds that are visually ambiguous. These sounds are grouped into the same class that represents a viseme, so there is a many-to-one mapping between phonemes and visemes. Both in the acoustic modality and in the visual modality, most of the vowels are distinguishable. The same is not true for consonants, however. For example, in the acoustic domain, the sounds /p/, /t/, and /k/ are very similar. The confusion sets in the auditory modality are usually distinguishable in the visual modality. One good example is the sounds /p/ and /k/, which can be easily distinguished by the visual cue of a closed mouth versus an open mouth. Therefore, for speech understanding, if we can extract the lip movements from the video of a talking person, such information can be utilized to improve speech understanding. This forms the basis for developing the multimodal speech recognition systems that, received a large interest in the last decade.

Audio-only speaker/speech recognition systems are far from being perfect especially under noisy conditions and reverberation. Performance problems are also observed in video-only speaker/speech recognition systems, where poor picture quality, changes in pose and lighting conditions, and varying facial expressions maybe harmful.

State-of-art multimodal speech recognition systems have been jointly using lip information with audio. In the speech recognition literature, audio is generally modelled by mel-frequency cepstral coefficients (MFCC). However for lip information, there are several approaches reported in the literature such as texture-based, motion-based, geometry-based and model-based [DL00]. In texture-based approaches, pure or DCT-domain lip image intensity are used as features. Motion-based approaches compute motion vectors to represent the lip movement during speaking. In geometry-based approaches shape features such as lengths of horizontal and vertical lip openings, area, perimeter, pose angle are selected for lip representation. For model-based approaches, processing methods such as active shape models, active contours or parametric models are used to segment the lip region [CEY06].

2.3.2 Multimodal speaker identification

Person identification is of major importance in security, surveillance, human-computer interfaces. Recently, person identification for smart environments has become another application area of significant interest [EJF06]. Sample application areas can be a smart video-conferencing system that can recognize the speaker; a smart lecture or meeting room, where the participants can be identified automatically and their behaviours can be analyzed throughout the meeting or the lecture. All these applications attempt the recognition of people based on audiovisual data.

The way the systems collect data divides multimodal speaker identification systems into two categories: near-field and far-field systems. In near-field systems both the sensor and the person to be identified focus on each other. In far-field systems the sensors monitor an entire space in which the person appears, occasionally collecting useful data (face and/or speech) about that person. Hence, far-field data streams are corrupted with noise: the video streams contain faces viewed from arbitrary angles, distances, under arbitrary illumination, and possibly, depending on the environment of the deployment, with arbitrary expressions. Similarly, the sound streams suffer from reverberations, large attenuations, and the coexistence of background sounds. The audiovisual environment changes dramatically as the person moves around the space. As a result, the faces collected are tiny (typically of 10 pixels between the eyes) and with gross variations in pose, expression, and illumination. The speech samples are also attenuated, corrupted with all sorts of background noises (occasionally entirely masked by them) and reverberations. Nevertheless, far-field systems have three features that allow them to offer usable recognition rates: the use of multiple sensors (many cameras and microphones); the abundance of training data that are audiovisual streams similar to those on which the system is expected to operate; the possibly long periods of time that the systems can collect data on which they are going to base their identity decision. Note the far-field multimodal person identification systems evaluation was in the scope of CHIL European project.

Multimodal speaker recognition systems existing in the literature are mostly bimodal, in the sense that they integrate multiple features from audio and face information as in [SP03] or from audio and lip information as in [WS01]. An example of a multimodal identification system that uses three different features - face, voice, and lip movement—to identify people is developed in [FD00]. Even if one modality is somehow disturbed—for example, if a noisy environment drowns out the voice—the other two modalities still lead to an accurate identification.

2.3.3 Multimodal emotion recognition

Emotions play an important role in natural human interactions, decision making and other cognitive functions. Current technologies allow exploring the human emotions using not only audio and video modalities, but also other modalities such as the human physiology. Emotion expressed in a piece of media, such as movies or songs, could be used for tasks of indexing and retrieval or automatic summarization. The information about the emotion that better represents a movie could, for example, be used to index that particular movie by genre-like categories (e.g. happiness vs. comedy or fear vs. thriller and horror, etc.). One of the challenging issues is to endow a machine with an emotional intelligence. Emotionally intelligent systems must be able to create an affective interaction with users: they must be endowed with the ability to perceive, interpret, express and regulate emotions [CKC07]. Recognising users' emotional state is then one of the main requirements for computers to successfully interact with humans.

Multimodal approaches to emotion recognition are currently gaining attention of research community. In [PHC10] authors perform emotion recognition by means of fusing information coming from both the visual and auditory modalities. Identification of the six "universal" emotions, i.e. anger, disgust, fear, happiness, sadness, and fear is addressed. Most of the works consider the integration of information from facial expressions and speech and there are only a few attempts to combine information from body movement and gestures in a multimodal framework. Authors in [GP06] for example fused at different levels facial expressions and body gestures information for bimodal emotion recognition. In [KR05] proposed a vision-based computational model to infer acted mental states from head movements and facial expressions.

2.4 Methodology

The design of a multimodal audio recognition system requires addressing three basic issues [CEY06]. The first one is to decide which modalities to fuse. The word "modality" can be interpreted in various ways; in meeting-room acoustic event detection it usually refers to a specific type of information that can be deduced from signals. In this sense, spectro-temporal content of audio signal and the acoustic localization information obtained from microphone array can be interpreted as two different modalities existing in audio signals. Likewise, video signal can be split into different modalities, face and motion being the major ones.

The second issue is how to represent the raw data for each modality with a discriminative and low-dimensional set of features and, in conjunction with this, to find the best matching metric in the resulting feature space for classification. This step also includes a training phase through which each class is represented with a statistical model or a representative feature set. Curse of dimensionality, computational efficiency, robustness, invariance, and discrimination capability are the most important criteria in selection of the feature set and the classification methodology for each modality.

The third issue is how to fuse different modalities. Different strategies are possible. In the so-called "early integration" modalities are fused at data or feature level, whereas in "late integration" decisions or scores resulting from each monomodal classification are combined to give the final conclusion. This latter strategy is also referred to as decision or opinion fusion and is effective especially in case the contributing modalities are uncorrelated and thus the resulting partial decisions are statistically independent. Multimodal decision fusion can also be viewed from a broader perspective as a way of combining classifiers, which is a well-studied problem in pattern recognition. The main motivation here is to compensate possible misclassification errors of a certain classifier with other available classifiers and to end up with a more reliable overall decision.

The general structure of a multimodal audio recognition system can be described with a block diagram, as it is shown in Figure 2.4.1. From the multimodal data, a number of characteristic features are extracted, the most important are selected which are then classified with some sort of pattern classifier. In the case of "late" fusion the recognition results from different detection systems can be posteriorly fused to obtain more reliable estimate.

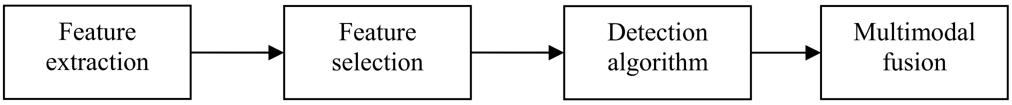


Figure 2.4.1. General block diagram of a sound detection system.

2.4.1 Feature extraction

One major issue in building an automatic audio recognition system is the choice of proper signal features that are likely to result in effective discrimination among different AEs. The appropriate choice of features is crucial in building a robust recognition system. The main source of additional information for the meeting-room AED is video data from multiple cameras. The visual component is considered to be a valuable supplementary information source in noisy environments because it remains unaffected by acoustic noise.

2.4.1.1 Audio features

Audio feature extraction serves as the basis for a wide range of audio technologies. In audio recognition many features are proposed which describe the acoustic content of the sound such as sub-band energies computed in short-time windows, time evolution parameters, modulation spectrum, level of harmonicity, etc [Pee03] [LSD01] [TN09]. Although in speech recognition the MFCC features (or alternative features which have a lot in common with them) became the de-facto standard for front-ends in many applications, the situation with audio recognition in general is not so clear yet. Very often authors do not present strong or clear arguments in favour of a particular feature set they propose, and the final decision about feature subset selection is mainly based on their prior knowledge. For instance, for music detection, the features that capture the harmonicity content of the waveform are preferable [SK04], while for classification of generic sounds the features which model the spectral envelope are widely used [LSD01]. For instance, in [SK04] harmonic concentration and harmonic energy entropy features are proposed, and in [INM06] a periodicity measure based on the fundamental frequency is estimated independently in each sub-band to obtain periodic and non-periodic features. In speech-song classification, authors from [Ger02] propose pitch as a feature due to the specific structure of a song.

The temporal evolution of the above mentioned features may be characterized by duration and dynamics features. In the later case, traditionally the first and the second temporal derivatives of the above mentioned features are used [Tem07] [SK04].

There are several attempts for grouping of audio feature. Broadly, acoustic features can be grouped into two categories: time-domain (or temporal features) and frequency-domain (or spectral features) [CNK09]. Time-domain features are computed from the audio waveforms directly and characterize the signal in time domain. Frequency-domain features are derived from the Fourier transform of the time signal over a frame, and characterize the spectral content of the signal. Similarly, in [UKR07] the authors classify features which commonly exploited for audio event detection into time domain features, transformation domain features, time-transformation domain features or their combinations. In [Pee03] the author distinguishes temporal, temporal shape, energy, energy shape, harmonic and perceptual features.

We can also distinguish frame-based and segment-based features. The frame-based features usually describe the spectrum of the signal within a short time period (10-30 ms), where the process is considered stationary. The concept of the audio frame comes from traditional speech signal processing, where analysis over a short time interval has been found to be appropriate. MFCCs and PLPs are examples of frame-based features routinely used in speech recognition, which represent the spectral envelope and also its temporal evolution. To extract the semantic content, we need to observe the temporal variation of frame features on a longer time scale segments. The length of the segment may be fixed (usually 0.5sec – 5 sec) or variable. Although fixing the segment size brings practical implementation advantages, the performance of a segmentation system may suffer either from the possibly high resolution required by the content or from the lack of sufficient statistics needed to estimate the segment features due to the limited time span of the segment. According to [KQG04] a more efficient solution is to extract global segments within which the content is kept stationary so that the classification method can achieve an optimum performance within the segment. The most usual segment-based features are the first and second order statistics of the frame-based features computed along the whole segment. Sometimes high-order statistics are taken into consideration, like skewness and kurtosis as well as more complex feature combinations that capture the dynamics of audio (e.g. the percentage of frames showing less-than-average energy), rhythm (e.g. periodicity from the onset detection curve), timbre or harmonicity of the segment [LT07].

Features may be characterized depending on the application where they are used. Two major groups could be discussed: Automatic Speech Recognition (ASR) features, which were initially designed for speech recognition applications, and the often so-called perceptual features, which have shown a high success in specific applications (recognition of certain classes of sounds or environments) [SK04] [Ger02] [PRA02]. In [Tem07], a set of perceptual features was used which showed to be successful for AED: frame energy, the silence ratio, sub-band energies, zero-crossing rate, high zero-crossing rate ratio, low short-time energy ratio, spectrum flux, band periodicity, noise frame ratio, fundamental frequency, spectral centroid, spectral roll-off and spectral band-width. Other research work [PRA02], proposes features for robust speech/music segmentation based on particular properties of those two classes: modulation entropy, 4 Hz modulation energy and duration features. Experiments showed that speech carries more modulation entropy as well as modulation energy than music, and for music the duration of stationary segments is higher; actually both facts are correlated.

In [Ara08], a feature vector is formed from the estimates of posterior probabilities over the set of classes. This vector is then used as input to state-of-the-art acoustic models (mainly HMM-GMM based systems). Given the discriminative nature of posterior features, the acoustic models can use significantly fewer parameters for a similar recognition performance.

In [CNJ08] the authors propose a novel method based on matching pursuit (MP) for feature extraction to analyze environment sounds. The MP-based method utilizes a dictionary from which to select features, and the results showed that this method is promising in classifying unstructured audio environments, especially when MP-based features are used in combination with state-of-the-art MFCC features.

Recently, the interest in technological development for designing automatic systems to mimic human capabilities has evolved and intensified. It is well known that an automatic speech recognition system performs far less reliably than a human listener under adverse conditions. There is still a lot to be learned in understanding the reason why a human subject would perceive and recognize sounds as reliably as witness. A major aim of psychoacoustic research is to establish functional relationships between the basic physical attributes of sound, such as intensity, frequency and changes in these characteristics over time, and their associated perceptions.

Many researchers have thus made various efforts on emulating human speech perception in order to achieve human-like performance. Some approaches model and rebuild the functionalities of the auditory nuclei. These are exemplified by research on basilar membrane and electronic cochlea [PH96]. Most of these types of systems have similar structures as the human auditory pathway. However, they model only a small fraction of perceptual functionality in the human auditory system. This is because the human auditory system is structurally very complex, consisting of many feedback paths from higher levels in the neural system. Presently, much of the structure and mechanism, especially interactions among auditory nuclei, remains unknown.

In [MJ08], both functional and structural aspects of perception/cognition which are essential to the understanding of most auditory phenomena are discussed. In this work the authors suggested 3 essential stages of auditory perception: the sensory stage, the primary neural processing stage, and the perceptual inference stage with intermediate signal representations. This architecture allows for a nearly independent construction of signal processing models within the three stages, and so to construct a chain of processing from a stimulus to the corresponding perceptual response.

In [PH96], a model of the auditory processing has been developed to analyse everyday sounds like music, speech and sound environments in which they occur. The model transforms a complex sound into a multi-channel activity pattern like that observed in the auditory nerve, and then it converts this neural activity pattern into an “auditory image” feature that is intended to represent human’s initial impression of the sound. The initial stage that generates the neural activity pattern is a cochlea simulation that is composed of two processing modules [PH96]:

1. A gamma-tone auditory filterbank, which performs a spectral analysis and converts the acoustic wave into a multi-channel representation of the basilar membrane motion.

2. A two-dimensional adaptation mechanism that “transduces” the membrane motion and converts it into a multi-channel representation of the neural activity pattern arriving at the cochlear nucleus.

The auditory gamma-tone filter bank incorporates two insights from the auditory physiology: 1) a higher frequency resolution for low frequencies 2) a higher temporal resolution for high frequencies. The gamma-tone filter bank-based feature extraction technique is also presented in [HAA07] for classification of everyday sounds like opening/closing the door, or footstep sounds. Two types of perceptual features were extracted, based on the combination of gamma-tone filters with the Hilbert transform, and with Meddis’ inner hair cell model. In the presented work, the authors compare psychoacoustic features with classical MFCC features, and the experiments show that, in general, the gamma-tone based representations outperform other pre-processing methods.

Other feature extraction techniques model even a smaller part of the psychoacoustic process, since they concentrate on particular transformations of the input signal. For instance, in [BC00] the feasibility of applying human speech perceptual characteristics to enhance the recognition accuracy and robustness in Mandarin vowel recognition is shown. In this study, a perceptual processing is applied to the Fourier spectrum to obtain the so called perceptual spectrum. The proposed perceptual speech processing is based on three perceptual characteristics, and consists of three independent processing steps: the masking effect, the minimum audible field renormalization, and the mel-scale resampling. Furthermore, to reduce the feature dimensionality, and at the same time retain the relevant information, a phonetic feature mapping to obtain the so called phonetic features was proposed. These new features are basically similarity measures between the input perceptual spectrum and a set of nine reference perceptual spectra.

In the PLP technique [Her90], the short-term spectrum of the speech is modified by several psycho-physically based transformations based on three concepts from the psycho-physics of hearing: the critical-bands spectral resolution; the equal-loudness curve, and the intensity-loudness power law. This technique is currently used in many speech recognition systems for feature extraction.

In [DPK96], a quantitative model for signal processing in the auditory system is described. This model combines several stages (preprocessing, adaptation and adding the internal noise) which simulate aspects of transformation in the auditory periphery with an optimal detector as the decision device.

The author in [Kle03] [Kle02] motivates an approach to feature extraction for automatic speech recognition which utilizes two-dimensional spectra-temporal modulation filters. Results from physiological and psychoacoustic studies indicate that spectrally and temporally localized time-frequency envelope patterns are the relevant basis for auditory perception. In this study the feature vector size is relatively high (more than 2000), so it is important to search for an efficient automatic feature selection procedure that yields smaller feature sets.

Although the modelling of the human perception system is an extremely difficult task, it may be still beneficial to model even a small part of transformations in the pathway between the entering of acoustic waveforms in the human's ear and the final perception in brain. Moreover, using psychoacoustic model is strongly motivated in numerous studies [HAA07] [BC00] [Her90] [Kle02], where authors present promising results in audio recognition.

Here, we present some of the commonly used audio features used in audio recognition applications. We use the following notation in feature definition:

$s(n)$ – signal value at the time index n;

N – frame length;

$f(i)$, $a(i)$ – frequency value at the frequency bin i and the corresponding Discrete Fourier Transform (DFT) amplitude, respectively;

$x(k)$ – value of mel-scaled logarithmic filter-bank energy at the sub-band frequency index k corresponding to the current frame.

Zero crossing rate (ZCR) Zero-crossings occur when successive samples have different signs, and the ZCR rate is the average number of times the signal changes its sign within a frame

$$ZCR = \sum_{n=0}^{N-1} I\{s(n)s(n-1) < 0\} \qquad (2.4.1)$$

where the indicator function *I{A}* is 1 if its argument *A* is true and 0 otherwise.

Short-time energy (STE) Short-time energy provides a convenient representation of the amplitude variation over time. It is defined as:

$$STE = \sum_{n=0}^{N-1} s(n)s(n) \qquad (2.4.2)$$

Sub-band log energies. The 4 sub-bands are equally distributed along the 20 mel-scaled filter bank energies (FBEs) (5 per sub-band). The energy of each sub-band is calculated as:

$$SBE(j) = \sum_{k=5j}^{5j+N-1} x(k) \quad \text{for } j = 0,...,3 \qquad (2.4.3)$$

where *N=5* is the number of log FBEs per sub-band.

Spectral centroid. The centroid is a measure of the spectral "brightness" of the spectral frame and is defined as the linear average frequency weighted by DFT amplitudes, divided by the sum of the amplitudes:

$$CE = \frac{\sum_{\forall i} f(i)\,a(i)}{\sum_{\forall i} a(i)} \qquad (2.4.4)$$

Spectral roll-off. This measure quantifies the frequency bin f_c at which the accumulative value of the frequency response magnitude reaches a certain percentage of the total magnitude. A commonly used threshold is *c*=95%.

$$\sum_{i=0}^{f_c} a(i) = \frac{c}{100} \sum_{\forall i} a(i) \qquad (2.4.5)$$

Spectral Flux. It is used to measure a spectral amplitude difference between two successive frames**.**

Spectral bandwidth measures the width of the range of signal's frequencies.

$$BW = \sqrt{\frac{\sum_{\forall i} (f(i) - CE)^2 a^2(i)}{\sum_{\forall i} a^2(i)}} \qquad (2.4.6)$$

where *CE* is the *spectral centroid* of the frame.

Chroma features. Chroma features [Fuj99] are the powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave. Since, in music, notes exactly one octave apart are perceived as particularly similar, knowing the distribution of chroma even without the absolute frequency (i.e. the original octave) can give useful musical information about the audio and may even reveal perceived musical similarity that is not apparent in the original spectra. In this thesis chroma features are computed for 12 semitones from a short-time FFT spectrogram with window-size 50 ms, rate 10 ms, and using Gaussian window.

Amplitude modulation features. The authors in [BAL05] propose an approach to sound classification that aims to mimic the human auditory system at least partially by making use of auditory features as known from auditory scene analysis. Among others, they propose amplitude modulation features. The possible way of describing amplitude modulations in various natural sound sources is using amplitude histograms that can be modelled by means of percentiles. The 50% percentile P_{50}, for example, shows the level below which the envelope is 50% of the time.

The distances between the percentiles may thus be the basis for more complex features. They are normalized to the 50% percentile; the distance between the 10% and the 90% percentiles for example is calculated as follows:

$$d_{90-10} = \frac{P_{90} - P_{10}}{P_{50}} \qquad (2.4.7)$$

A number of features are presented in the following that appear to be valuable for the description of the form of the histogram:

1) *Width*. The width of the histogram is well described by the distance between the 90 % and the 10% percentile:

$$width = d_{90-10} \qquad (2.4.8)$$

2) *Symmetry.* The symmetry can be investigated by looking at the difference:

$$symmetry = (d_{90-50}) - (d_{50-10}) \qquad (2.4.9)$$

The symmetry is near zero for symmetrical distributions, positive for left sided distributions, and negative for right sided distributions. Impulse-like signals are asymmetric right-sided due to the signal pauses.

3) *Skewness*. The skewness of the histogram can be regarded as the difference between the 50% percentile and the median:

$$skewness = P_{50} - \tilde{x} \qquad (2.4.10)$$

The median is estimated by the mean between the 10 % and the 90 % percentile:

$$\tilde{x} = \frac{P_{90} + P_{10}}{2} \qquad (2.4.11)$$

For asymmetrical distributions the difference between P_{50} and the approximated median should be large, for symmetrical distributions approximately zero.

4) *Kurtosis*. The kurtosis corresponds to the approximation

$$kurtosis = \frac{P_{70} - P_{30}}{2(P_{90} - P_{10})} \qquad (2.4.12)$$

which sets the middle 50% interval in relation to the range of the distribution, indicating whether the distribution has a narrow or a broad peak.

5) Lower half. The distributions in the lower half of the histogram are expressed by the difference*:*

$$lower\ half = (d_{50\text{-}30}) - (d_{30\text{-}10}) \qquad (2.4.13)$$

The lower half of the distribution allows to characterize right-sided distributions by encoding the relations between the lower and upper half (that is, below and above P_{30}) of the lower half of the total distribution (that is, below P_{50}). For impulse-like signals, this feature will have a large value, for continuous signals it will be approximately zero.

We will call these features as perceptual throughout the work, since it has a more perceptually-oriented profile than the conventional features taken from ASR.

Among ASR feature which are very popular in audio recognition tasks the MFCC features are the most widely used. In this thesis, **frequency-filtered (FF) log filter-bank energies** [NMH01] are employed. The feature extraction procedure consists of applying, for every frame, a short-length FIR filter to the vector of log filter-bank energies vector, along the frequency variable. The transfer function of the filter is $z - z^{-1}$, and the end-points are taken into account. That type of features has been successfully applied not only to speech recognition but also to other speech technologies like speaker recognition [LH08].

2.4.1.2 Video features

Considering the visual signal, feature extraction depends heavily on the type of video data: synchronized recordings of multiple calibrated cameras or close-up video recordings of subject's face. Traditional methods for processing video have focused on analyzing individual frames independently, or possibly adjacent frames such as optical flow. Features are typically extracted for each frame independently, and then subsequently linked together temporally. Other methods like the one described in [Ke08] consider video as three-dimensional volumes, and thus the fundamental processing unit should be 3D blocks consisting of many frames.

Visual features can be extracted on object color, texture, shape and motion [ALM03]:

1) Color is the most widely used visual feature in video retrieval. Color features can include color histogram, dominant color, mean and standard deviation of colors.

2) Texture also is an important feature of a visible surface, where repetition or quasi-repetition of a fundamental pattern occurs.

3) Shape features that are related to the shape of the objects in the image are usually represented using traditional shape analysis such as moment invariants, Fourier descriptors, etc.

4) Motion is another useful visual cue. Theoretically, it is invariant to changes of color and lighting. Motion features include motion histogram/phase correlation, dominant motion, and model parameters for global motion description.

A preliminary inspection show that visual features are not very robust regarding changes in illumination, variability in object viewpoint, unpredictable object motion and so on.

For the task of event detection, shape and motion features are preferable since they are discriminative and robust to variations. Both of them capture how people and other surrounding objects deform and move through space-time, thus enabling to recognize an event. Robust detection of events thus requires robust tracking of an object's state. In particular, a very promising approach is presented in [CSC08] which uses a video 3D tracking algorithm, where a Particle Filtering (PF) technique is used to estimate the location of each person inside the room at a given time *t*. Using the 3D tracking algorithm, apart from the position of the participants in the room, their velocity and acceleration can be estimated, which may be useful to describe activities as standing, walking or running. Moreover, the *z*-component may give indications about person's sitting/standing.

Video features could be extracted on various levels: low-level, middle-level, object-level [WCC05]. For low-level features such as dominant color, motion vectors are acquired directly from the input videos by using simple feature extractions, which usually possess limited capabilities in presenting the semantic contents of the video events. In contrast, object-related features are attributes of the objects such as ball location and player shapes, which greatly facilitate the high-level domain analysis. However, their extraction is usually difficult and computationally costly for real-time implementation. Middle-level features offer a reasonable trade-off between the computational requirements and the resulting semantics.

In [XMZ02], video features are based on energy redistribution between consecutive frames in the video sequence and they were useful for classifying basketball video. In [ZWR03], other motion features such as center of motion, wideness of motion, intensity of motion were effectively utilized for off-line segmentation and recognition of actions in meeting scenarios. In [Ke08], volumetric features based on video's optical flow for visual event detection are presented. Video is thought as a group of 3D volumes and decomposed into 3D subregions. The recognition of spatio-temporal events in video is done by means of matching of individual volumes. In [HDG06], the problem of segmentation and recognition of sequences of multimodal human interactions in meetings is addressed. Authors propose person-specific video features: head vertical centroid, head eccentricity, right hand horizontal centroid, right hand angle, right hand eccentricity, head and hand motion, global motion features from each seat. These features were calculated from predefined positions in the meeting-room. Used in conjunction with audio and semantic features, they showed a superior performance. In laughter detection [PP08], the video features capture the facial expression dynamics. These features are 20 facial points (corners/extremities of the eyebrows, the eyes, the nose, the mouth and the chin) which are tracked using a particle filtering tracking scheme. Then these features were transformed using Principle Component Analysis (PCA) to reflect rigid-movement aspects of data.

2.4.2 Feature selection

Most previous efforts utilize a combination of some, or even all, of the aforementioned features, to characterize audio signals. However, adding more features is not always helpful. As the feature dimension increases, data points become sparser and there are potentially irrelevant features that could negatively impact the classification result. This in turn leads to the issue of selecting an optimal subset of features from a larger set of possible features to yield the most effective subset for acoustic event detection tasks.

The feature selection problem still remains very important and challenging in pattern recognition. Selecting proper features is key to effective system performance. When we have two or more classes, feature selection consists of choosing those features which are most effective for showing class separability [Fuk72]. There are many research works devoted to this problem [GE03] [KJ97] [GS08]. Two main conclusions could be formulated as followed:

1. *The probability of misclassification of a decision rule does not increase as the number of features increases, as long as the class-conditional densities are completely known.*

This means that an additional feature can never decrease the performance of the optimum Bayes classifier. But it is observed in practice that the added features actually degrade the performance of a classifier if the number of training samples that are used to design the classifier is small relative to the number of features, or when the class-conditional probability with many parameters is difficult to estimate.

2. *In general no non-exhaustive sequential feature selection procedure can be guaranteed to produce the optimal subset* [CC77].

If d is a number of features, then any ordering of the classification errors of each 2^d feature subsets is possible. Only exhaustive search can guarantee the optimal feature subset, so the research efforts are directed towards developing techniques which can filter irrelevant and highly correlated features in order that the search space decreases. Sometimes certain heuristics are proposed to decrease the search space [GS08].

In [KJ97], the concept of *weak* and *strong* relevance is presented: a feature X is *strongly* relevant if removal of X alone results in performance deterioration of an optimal Bayes classifier. A feature X is *weakly* relevant if it is not strongly relevant and there is exists a subset of features, S, such that the performance of a Bayes classifier on S is worse than the performance on $S \cup \{X\}$. A feature is irrelevant if it is not strongly or weakly relevant. It is argued that the optimal feature set should include strong and possibly some weak relevant features.

A search of the optimal feature set requires a state space, an initial state, a termination condition, and a search engine [KJ97]. The state space include possible combinations of features, and the search is terminated after finding a feature set with the highest value of the evaluation (objective) function *J*(.). Depending on the way of calculating the objective function, the feature selection approaches are divided into:

1. Filter approaches (score feature subsets independently of the chosen classifier).

2. Wrapper approaches (utilize the learning machine of interest as a black box to score subsets of features).

3. Embedded methods (feature set scoring is done during the process of training a learning machine).

Most filter approaches attempt to identify and remove as much irrelevant and redundant information as possible prior to learning, as a pre-processing step. The main disadvantage of the filter approach is that it totally ignores the effects of the selected feature subset on the performance of the classification algorithm.

The feature selection search engine conducts the search in the space of all possible states. In [JDM00], the following search engines are summarized:

1. Exhaustive Search

2. Branch-and-Bound Search

3. Best Individual Features

4. Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS)

5. Plus "*l*-take away *r*" Selection

6. Sequential Forward Floating Search (SFFS) and Sequential Backward Floating Search (SBFS).

Sequential forward and backward selection are widely used to select features, with SFS being used more often due to the lesser magnitude of calculations involved [Pir04]. Two forms of sequential selection are described in [KJ97]: best first search and hill-climbing search. In [KJ97], the authors introduced a way to change the search space topology by creating dynamic operators (compound operators) that reduce the search by means of dynamically selecting promising feature sets during evaluation. When the number of features is very high, the search of the best individual features is used as a pre-processing step.

In literature, linear transformation of the original features is often applied prior to feature selection. More precisely, if $X=[x_1 \ldots x_n]$ is the feature vector in the original feature domain, after applying a transformation matrix A, we obtain the feature in the transformed domain $Y=[y_1 \ldots y_m]$

(m<n). The main purpose of this transformation is, firstly, to find a new feature space where classes show better separability, and secondly, to reduce the feature space dimensionality from n to m. In PCA [DHS00], the transformation matrix A transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each subsequent component accounts for as much of the remaining variability as possible. In the Independent Component Analysis (ICA) approach [DHS00] [Com94], the transformation matrix is selected in such a way that features after transformation become maximally independent. In the Linear Discriminant Analysis (LDA) approach [Fuk72], the transformation matrix A is found through an optimization criterion of separability of classes which is formulated in terms of both the within-class and the between-class scatter matrices. Sometimes, it might be desired to pick a subset of the original features rather than find a mapping that uses all the original features [CXZ07]. Other linear transforms for dimensionality reduction are presented in [VGB08]: Maximum Likelihood Linear Transform (MLLT), Heteroscedastic Linear Discriminant Analysis (HLDA), and Smoothed HLDA (SHLDA). The experiments performed on the ASR AURORA2 setup, a standard digit speech recognition task, showed that, when using approaches such as MLLT, HLDA, SHLDA, PCA and concatenated schemes, no remarkable improvement is achieved compared to LDA alone.

The method presented in [KS96] is based on eliminating features if they give a little or no additional information to remaining features in the original feature domain. The main idea is to find a subset of original features $G \subset F$ that keep the conditional distribution of class C almost unchanged: $Pr(C|F) \approx Pr(C|G)$. This means that all features in F, which are not included in G are non-informative. To compare two probability distributions, the *KL*-distance measure $D(.)$ is proposed. If f_G is the projection of the variable f from F onto the variables in G then $\delta = D(Pr(C|f), Pr(C|f_G))$ should be as small as possible. This theoretically justified model for optimal feature set selection requires finding the subset G in a practical way. The authors propose the idea of the Markov blanket to find all non-informative features F_i. In the study, the authors heuristically choose the Markov blanket based on selecting a set of K features which are strongly correlated with F_i.

In [Fuk72], the author proposes to find the best feature set with reference to the Bayes classifier. Then the class separability becomes equivalent to the probability of error due to the Bayes classifier, which is the best we can expect. A major disadvantage of the probability of error as a criterion is the fact that an explicit mathematical expression is not available, except for a very few special cases. Even for normal distributions, the calculation of the error requires a numerical integration. To avoid numerical integration, the upper and lower bounds of probability of error are used instead. In this study, the Bhattacharyya distance and the divergence are discussed as a feature selection criterion for normal distribution, which have a direct relationship to the probability of error. The extension to a multi-class problem is proposed.

In [HS98], the authors present a feature selection heuristic that takes into account the usefulness of individual features for predicting the class label along with the level of inter-correlation among them. The hypothesis on which the heuristic is based is as follows: "Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other". Authors propose to evaluate the objective function which takes into account the features average correlation with class and average inter-correlation. The entropy measure is used to estimate the above mentioned correlations. A similar idea is presented in [PLD05], where a criterion function named "minimal-redundancy-maximal-relevance" (mRMR) is constructed which tries to find features with maximal statistical dependency of target class on the data distribution, and with minimal mutual redundancy among them.

In [ZZH08], the authors rank features according to their individual Bayes accuracy on the training set. Firstly the features are decorrelated using the PCA transformation, and afterwards the discriminant capability of each feature is measured the using *hard_bayesian* and *soft_bayesian* functions defined as:

$$F_{hard} = \frac{1}{T}\sum_{t=1}^{T} \delta(\arg\max_{k} P(x_t \mid \omega_k) - l(t)) \qquad (2.5.1)$$

$$F_{soft} = -\sum_{t=1}^{T} Rank_t(l(t)) \qquad (2.5.2)$$

where T is the number of training instances; $l(t)$ is the true label for t^{th} instance; $\delta(.)$ is the Dirac delta function; ω_k is the class label; $Rank_t(.)$ is the likelihood of the true label $l(t)$ on the data point $x_{t.}$

2.4.3 Detection approaches

Detection of acoustic events can be performed in three different ways. The first one is based on detecting the sound boundaries and then classifying each end-pointed segment. Hereafter we refer to it as the detection-and-classification approach. For example, in [Pfe01] an approach based upon exploration of relative silences has been proposed. A relative silence is considered as a pause between important foreground sounds. A different type of segmentation algorithm, which does not require any a-priori information about the particular acoustic classes, is based on the BIC [CG98]. It assumes that the sequence of acoustic feature vectors is a Gaussian process, and measures the likelihood that two consecutive acoustic frames were generated by two processes rather than a single process.

The second approach consists of classifying consecutive fixed-length audio segments. We will refer to it as the detection-by-classification approach. A raw segmentation output is obtained in that case as a direct byproduct of the sequence of segment labels given by the classifier. However, to improve the segmentation (detection) accuracy, some kind of smoothing is required, assuming it is improbable that sound types change suddenly or frequently in an arbitrary way. Many publications give preference to the latter approach due to its natural simplicity. As an example, [Sau96] used multivariate Gaussian classifier to obtain a sequence of decisions, [LZJ02] applied a KNN-based classifier, and [BFM02] used an MLP-based classifier in the experiments.

In the third approach, segmentation and classification are done jointly. For instance, in its decoding step, the HMM-based method attempts to find the state sequence (and, consequently, acoustic class sequence) with the highest likelihood given a sequence of observed feature vectors. The most common procedure for doing that is Viterbi decoding, which uses a dynamic programming algorithm to find in recursive way the most probable sequence of HMM states. The HMM-based audio segmentation approach borrowed from speech/speaker recognition applications has been successfully applied in [ZK99], [AMB03], [LV10] and many other works.

All above mentioned detection approaches require a classification algorithm. A classifier that assigns the class label with the largest posterior probability, Bayes classifier, is the most natural choice, but in real-life problems we do not know the true prior probabilities nor the class conditional pdfs, so we can only design flawed versions of the Bayes classifier. Statistical pattern recognition provides a variety of classifier models that are effectively used in audio recognition tasks. There is no consensus on a single taxonomy of classification methods. In [Lip91] the author lists five types of classifiers: probabalistic (linear discriminant classifier, quadratic discriminant classifier, Parzen), global (multilayer perceptron), local (radial basis function neural networks (RBF)), nearest-neighbor type (k-nn, learning vector quantization neural networs) and rule-forming (binary decision trees, rule-based systems). In the work [HKL97] the authors consider another grouping: classifiers, based on density estimation and classifiers based on regression.

In the following sub-Sections three widely used classification algorithms employed in this thesis are briefly described.

2.4.3.1 Gaussian Mixture Models (GMMs)

Gaussian mixture models are quite popular in speech and speaker recognition. In the design step, we have to find the probability density functions that most likely have generated the training patterns of each of the classes, assuming that they can be modelled by mixtures of Gaussians.

In the GMM, the likelihood function is defined as

$$p(x) = \sum_{i=1}^{K} p_i N(x; \mu_i, \Sigma_i) \qquad (2.4.1)$$

where K is the number of Gaussians, the weights p_i verify

$$\sum_{i=1}^{K} p_i = 1 \text{ and } p_i \geq 0, \forall i \qquad (2.4.2)$$

and $N(x; \mu, \Sigma)$ denotes the multivariate Gaussian distribution

$$N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{|x|}{2}} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu) \right) \qquad (2.4.3)$$

being μ the mean vector and Σ the covariance matrix (often considered diagonal). As the goal is to maximize the likelihood (ML), the parameters of the GMM ($w_i, \mu_i, \text{and} \Sigma_i$) are obtained via the Expectation-Maximization (EM) algorithm [RJ93]. A GMM-based classifiers can handle an arbitrary number of classes.

2.4.3.2 Hidden Markov Models (HMMs)

An HMM is a doubly stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observations.

An HMM is characterized by a set of parameters *λ=(A, B, Π)* [HLW05], where

- $A = \{a_{i,j}\}$ is the state transition probability matrix, and $a_{i,j}$ is the transition probability from state i to state j satisfying $0 \leq a_{i,j} \leq 1$, $\sum_{j=1}^{N} a_{i,j} = 1$, and N is the number of states in the model;
- $B = \{b_i(x_t)\}$ is the observation probability matrix, where $b_i(x_t)$ is the observation probability of feature x_t at state i;
- $\Pi = \{\pi_i\}$ is the initial state distribution.

An HMM can be continuous or discrete. In the continuous model, observation probabilities of the feature vectors are characterized using a parameterized form. A common approach is to use Gaussian Mixture Model (GMM), by which the probability is described as $b_j(x_t) = \sum_{k=1}^{K_j} p_{jk} N(x_t, \mu_{jk}, \Sigma_{jk})$, where K_j is the number of mixtures in state j, $p_{j,k}$ is the mixture coefficient for the k^{th} mixture and $N(.)$ is a Gaussian function with mean vector $\mu_{j,k}$ and covariance matrix Σ_{jk}.

An HMM is trained for every class and the training process follows the Baum-Welch method [RJ93]. The initial parameters of A and B are chosen randomly and the initial values of Π are uniformly distributed for each state. After training, we have $\lambda_1, \lambda_2, \ldots, \lambda_K$, where K is the number of acoustic classes. For each testing sequence X, the likelihood $P(X, \lambda_i)$, $i = 1, \ldots, K$ is computed for each class, and the testing sequence is classified to the class with the maximum likelihood.

A “one-pass” technique [NO99] developed for speech recognition is used for detection, which simultaneously determines the optimal state sequence and AE class sequence. To hypothesize an AE sequence $\Omega = c_1, c_2, \ldots, c_L$, where L is the number of scene transitions in the sequence, we can imagine a super HMM that is obtained by concatenating the HMM’s for different AE classes. The search space can be described as a super network consisting of all states of all classes where the best state transition path has to be found. The search has to be performed at two levels simultaneously: at the state level within a class and at the class level. The paths at the two levels can be searched efficiently by dynamic programming.

The one-pass dynamic programming method searches the optimal AE sequence for a given observation sequence $X= \{x_1, x_2, \ldots, x_T\}$. The algorithm requires two arrays:

1) $Q(t,s;c)$ is the score of the best path up to time t that ends in state of class c.

2) $B(t,s;c)$ is the start time of the best path up to time t that ends in state s of class c.

As illustrated in Figure 2.4.1, the path is searched within the class and among the classes. Within the class, the recurrence equation is as follows:

$$s_m = \arg\max_{0 \le s' \le N(c)} \{p(x_t, s \mid s'; c) \cdot Q(t-1, s'; c)\} \qquad (2.4.4)$$

$$Q(t,s;c) = p(x_t, s \mid s_m; c) \cdot Q(t-1, s_m; c) \qquad (2.4.5)$$

$$B(t,s;c) = B(t-1, s_m; c), \qquad (2.4.6)$$

$1 \le s \le N(c)$

where $N(c)$ is the number of states in class c, s_m is the optimum predecessor state for the hypothesis $(t,s;c)$.

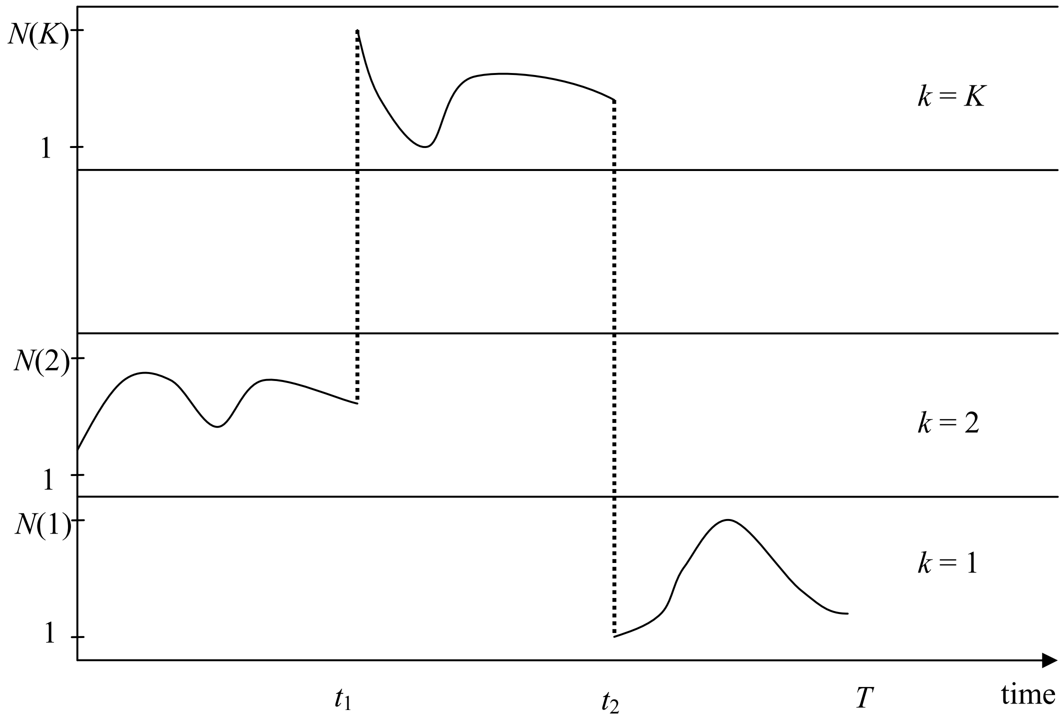


Figure 2.4.1. Illustration of transition between states within a class and between classes. N(k) means the number of states in the class k

In (2.4.4) the partial score $p(x_t, s \mid s'; c)$ between state *s´* and *s* at time *t* is defined as:

$$p(x_t, s \mid s'; c) = a_{s's} b_s(x_t) \qquad (2.4.7)$$

where $a_{s's}$ is the state transition probability from state *s´* to *s* and $b_s(x_t)$ is the observation probability of feature x_t at the state s in the class *c*.

To hypothesize the potential AE boundary, the termination quantity *H*(*t;c*), a class traceback pointer *R(t;c)*, and a time traceback pointer *F*(*t;c*) are introduced as:

$$S_b = \arg\max_{1 \le s \le N(b)} Q(t, s; b) \qquad (2.4.8)$$

$$R(t;c) = \arg\max_{1 \le b \le K, b \ne c} \{p(c \mid b) \cdot Q(t, S_b; b))\} \qquad (2.4.9)$$

$$H(t;c) = p(c \mid R(c;t)) \cdot Q(t, S_b; R(c;t)) \qquad (2.4.10)$$

$$F(t;c) = B(t, S_b, R(c;t)) \qquad (2.4.11)$$

where $p(c|b)$ is the class transition probability of class b to class c. To allow for the transition between different classes, a special state s = 0 is introduced: and is passed on as both the score and the time index:

$$\begin{aligned} Q(t-1, s=0; c) &= H(t-1, c) \\ B(t-1, s=0; c) &= t-1 \end{aligned} \qquad (2.4.12)$$

Figure 2.4.1 illustrates the time alignment, which gives the optimal AE sequence. In this example, the optimal AE sequence is *(2; K; 1)* and transitions occur at t_1 and t_2. $Q(t,s;c)$ and $B(t,s;c)$ are determined for every state within each class at every time instance t. Then, $H(t;c)$ is computed, and $R(t;c)$ and $F(t; c)$ are recorded for all K classes. Before computing $Q(t,s;c)$, the score value and the backtrack time value for the potential AE change $Q(t$ - 1, s=0$;c)$ and $B(t$ - 1,c=0$;c)$ are set. Note that the AE sequence and the state sequence are determined simultaneously. The process starts at time t = 1 and ends at t = T in a strictly left-right fashion. When time T is reached, the optimum AE sequence Ω_l^* and the time for the AE transition T_l^* can be found by tracing back $R(c;t)$ and $F(c;t)$, respectively.

2.4.3.3 Support vector machines (SVMs)

Kernel-based algorithms have been recently developed in the Machine Learning community, where they were first introduced in the Support Vector Machine (SVM) algorithm. The attractiveness of this algorithm is due to their elegant treatment of nonlinear problems and their efficiency in high dimensional problems. Support Vector Machines (SVMs) have been shown to provide better performance than more traditional techniques in many problems, thanks to their ability to generalize. The SVM model relies on two assumptions. First, transforming data into a high-dimensional space may convert complex classification problems (with complex decision surfaces) into simpler problems that can use linear discriminant functions. Second, SVMs are based on using only those training patterns that are near the decision surface assuming they provide the most useful information for classification.

Consider the problem of separating the set of training vectors belonging to two separate classes, $(x_1; y_1)$, …, $(x_l; y_l)$, where $x_i \in R^n$ is a feature vector and $y_i \in$ {-1, +1} a class label, with a hyperplane of equation wx + b = 0. Of all the boundaries determined by w and b, the one that maximizes the margin (Figure 2.4.2) would generalize well as opposed to other possible separating hyperplanes.

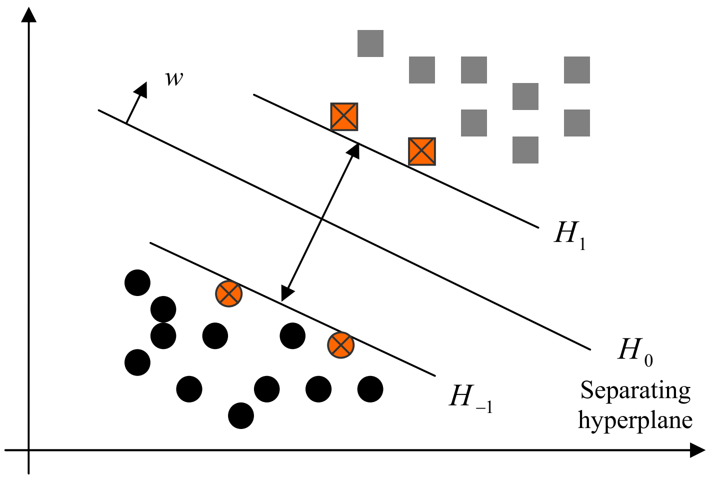


Figure 2.4.2. Two-class linear classification. The support vectors are indicated with crosses

A separating hyperplane in canonical form must satisfy the following constraints, $y_i\ [(w{\bullet}x_i) + b] \geq 1;\ i = 1,\ \ldots,\ l$. The margin is $\frac{2}{\|w\|}$ according to its definition. Hence the hyperplane that optimally separates the data is the one that minimizes $\Phi(w) = \frac{1}{2}\|w\|^2$

The solution to the optimization problem can be obtained as follows [Vap98]: first, find the maximization solution to the following problem

$$\bar{\alpha} = \arg\max_{\alpha} \sum_{i=1} \alpha_i - \frac{1}{2}\sum_{i,j=1} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{2.4.13}$$

subject to

$$\alpha_i \geq 0,\ (i = 1,\ldots,l),\ \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.4.14}$$

Then calculate

$$\bar{w} = \sum_{i=1}^{l} \bar{\alpha}_i y_i x_i\ ,\ \bar{b} = -\frac{1}{2}\bar{w} \cdot [x_+ + x_-] \tag{2.4.15}$$

where x_+ is a support vector of the "+" class and x_- is a support vector of the "-" class. Now, a new data point x is classified by the sign of

$$f(x) = sign(\sum_{i=1}^{l} \bar{\alpha}_i y_i (x_i \cdot x) + \bar{b}) \tag{2.4.16}$$

In non-separable cases, slack variables $\xi_i \geq 0$, which measure the mis-classification errors, can be introduced and a penalty function, $F(\xi) = \sum_{i=1}^{l} \xi_i$, added to the objective function. The optimization problem is now treated as to minimize the total classification error as well as the bound on the VC dimension [Vap98] of the classifier. The solution is identical to the separable case except for a modification of the Lagrange multipliers as $0 \leq \alpha_i \leq C$, $i=1, ..., l$. We refer to [Vap98] for more details on the non-separable case.

In linearly non-separable but nonlinearly (better) separable case, the SVM replaces the inner product $\mathbf{x} \cdot \mathbf{y}$ by a kernel function K(**x; y**), and then constructs an optimal separating hyperplane in the mapped space. According to the Mercer theorem [Vap98], the kernel function implicitly maps the input vectors, via Φ associated with the kernel, into a high dimensional feature space in which the mapped data is linearly separable. Possible choices of kernel functions include (1) Polynomial $K(\mathbf{x}; \mathbf{y}) = ((\mathbf{x} \cdot \mathbf{y} + 1))^d$, where the parameter d is the degree of the polynomial; (2) Gaussian Radial Basis Function $K(\mathbf{x}; \mathbf{y}) = \exp(-\frac{(x-y)^2}{2\sigma^2})$, where the parameter σ is the width of the Gaussian function; and (3) Multi-Layer Perception $K(\mathbf{x}; \mathbf{y}) = \tanh(k(\mathbf{x} \cdot \mathbf{y}) - \mu)$, where the k and μ are the scale and offset parameters. However, Exponential Radial Basis Function (ERBF) has been empirically observed to perform better than above three [Gun98].

2.4.4 Multimodal fusion approaches

As it has been already mentioned, using jointly audio and visual information can significantly improve the accuracy for AED with respect to using audio or visual information only. This is because multimodal features can resolve ambiguities that are present in a single modality. Feature sets derived from different modalities are usually governed by different laws, have different characteristic time-scales, and highlight different aspects of the AEs. Another motivation behind audio-visual AED is the bimodal characteristics of perception and production systems of human beings.

The effective combination of acoustic and visual information for AED is a challenging problem. Several approaches to audio and video fusion have been suggested in the literature. These can be classified into three main groups: data fusion, feature fusion and decision fusion. Data fusion is rarely found in multi-modal systems because raw data is usually not compatible among modalities. For instance, audio is represented by one-dimensional high rate data, whereas video is organized in two-dimensional frames over time at a much lower rate.

Concatenating features from all the modalities is likely to improve the classification accuracy. One straightforward approach is to compute audio and video features at the same time scale and put all features for each time interval into one super feature vector. Theoretically, this method can fully exploit the correlation between features from different modalities and lead to the highest classification accuracy. However, the very high dimension of the feature space makes it necessary to obtain a very large set of training data and increase the degree of freedom in the observation probability distribution (the number of Gaussian mixtures in the continuous HMM-GMM case). Integration of audio and visual features in a HMM classifier has been studied previously, for instances, for speech recognition [CZH98], and for speech-to-lip-movement synthesis. Authors incorporate dynamic visual features extracted from the speaker's lips. It is motivated by the ability of the hearing-impaired to lip-reading. Four different of feature integration methods are compared in [WHL99]: direct concatenation of feature vectors, product HMM (product of likelihood values from individual modality), two-stage HMM, and integration by neural network. The product HMM based on product of likelihood values from individual modality showed the best result on average.

An alternative to feature-level fusion is to model each different feature set separately, design a specialized classifier for this feature set, and combine the classifier output scores. Each such classifier acts as an independent "expert", giving its opinion about the unknown audio segment. The fusion rule then combines the individual experts' match scores. This approach is referred here as decision-level fusion. By combining classifiers we are aiming at a more accurate classification decision at the expense of increased complexity.

There are two main strategies in combining decisions: fusion and selection. In classifier fusion, each ensemble member is supposed to have knowledge of the whole feature space. In classifier selection, each ensemble member is supposed to know well a part of feature space and be responsible for objects in this part. Therefore in the fusion approach, usually applied combiners such as average and majority vote whereas in the selection approach usually selected only one classifier to label the input x. There are combination schemes lying between the two "pure" strategies. Such a scheme, for example, is taking the average of the outputs with coefficients that depend on the input x. Classifier selection has not attracted as much attention as classifier fusion. This might change in future [Kun04]. Cascade classifiers also seem to be relatively neglected in the literature.

Some combiners do not need training after the classifiers in ensemble have been trained individually. An example is the majority vote combiner. Other classifiers need additional training, for example, the weighted average combiner. A third class of combiners develop the combiner during the training of the individual classifiers, for example, Adaboost [Kun04].

The possible ways of combining the outputs of the L classifiers in an ensemble depend on what information we obtain from the individual members. We can distinguish two types of classifier outputs:

- Label output. Each classifier D_i produces a class label $s_i \in \Omega$, $i = 1, \ldots, L$. Thus, for any object $x \in \Re^n$ to be classified, the L classifier outputs define a vector $\mathbf{s} = [s_1, \ldots, s_L]^T \in \Omega^L$. There is no information about the certainty of the guessed labels, nor any alternative labels suggested. The methods that are based on combining label outputs include majority vote, weighted majority vote, naive Bayes combination, probabilistic tree and others.
- Continuous valued outputs. Each classifier D_i produces a c-dimensional vector $[d_{i,1}, \ldots, d_{i,c}]^T$. The value $d_{i,j}$ represents the support for the hypothesis that vector x submitted for classification comes from class ω_j. The methods that combine the continuous valued outputs include non-trainable combiners: average, minimum/maximum/median, product. Among trainable combiners are: weighted average, fuzzy integral, decision template etc.

In the following we present the examples of multimodal fusion approaches found in the literature. In paper [HLW99], authors examine different techniques to integrate audio and visual information for classification in video based on HMMs. In that work, one HMM for each class and modality is trained, and a 3-layer perceptron is used to combine the outputs. The fusion scheme in [XDX03] was proposed in the framework of detection events in sports video, where the audio features were considered as the main cue and the decisions based on motion, texture and colour were considered as auxiliary information to refine results based on audio. For laughter detection in [PP08] the decision level fusion is done by means of a sum operator. Two statistic fusion schemes: the logistic regression and the Bayesian belief network (BBN) are proposed in [LT05] for fusion different modalities. Experimental results in sports video domains suggest that the proposed framework is promising. In [CAW06] a gunshot event recognition system based on audio and visual feature analysis is presented. Gunshot events are among the most important events for an automatic surveillance system to recognize.

2.5 Chapter Summary

In this chapter we have quickly reviewed the work done so far in the area of acoustic event detection focusing on multimodal approaches. Firstly, the task of acoustic event detection has been overviewed. Also, a literature review of different multimodal audio recognition applications have been presented, where the application domain has been subdivided into multimodal speech recognition, speaker identification and emotion recognition. Secondly, the feature extraction and selection approaches as well as detection and fusion techniques that have been used in the area of audio recognition have been discussed. Finally the relevant reported works have been presented.

Chapter 3. Multimodal feature extraction

3.1 Chapter Overview

This Chapter presents features extracted from audio and video modalities which are employed in this thesis.

In Section 3.2 a set of spectro-temporal features that have shown so far its usefulness for meeting-room AED and broadcast AS tasks is presented. First, a feature grouping scheme is introduced to put all these features in a meaningful framework. Second, features coming from acoustic source localization system, which, in combination with usual spectro-temporal audio features, yield further improvements in recognition rate, are described in Section 3.3. In Section 3.4 a variety of features is extracted from video recordings by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of several acoustic events. Since the video modality is not affected by acoustic noise, the proposed features show to be useful for AED in spontaneous scenario recordings. Both the localization and the video features are new for the task of meeting-room AED.

3.2 Conventional acoustic features

In audio recognition we face the problem of the large number and variety of features proposed in the literature. Those features exploit acoustic content such as sub-band energies computed in short-time windows, time evolution parameters, modulation spectrum, level of harmonicity, etc. There are several attempts of grouping all these features into a meaningful structure can be found in literature and several of them are discussed in sub-Section 2.4.2. However, the main drawback of the most proposed grouping schemes is uncertainty about which group each particular feature belongs to. In the following we propose an alternative meaningful feature grouping scheme that can be used to put together a large variety of audio features. We argue that the features located in the same cell of this structure capture similar acoustic information of sounds.

3.2.1 Feature grouping

As we have already mentioned previously, there is a great variety of audio features proposed in the literature, but they are not well organized in some meaningful structure. In fact, some of the features exhibits a certain redundancy, like, for instance, MFCC and LPC features, that are two alternative ways of spectral envelope representation, but other are quite different like MFCC and, for instance, the fundamental frequency.

Feature extraction process consists in extraction of relevant information from audio which is used posteriorly for recognition. The question is which information should be considered as relevant, i.e. what makes sounds different for human perception? Let's review briefly the auditory model of human perception. The acoustic wave is transmitted from the outer ear to the inner ear where the ear drum and bone structures convert the sound wave to mechanical vibrations which ultimately are transferred to the basilar membrane inside the cochlea. The basilar membrane vibrates in a frequency-selective manner along its extent and thereby performs a rough (non-uniform) spectral analysis of the sound. Different frequencies presented in audio signal invoke particular neurons. The information embedded in the firing of action potentials at auditory nerves is transmitted to the higher stages of human auditory system. Thus the spectral content is one of the most important characteristic of sound from perceptual point of view. One the other hand, several studies have indicated that the exploitation of temporal information plays an important role in human sound processing. For instance, magneto-encephalographic studies in humans [LSH97] have suggested that amplitude modulations are explicitly coded in the auditory cortex; this motivates incorporating a signal decomposition in both time and frequency domain [BAK11].

Based on that observation, most of the feature extraction methods analyze audio from two perspectives: *spectral content* and the *time evolution* of spectral components. Taking into account that audio signal in general is non-stationary stochastic process, the spectral content analysis is usually performed on frame basis where the process can be considered stationary. The frame spectrum can be described from 2 perceptually relevant points of view: the shape of the *spectral envelope* and its *harmonic structure*. The time evolution of spectral components can be characterized using *statistical* and *structural* approaches. Note in some audio recognition systems time evolution is explicitly modeled during recognition stage (e.g. HMM approach).

In the proposed grouping scheme (Table 3.2.1) each feature is described in 2 dimensions: time (horizontal) and frequency (vertical). Different features can be put into specific cell of the table based on the type of information the feature is capturing. If some feature captures only spectral information without time evolution, it is put into the cell of the first column in the table.

Table 3.2.1. Proposed feature grouping scheme.

		Time domain		
		None (frame-based features)	Statistical	Structural
Frequency domain	Spectral envelope	FBE, MFCC, FF, LPC, PLP, Sp. Centr, Sp. Roll-off, Sp. width, Brightness, Spectral Slope, F_1, F_2	Mean, Variance, Kurtosis, Percentile values, Min/max of spectral envelope parameters	Delta and delta-delta, Autocorrelation features, 4 Hz modulation energy, Attack time, Attack slope of spectral envelope parameters
	Harmonicity	Fundamental frequency, Pitch, Tonality, CRRM, Harmonic to noise ratio	Mean, Variance, Kurtosis, Percentile values, Min/max of harmonicity parameters	Delta and delta-delta, Autocorrelation features, 4 Hz modulation energy, Attack time, Attack slope of harmonicity parameters

The spectral features are decomposed in 2 sub-categories: *spectral envelope* features and *harmonicity* features. *Spectral envelope* features are the most popular and they describe the shape of the spectral envelope. Spectral envelope can be parameterized in different ways. The most straightforward approach is using the integrated energies in critical sub-bands (filter bank energies, FBE). Usually the filter bank energies are further transformed since they exhibit certain correlation. The examples of such transformations are: log + DCT for MFCC feature extraction, derivative filter transformation for FF LFBE feature extraction. Sometimes the precise shape of the envelope is not of particular interest, so the distribution of spectral energy between low/high frequencies (spectral slope), sharpness/flatness (spectral width), the "middle" point of spectrum (spectral centroid, spectral roll-off point), position of formants (F_1, F_2), etc are used.

A set of *harmonicity* features is another perceptual group of features that describe the spectrum of a signal. The examples are: fundamental frequency, pitch, energy of harmonics, harmonic to noise ratio, position of harmonics along the frequency axis, inharmonicity (amount of partials that are not multiples of fundamental frequency), tonality (ratio of the harmonic to inharmonic parts of the spectrum) etc. Another example is cepstrum resynthesis residual magnitude (CRRM) that is 2-norm of the vector residual after cepstral analysis, smoothing, and resynthesis [SS97]. For instance, in the case of voiced speech, cepstral analysis filters out the pitch "ripple" from the signal, giving higher values for the residual.

The *time evolution* of frame-based spectral parameters can by analyzed using *statistical* or *structural* approaches. In statistical approach the descriptive statistics is usually used: it describes a large number of values in a sensible way. The examples of descriptive statistics are the mean, median value, variance, range, the minimum and maximum variables, histogram, quantiles and percentiles, kurtosis and skewness, percentage of low/high values etc. Statistical analysis of a set of variables is performed irrespectively of position of variables along time within analysis segment.

Although statistical characteristics of audio features are widely used for audio representation in most of current audio analysis systems and have been proved to be effective, they lead to ambiguities in some cases. In *structural time evolution* analysis the order of variables is significant thus features form a part of time series. Structural analysis accounts for the fact that data points taken over time may have an internal morphological structure (i.e. shape, autocorrelation) that should be accounted for. A structural analysis generally reflects the fact that observations close together in time will be more closely related than observations further apart. Structural feature extractors are difficult to apply to new domains because implementation of feature extraction requires domain knowledge. The simple version of *structural time evolution* features are delta and delta-delta coefficients of the standard ASR features. Other features from that category include attack time (temporal duration of the attack phase), attack slope (ratio between the magnitude difference at the beginning and the ending of attack period, and the corresponding time difference), 4 Hz modulation energy (speech tends to have more modulation energy at 4Hz than music does), amplitude modulation features, etc.

3.2.1 Feature extraction for AED and AS

3.2.1.1 Frequency filtered log filter bank energies

Sixteen FF LFBE coefficients [NMH01] along with their first temporal derivatives are extracted to describe every audio signal frame. Therefore, a 32-dimensional feature vector is used. The frame length is 30 ms with 20 ms shift, and a Hamming window is applied.

3.2.1.2 Set of perceptual features

Motivated by the discriminative ability of perceptual features reported in state-of-the-art literature, in certain experiments a set of perceptual features is used in combination with previously described features. The number of coefficients is presented in brackets:

- Zero crossing rate (1)
- Short-time energy (1)
- Sub-band energies (4)
- Spectral flux (4)
- Spectral centroid (1)
- Spectral bandwidth (1)
- Chroma parameters (12)

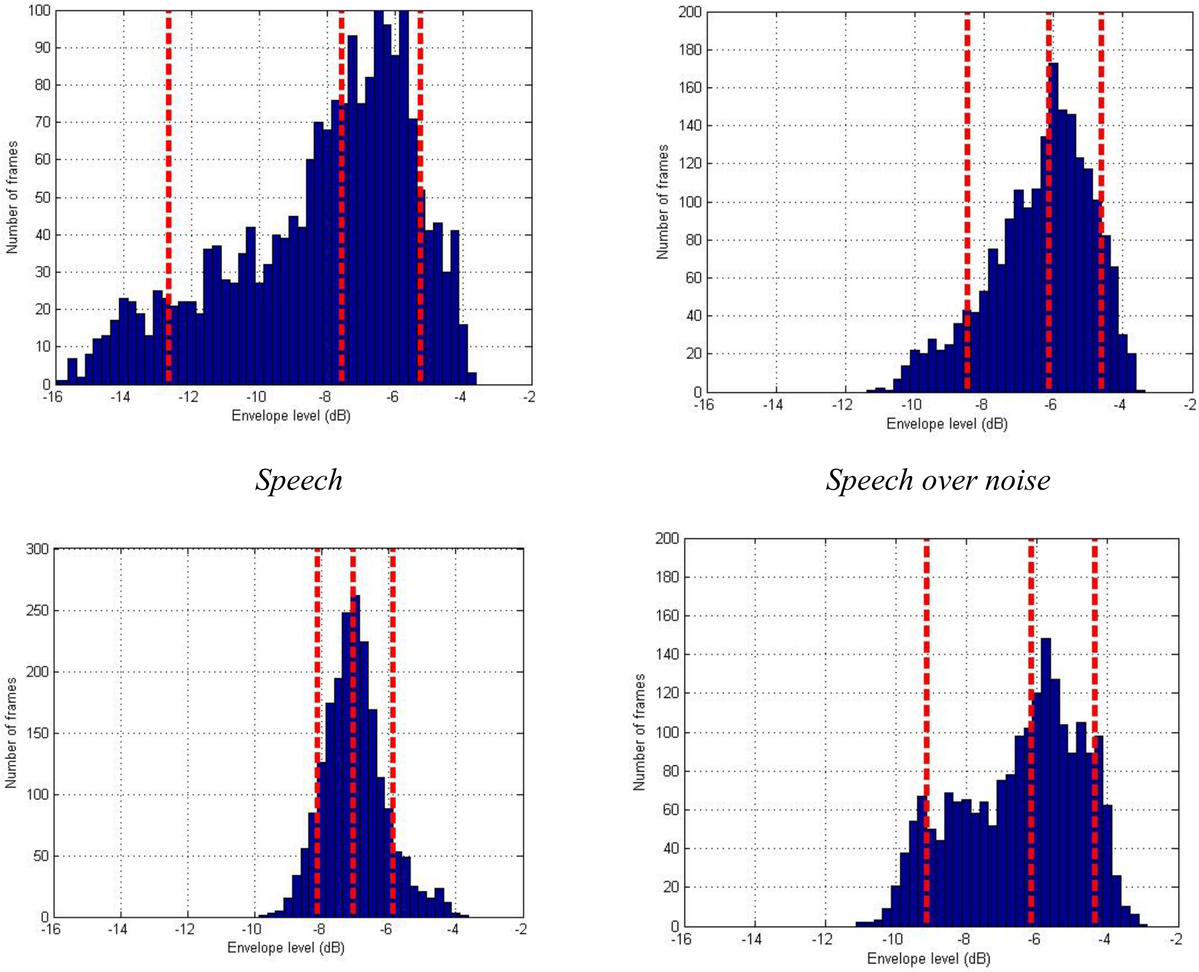
3.2.1.3 Amplitude modulation features

In order to improve the AS results in the broadcast news domain the following 5 amplitude modulation features are extracted (described in chapter 2):

- Width
- Symmetry
- Skewness
- Kurtosis
- Lower half

All these features are computed over a 10 sec window with 1 sec shift.

The amplitude histogram of continuous signals, like noisy speech noise and certain kinds of music possibly overlapped with noise, shows a narrow and symmetrical distribution, whereas the distribution is broad and asymmetric for speech.



Speech

Speech over noise

Music

Speech over music

Figure 3.2.1. Amplitude envelope histogram of "Clean speech", "Speech with noise in background", "Music", "Speech with music in background"

The examples in Figure 3.2.1 show the amplitude histogram of speech, speech over noise, music and speech over music classes employed in AS task. The histograms were built over twenty seconds of the energy envelope of each signal. Due to the pauses in the speech signal, its level varies very much over time, resulting in a broad and asymmetrical amplitude histogram. The level of speech over noise varies less, that is, the amplitude histogram has a narrow and symmetrical form. Music histogram has even narrower amplitude histogram and also symmetric form. In addition to the histograms, 10%, 50% and 90% percentiles are also drawn in the figures. The 10% percentile, for example, shows the level below which the envelope is 10% of the time. The asymmetrical distribution in the speech signal results in a much larger distance between the 10% and the 50% percentile than between the 50% and 90% percentile, or, in other words, the 50% percentile is far away from the arithmetical mean of the 10% and 90% percentile. For music signals, the 50% percentile is more or less in the middle of the 10% and the 90 % percentile, representing the symmetrical distribution.

3.2.1.4 Spectral slope features

In AS task we have "telephone speech over music" as one of the classes of interest which composed of the music that spans all the frequency range up to 8 kHz (16 kHz sampling frequency), and telephone speech which is in low frequency range (up to 3.4 kHz). New features, called spectral slope features, are proposed to enhance the detection accuracy. To compute a spectral slope, two different couples of sub-bands are defined. These sub-bands have been chosen to discriminate between "telephone speech over music" and the rest of audio based on the slope of the spectrum in the region around 4000 Hz, the end of the band of telephone speech, beyond which only music frequency components exist. The first couple is made of the sub-bands [1000 – 3000] Hz and [3000 – 7000] Hz and the second is consists of the sub-bands [3000 – 3500] Hz and [3500 – 4000] Hz (see Figure 3.2.2). These sub-bands aim to parameterize the energy in the region where the energy drop should appear for the "telephone speech over music" class.

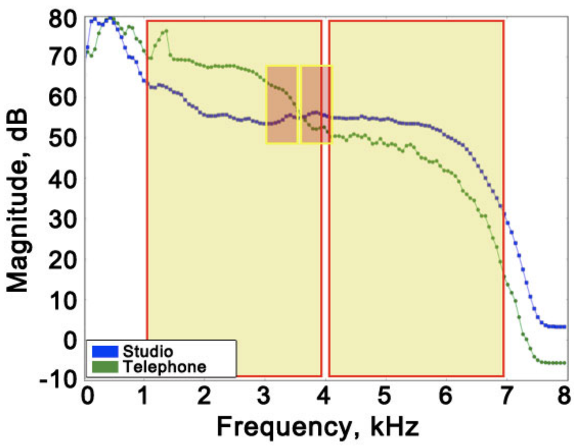


Figure 3.2.2. Sub-band couples for the spectral slope superposed over periodograms corresponding to "speech" and "telephone speech over music" classes

A spectral slope feature vector *ss* is computed for each couple as:

$$ss = (S_1, S_2, \frac{S_1}{S_2}) \qquad (3.2.1)$$

where S_1, S_2 are total energies of the first and second sub-band respectively.

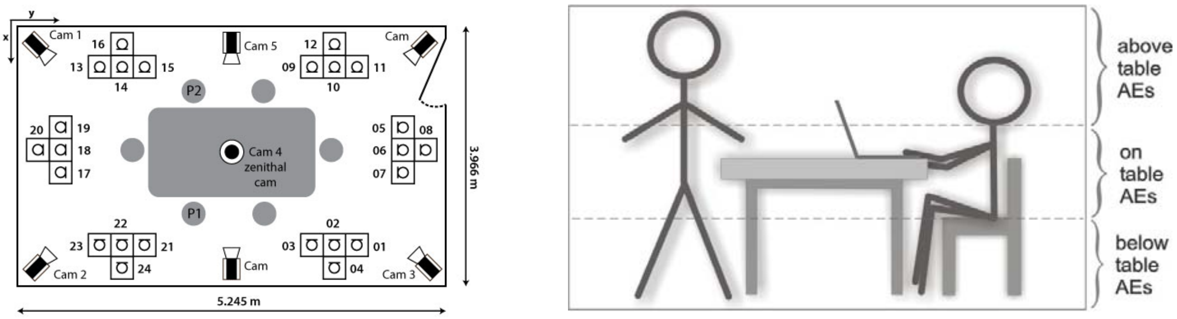
Experimental results have shown that the dynamics of the spectral slope features are helpful for the detection of the "telephone speech over music" class. Thus the deltas and accelerations are added to the final feature vector. Finally a set of 18 values is obtained for each frame.

3.3 Acoustic localization features

In order to enhance the AED accuracy in meeting-room environments, acoustic localization features are used in combination with audio spectro-temporal features. In our case, as the characteristics of the room are known beforehand (Figure. 3.3.1 (a)), the position (*x, y, z*) of the acoustic source may carry useful information. Indeed, some AEs can only occur at particular locations, like door slam and door knock can only appear near the door, or footsteps and chair moving events take place near the floor. Based on this fact, we define a set of meta-classes that depend on the position where each AE can be detected. The proposed meta-classes are: “near door” and “far door”, related to the distance of the acoustic source to the door, and “below table”, “on table” and “above table” meta-classes depending on the *z*-coordinate of the detected AE. The meta-class categorization of meeting-room AEs is presented in Table 3.3.1. The height-related meta-classes are depicted in Figure 3.3.1 (b) and their likelihood function modelled via Gaussian Mixture Models can be observed in Figure 3.3.2 (b). Thus in our experiments two acoustic source localization features are extracted: *d* (distance from the door) and *z* (height of the detected acoustic source). It is worth noting that the *z*-coordinate is not a discriminative feature for those AEs that are produced at the similar height.

Table 3.3.1 Meta-class categorization of the meeting-room acoustic events

Based on distance from the door		**Based on z-coordinate**		
near door	*Far door*	*below-table*	*on-table*	*above-table*
Door knock	10 AEs not included into “near door”	Steps	Paper wrapping	Speech
Door slam		Chair moving	Keyboard typing	Cough
			Key jingle	Laugh
			Phone ring	Applause
			Cup clink	



(a) (b)

Figure 3.3.1. (a) The top view of the room. (b) The three categories along the vertical axis

3.3.1 Feature extraction approach

The acoustic localization system used in this work is based on the SRP-PHAT [DSB01] localization method, which is known to perform robustly in most scenarios. The SRP-PHAT algorithm is briefly described in the following. Consider a scenario provided with a set of N_M microphones from which we choose a set of microphone pairs, denoted as Ψ. Let X_i and X_j be the 3D location of two microphones *i* and *j*. The time delay of a hypothetical acoustic source placed at $x \in R^3$ is expressed as:

$$\tau_{x,i,j} = \frac{\|x - x_i\| - \|x - x_j\|}{s} \tag{3.3.1}$$

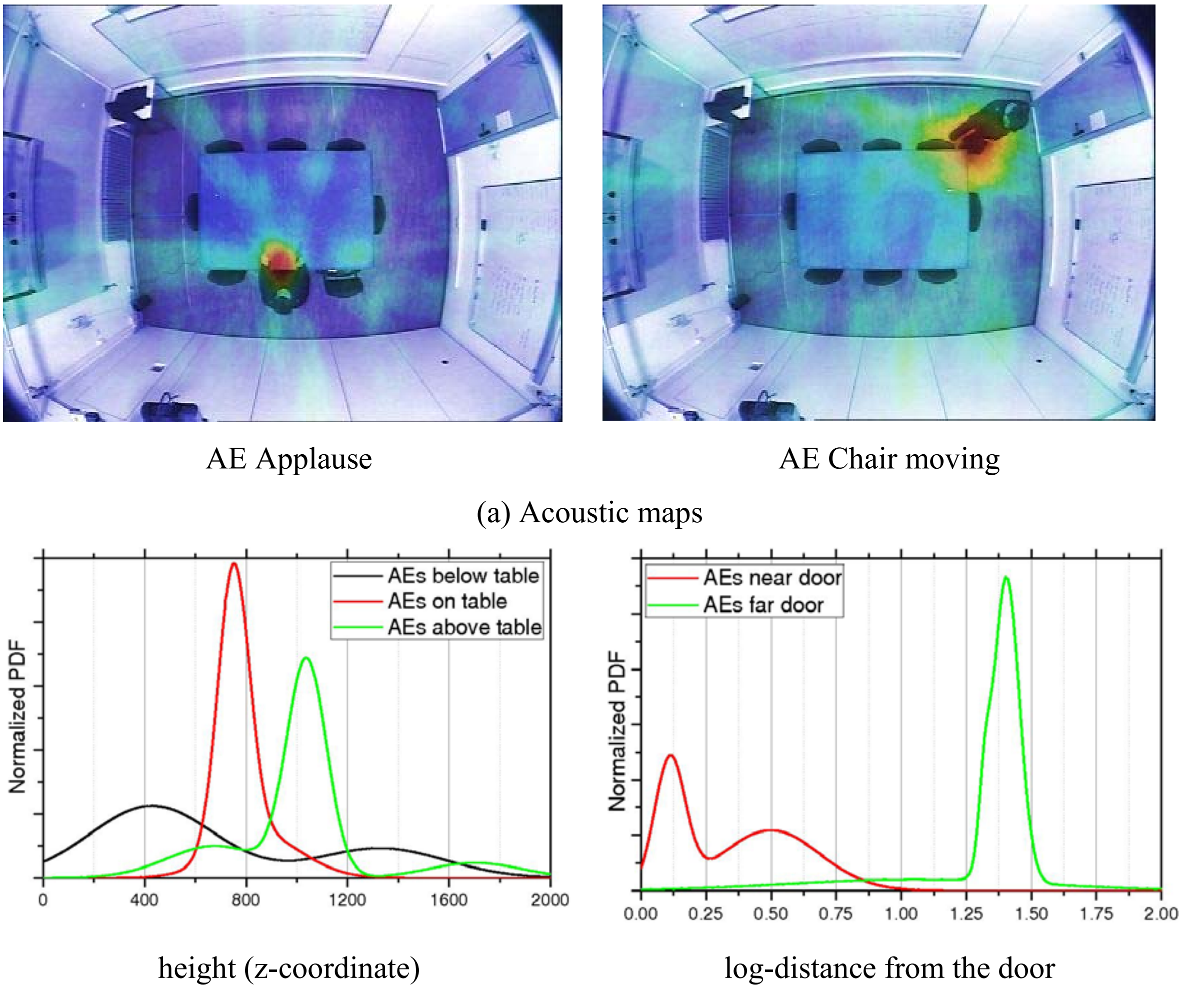
where *s* is the speed of sound. The 3D space to be analyzed is quantized into a set of positions with typical separations of 5 to 10 cm. The theoretical TDoA $\tau_{x,i,j}$ from each exploration position to each microphone pair is pre-calculated and stored. PHAT-weighted cross-correlations of each microphone pair are estimated for each analysis frame [OS97]. They can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density $G_{i,j}(f)$ as follows:

$$R_{i,j}(\tau) = \int_{-\infty}^{\infty} \frac{G_{i,j}(f)}{|G_{i,j}(f)|} e^{j2\pi f\tau} df \tag{3.3.2}$$

The contribution of the cross-correlation of every microphone pair is accumulated for each exploration region using the delays pre-computed in Eq. 3.3.2. In this way, we obtain an acoustic map at every time instant, as depicted in Figure 3.3.2 (a). Finally, the estimated location of the acoustic source is the position of the quantized space that maximizes the contribution of the cross-correlation of all microphone pairs:

$$\hat{x} = \underset{x}{argmax} \sum_{i,j \in \Psi} R_{i,j}(\tau_{x,i,j}) \tag{3.3.3}$$

The sum of the contributions of each microphone pair cross-correlation gives a value of confidence of the estimated position, which is assumed to be well-correlated with the likelihood of the estimation.



AE Applause

AE Chair moving

(a) Acoustic maps

height (z-coordinate)

log-distance from the door

(b) AE localization distributions

Figure 3.3.2. Acoustic localization. In (a), acoustic maps corresponding to two AEs overlaid to a zenithal camera view of the analyzed scenario. In (b), the likelihood functions modelled by GMMs

3.4 Video features

Acoustic event detection is usually addressed from an audio perspective only. Typically, low acoustic energy AEs as paper wrapping, keyboard typing or footsteps are hard to be detected using only the audio modality. The problem becomes even more challenging in the case of signal overlaps. Since the human-produced AEs have a visual correlate, it can be exploited to enhance the detection rate of certain AEs. Therefore, a number of features are extracted from video recordings by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of five classes of AEs. From the audio perspective, the video modality has an attractive property: the disturbing acoustic noise usually does not have a correlate in the video signal. In this section, several video technologies which provide useful features for our AED task are presented.

3.4.1 Person tracking features

Multiple cameras are employed to perform tracking of multiple interacting people in the scene, applying the real-time performance algorithm presented in [CSC08]. This technique exploits spatial redundancy among camera views towards avoiding occlusion and perspective issues by means of a 3D reconstruction of the scene. Afterwards, an efficient Monte Carlo based tracking strategy that exploits particle filtering (PF) [AMG02] approach retrieves an accurate estimation of the location of each target at every time instant *t*. Two main factors are to be taken into account when implementing a particle filter: the likelihood function and the propagation strategy.

Likelihood function $p(z_t|x_t)$ can be defined as the likelihood of a particle belonging to the volume that corresponds to a person. For a given particle *j* occupying a voxel x_t, its likelihood is formulated as:

$$p(z_t \mid x_t^j) = \frac{1}{\left|C(x_t^j, q)\right|} \sum_{p \in C(x_t^j, q)} d(x_t^j, p) \qquad (3.4.1)$$

where $C(\cdot)$ stands for the neighborhood over a connectivity q domain on the 3D orthogonal grid and $|C(\cdot)|$ represents its cardinality. Typically, connectivity in 3D discrete grids can be 6, 14 and 26; in our research q=26 provided accurate results. Function $d(\cdot)$ measures the distance between a foreground voxel p in the neighbourhood of the particle.

Challenges in 3D multi-person tracking from volumetric scene reconstruction are basically twofold. First, finding an interaction model in order to avoid mismatches and target merging. Several approaches have been proposed [Lan06] but the joint PF presented in [KBD03] is the optimal solution to multi-target tracking using PFs. However, its computational load increases dramatically with the number of targets to track since every particle estimates the location of all targets in the scene simultaneously. The proposed solution is to use a split PF per person, which requires less computational load at the cost of not being able to solve some complex cross-overs. However, this situation is alleviated by the fact that cross-overs are restricted to the horizontal plane in our scenario (see Figure 3.4.1).

Let us assume that there are *M* independent PF trackers, being *M* the number of humans in the room. Nevertheless, they are not fully independent since each PF can consider voxels from other tracked targets in either the likelihood evaluation or the 3D re-sampling step resulting in target merging or identity mismatches. In order to achieve the most independent set of trackers, we consider a blocking method to model interactions. Many blocking proposals can be found in 2D tracking related works [KBD03] and we extend it to our 3D case.

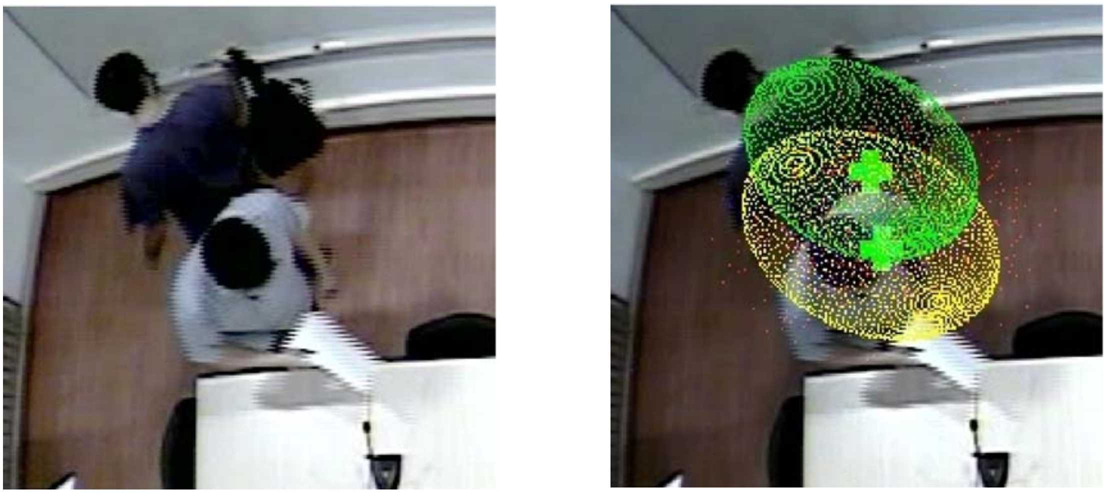


Figure 3.4.1. Particles from the tracker A (yellow ellipsoid) falling into the exclusion zone of tracker B (green ellipsoid) will be penalized

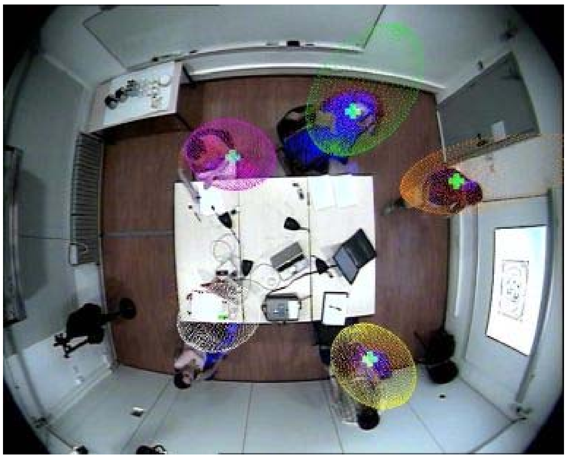
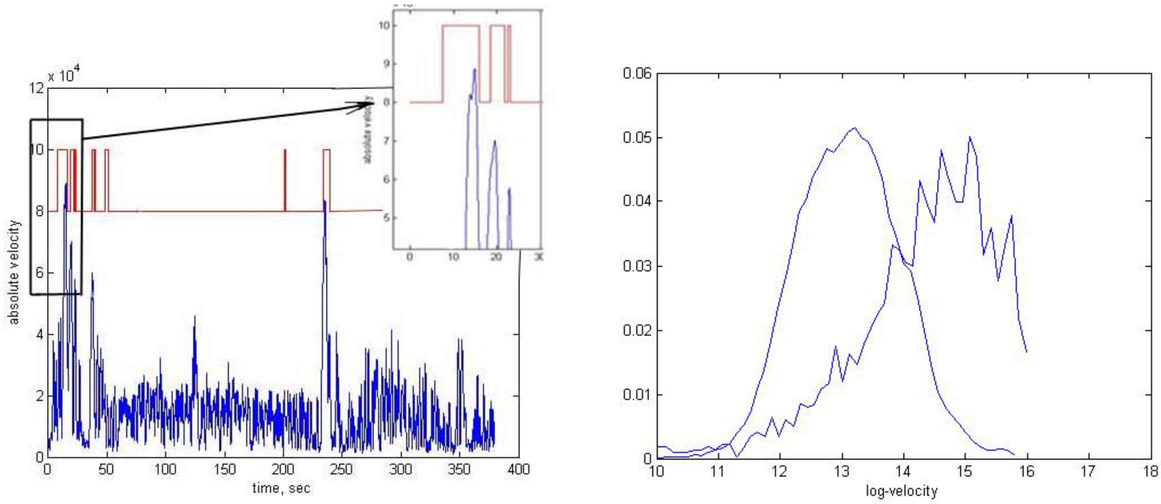


Figure 3.4.2: The output of the employed algorithm in a scenario involving multiple targets

The combination of the estimated 3D location together with geometric descriptors allows discarding spurious objects such as furniture and a simple classification of the person's pose as standing or sitting. The performance of this algorithm over a large annotated database [LCC07] showed the effectiveness of this approach. An example of the performance of this algorithm is shown in Figure 3.4.2.



(a) (b)

Figure 3.4.3. In (a), values of the velocity during one development seminar (bottom) and reference "steps" labels (top). In (b), the histograms of log-velocities for "non-steps" (left hump) and "steps" (right hump)

The output of the 3D tracking algorithm is the set of coordinates of all the people in the room, which are given every 40ms (25 fps). From those coordinates, we have to generate features that carry information correlated with “Steps”. The movements of people in the meeting room can be characterized by a velocity measure. In a 2D plane, the velocity can be calculated in the following way:

$$v = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} \qquad (3.4.2)$$

where *dx/dt* and *dy/dt* are the values of velocity along *x* and *y* axes, respectively. Those values are calculated using a smoothed derivative non-casual filter *h* applied to the vector of positions of each person in the room. We tried several shapes of the impulse response of the derivative filter; best results were obtained using a linear non-casual filter with the impulse response *h(n)* = [-*m* … -2 -1 0 1 2 … *m*] (zero corresponds to the current value and *L=2*m+1* is the length of the filter).

Usually more than one person is present in the room, and each person has its own movement and velocity. The maximum velocity among the participants in the seminar is used as a current feature value for “Steps”/ “non-Steps” detection.

Figure 3.4.3 (a) plots the maximum value of velocity among participants for a 6-min seminar along with the corresponding ground truth labels. From it we can observe that there is certain degree of correspondence between peaks of velocity and true “Steps”.

The normalized histograms of the logarithm of velocity for “Steps” and “non-Steps” obtained from development seminars are depicted in Figure 3.4.3 (b), from which can be seen that “Steps” are more likely to appear with higher values of velocity. The jerky nature of the “Steps” hump results from a more than 10 times scarcer representation of “Steps” with respect to “non-Steps” in the development database.

To have a better detection of “Steps” the length *L* of the derivative filter *h(n)* and several types of windows applied on *h*(*n*) were investigated. According to the results shown in Figure 3.4.4, the best detection of “Steps” on development data is achieved with a 2-sec-long derivative filter and a Hamming window.

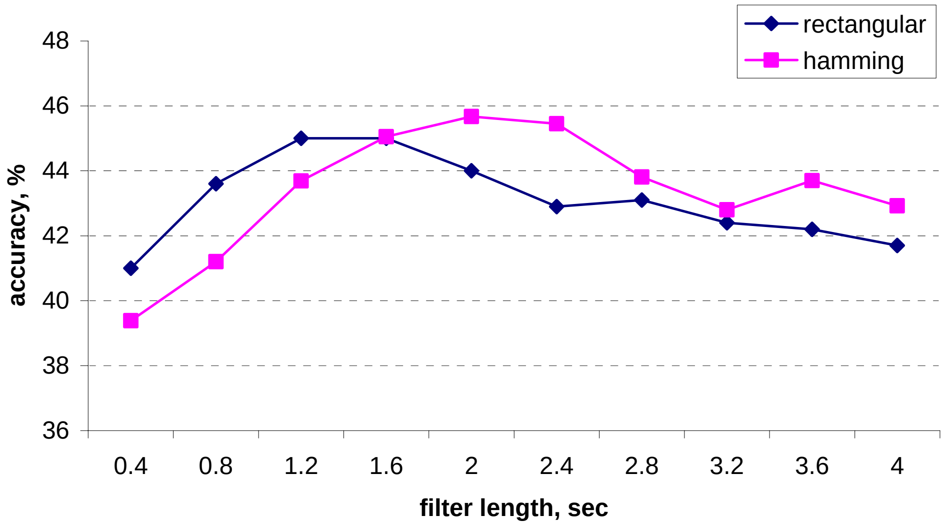


Figure 3.4.4. Detection of "Steps" on the development database as a function of the length of the derivative filter (in seconds)

3.4.1.1 Feature for detection "Chair moving" AE

Once the position of the target is known, an additional feature associated with the person can be extracted: height. When analyzing the temporal evolution of this feature, sudden changes of it are usually correlated with chair moving AE, that is, when the person sits down or stands up. In Figure 3.4.5 the temporal evolution of the height position of the participant (blue solid line) is depicted along with ground truth labels of "chair moving" AE (red dashed line). One can observe a certain correlation between these two curves. The video features for "chair moving" AE detection are obtained using a smoothed derivative non-casual filter *h* applied to the vector of height positions of each person in the room. There are two possible cases when "chair moving" AE may appear:

1) The person is currently sitting and wants to stand up. In this case the person usually first stands up and then moves a chair towards the table. This corresponds to the case when "chair moving" sound appears with some delay (in average, about 1.5 seconds).

2) The person is currently standing and wants to sit down. In this case the person first moves a chair and then takes a seat (in average, about 1.5 seconds delay).

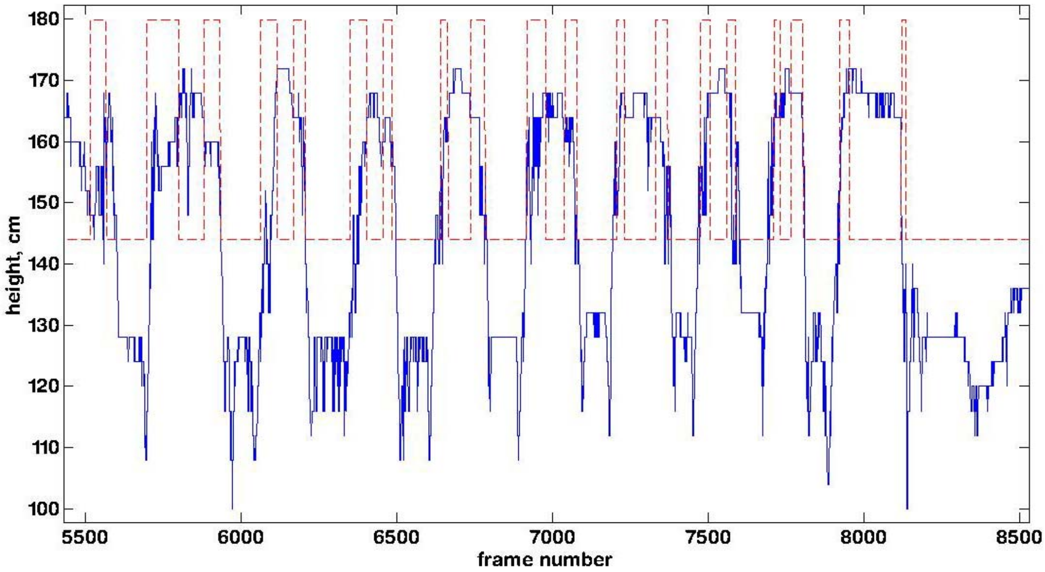


Figure 3.4.5. Values of the height position of the participant during the development seminar (solid blue curve) and reference "chair moving" labels (dashed red curve)

These two types of delays are incorporated in feature extraction process as described in following. If $f(t)$ is the height position of the participant along the time, the derivative of height position of the person is computed as $g\ (t) = f(t)*h$. Note, $g(t)$>0 if the persons stands up, $g(t)<0$ if the person sits down and $g(x) \approx 0$ the rest of the time. Two new auxiliary functions are introduced:

$$g_1(t) = \begin{cases} g(t), \text{ if } g(t) > 0 \\ 0, \quad \text{otherwise} \end{cases} \qquad (3.4.3)$$

$$g_2(t) = \begin{cases} -g(t), \text{ if } g(t) < 0 \\ 0, \quad \text{otherwise} \end{cases} \qquad (3.4.4)$$

The feature vector for "chair moving" AE detection is obtained as $G(t) = g_1(t + \tau) + g_2(t - \tau)$, where $\tau = 1.5$ sec in our experiments.

3.4.2 Colour-specific MHI

Some AEs are associated with motion of objects around the person. In particular, we would like to detect a motion of a white object in the scene that can be associated to paper wrapping (under the assumption that a paper sheet is distinguishable from the background colour). In order to address the detection of white paper motion, a close-up camera focused on the front the person under study is employed. Motion descriptors introduced by [BD99], namely the motion history energy (MHE) and image (MHI), have been found useful to describe and recognize actions.

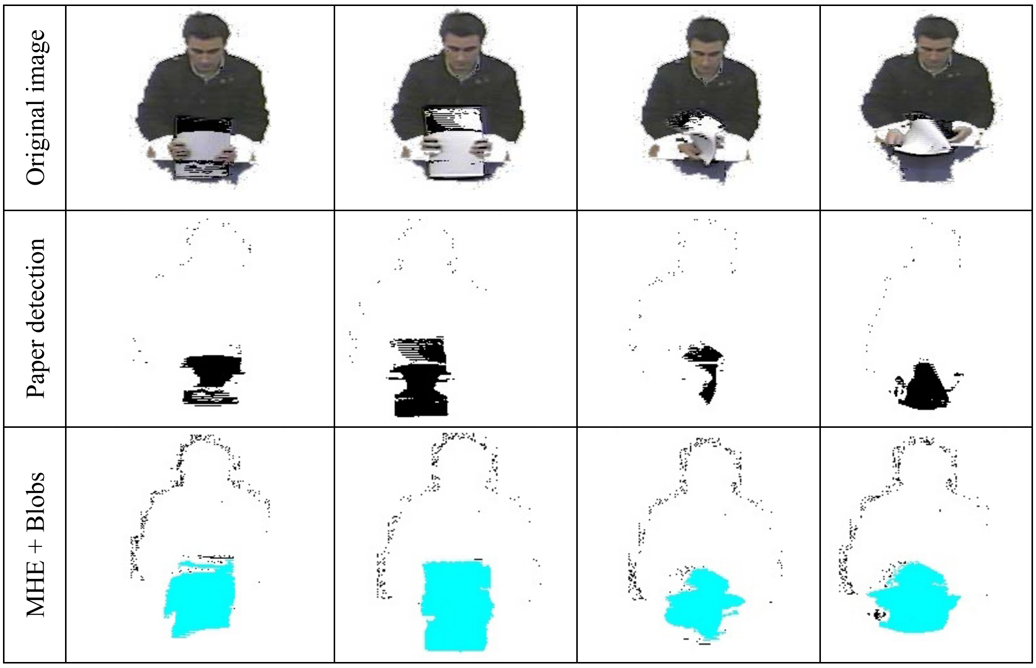


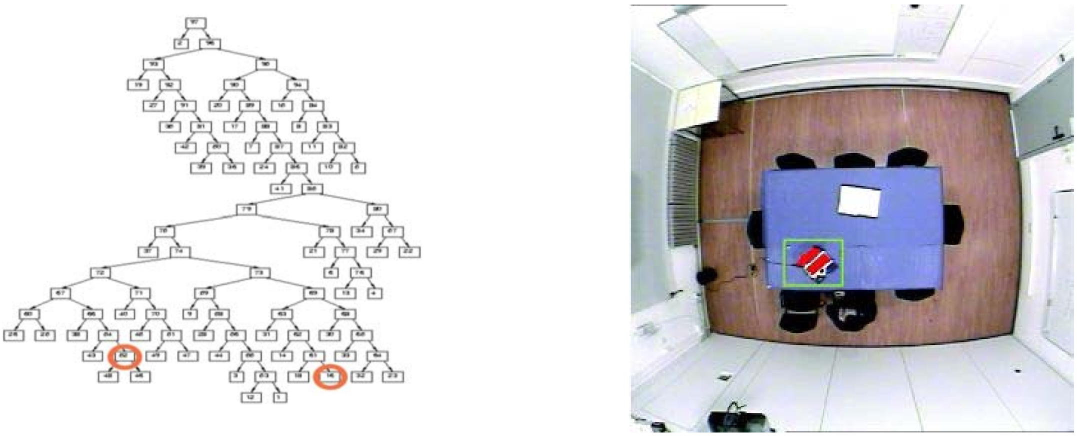
Figure 3.4.6. Paper wrapping feature extraction

However, in our work, only the MHE feature is exploited, since the MHI descriptor encodes the structure of the motion, i.e. how the action is executed; this cue does not provide any useful information to increase the classifier performance. Every pixel in the MHE image contains a binary value denoting whether motion has occurred in the last τ frames at that location. In the original technique, silhouettes were employed as the input to generate these descriptors but they are not appropriate in our context since motion typically occurs within the silhouette of the person. Instead, we propose to generate the MHE from the output of a pixel-wise colour detector, hence performing a color/region-specific motion analysis that allows distinguishing motion for objects of a specific color. For paper motion, a statistic classifier based on a Gaussian model in RGB is used to select the pixels with whitish colour. In our experiments, $\tau = 12$ frames produced satisfactory results. Finally, a connected component analysis is applied to the MHE images and some features are computed over the retrieved components (blobs). In particular, the area of each blob allows discarding spurious motion. In the paper motion case, the size of the biggest blob in the scene is employed to address paper wrapping AE detection. An example of this technique is depicted in Figure 3.4.6.

3.4.3 Object detection

Detection of certain objects in the scene can be beneficial to detect some AEs such as phone ringing, cup clinking or keyboard typing. Unfortunately, phones and cups are too small to be efficiently detected in our scenario, but the case of a laptop can be correctly addressed. In our case, the detection of laptops is performed from a zenithal camera located at the ceiling. The algorithm initially detects the laptop's screen and keyboard separately and, in a second stage, assesses their relative position and size. Captured images are segmented to create an initial partition of 256 regions based on colour similarity. These regions are iteratively fused to generate a Binary Partition Tree (BPT), a region-based representation of the image that provides segmentation at multiple scales [SG00]. Starting from the initial partition, the BPT is built by iteratively merging the two most similar and neighbouring regions, defining a tree structure whose leaves represent the regions at the initial partition and the root corresponds to the whole image (see Figure 3.4.7 (a)). Thanks to this technique, the laptop parts may be detected not only at the regions in the initial partition but also at some combinations of them, represented by the BPT nodes. Once the BPT is built, visual descriptors are computed for each region represented at its nodes. These descriptors represent colour, area and location features of each segment.

The detection problem is posed as a traditional pattern recognition case, where a GMM-based classifier is trained for the screen and keyboard parts. A subset of ten images representing the laptop at different positions in the table has been used to train a model based on the region-based descriptors of each laptop part, as well as their relative position and sizes. An example of the performance of this algorithm is shown in Figure 3.4.7 (b). For further details on the algorithm, the reader is referred to [GM07].



(a) (b)

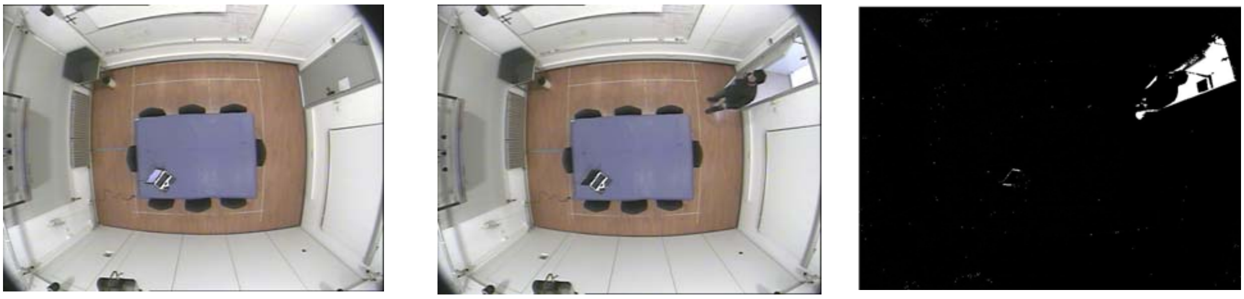
Figure 3.4.7. Object detection. In (a) the binary tree representing the whole image as a hierarchy. Regions corresponding to the screen and keyboard regions are identified within the tree. In (b), the detection of a laptop from zenithal view

3.4.4 Door activity features

In order to visually detect door slam AE, we considered exploiting the a priori knowledge about the physical location of the door. Analyzing the zenithal camera view, activity near the door can be addressed by means of a foreground/background pixel classification [SG99]. The amount of foreground pixels in the door area will indicate that a person has entered or exited, hence allowing a visual detection of door slam AE. Although changes in scene lighting can cause problems for many backgrounding methods in outdoor conditions, the lighting in the meeting room scenario is usually remains constant.

In Figure 3.4.8 the method for foreground pixel extraction is described. First, the background image of the room from zenithal camera view is stored (Figure 3.4.8 (a)). Second, for each video frame (Figure 3.4.8 (b)) the foreground is detected by subtracting each pixel of background and current frame of the sequence and making thresholding Figure 3.4.8 (c).

The proportion of foreground pixels in the "door area" along the time is depicted in Figure 3.4.9. Although this algorithm is robust enough, the false alarms may appear when some person is moving in the "door area".



(a) (b) (c)

Figure 3.4.8. Door activity feature extraction. (a) Background image; (b) Image corresponding to the person enters the room; (c) Foreground pixels depicted in white

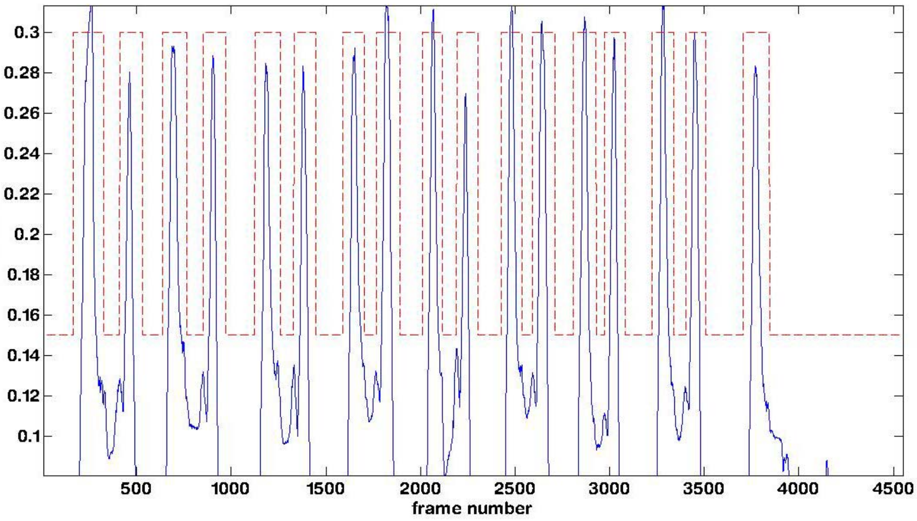


Figure 3.4.9: The proportion of foreground pixels in "door area" along the time (blue solid line) with the reference "door slam" labels (red dashed line)

3.5 Chapter Summary

In this Chapter we have presented a meaningful framework for grouping of audio features, and described those features that so far have shown its usefulness for the acoustic event detection task. A new feature set called spectral slope was designed to detect a particular class of interest in broadcast news audio segmentation. Additional features that describe the spatial location of the produced AE in the 3D space were presented which are new for the meeting-room AED task. Moreover, a number of features were extracted from the video signals by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of several AEs which are also new for the meeting-room AED.

Chapter 4. Multimodal meeting-room acoustic event detection

4.1 Chapter Overview

In the previous chapter a variety of multimodal features extracted from audio and video modalities was presented. The second step in building a multimodal AED system is choosing a multi-class detection approach and an appropriate way of fusion of different modalities.

The databases and metric to evaluate the accuracy are presented in Section 4.2. Two basic detection strategies are described in chapter 4.3: HMM-GMM-based and SVM-based. SVM is a discriminative two-class classification technique that can be easily extended to multi-classification problem. Feature-level and decision-level fusion strategies are described in this Section 4.4. Feature-level fusion strategy relies on synchronization of different modalities and concatenation of features into one super-vector. In decision-level fusion approach an individual detectors are built for audio and video modalities and the output scores are combined using two statistical approaches: weighted arithmetical mean (WAM) and fuzzy integral (FI). Fuzzy integral is a meaningful formalism for combining classifier outputs that can capture interactions among various sources of information. Experimental results are presented in Section 4.6 and Section 4.7 concludes this chapter.

4.2 Databases and metric

4.2.1 Description of the databases

State-of-the-art empirical and statistical data driven methods in audio recognition depend to a large extend on sufficient and appropriate sample data, often covering a particular domain, acoustic environment, recording channel or modality. One of the problems when dealing with multimodal AED task in the meeting-room environment is lack of the annotated data to evaluate the performance of the proposed techniques. There exists a relatively large database of sounds, like RWCP sound scene database [NHA02], but only a small part of the sounds included in that database can be considered as usual or at least possible in a meeting room and only the audio modality is available for those sounds. Another relatively large and multimodal AMI corpus [AMI] contains only a limited number of AE instances that is not appropriate to develop AED technologies.

For meeting-room environments, the task of AED is relatively new; however, it has already been adopted as a semantically relevant technology in CHIL European project (2004-2007) and two international evaluation campaigns: in CLEAR (Classification of Events, Activities, and Relationships evaluation campaigns) 2006 [CLE06], by three participants, and in CLEAR 2007 [CLE07], by six participants. To support these evaluations a large multimodal and multi-site corpus for AED in meeting-room environment has been created.

In this thesis a CLEAR’07 evaluation corpus is used. It consists of 25 interactive seminars, approximately 30 min-long that have been recorded by AIT (Athens Information Technology), ITC (Instituto Trentino di Cultura), IBM, UKA (Universität Karlsruhe), and UPC (Universitat Politècnica de Catalunya) in their smart-rooms. Each seminar usually consists of a presentation of 10 to 30 minutes to a group of three to five attendees in a meeting room. During and after the presentation, there are questions from the attendees with answers from the presenter. There is also activity in terms of people entering/leaving the room, opening and closing the door, standing up and going to the screen, discussion among the attendees, coffee breaks, etc. Each meeting can be conditionally decomposed into acoustic scenes: “beginning”, “meeting”, “coffee break”, “question/answers”, and “end”. The recorded interactive seminars contained a satisfactory number of acoustic events, so it was possible to perform AED tests that are statistically meaningful. The development part of the database consists of five interactive seminars (one from each site). In total, development data consists of 7495 seconds, where 16% of total time is AEs, 13% is silence, and 81% is “Speech” and “Unknown” classes. The remaining interactive seminars have been conditionally decomposed into 5 types of acoustic scenes: “beginning”, “meeting”, “coffee break”, “question/answers”, and “end”. After observing the “richness” of each acoustic scene type in terms of AEs, 20 5-minute segments have been extracted maximizing the AE time and number of occurrences per AE class. These data belong to the test part of the database and consist of 6001 seconds, where 36% are AE time, 11% are silence, and 78% are “Speech” and “Unknown” classes. Noticeably, during about 64% of time, the AEs are overlapped with “Speech” and during 3% they are overlapped with other AEs. In terms of AE occurrences, more than 65% of the existing 1434 AEs are partially or completely overlapped with “Speech” and/or other AEs.

Since the employed cameras in CLEAR’07 evaluation corpus do not provide a close view of the subjects under study, a new database has been recorded at UPC smart-room with 5 calibrated cameras and 6 T-shaped 4-microphone clusters (Appendix A). This database includes two kinds of datasets: 8 recorded sessions of isolated AEs, where 6 different participants performed 10 times each AE, and a spontaneously generated dataset which consists of 9 scenes about 5 minutes long with 2 participants that interact with each other in a natural way: discuss certain subject, drink coffee, speak on the mobile phone, etc. Although the interactive scenes were recorded according to a previously elaborated scenario, we call this type of recordings “spontaneous” since the AEs were produced in a realistic seminar style with possible overlap with speech. Manual annotation of the data has been done to get an objective performance evaluation. This database is publicly available from the author and the detailed description of this database is presented in Appendix A.

The above mentioned databases include 15 semantic classes (classes of interest), i.e. types of AEs 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”, “silence”, “unknown”. Among them, there are 2 AEs, “silence” and “unknown”, which are not evaluated. Along with the audio-visual data, the audio database of isolated AEs recorded at UPC in 2004 was used. The details of the databases in terms of the number of occurrences per AE class are shown in Table 4.2.1.

Using “UPC iso multimodal” database a new corpus with signal overlaps has been artificially generated. We assume a meeting scenario where there are two simultaneous acoustic sources in the room: one is always speech and the other is a specific AE. Taking into account this assumption, the UPC smart-room has been considered ideally subdivided in the two areas: left and right. In the left part the speaker produces speech, and in the right part the listener produces different types of AEs. Following this assumption, speech of the speaker was recorded from the left part of the room and posteriorly it was artificially overlapped with the multimodal database of isolated AEs. To do that, for each AE instance, a segment with the same length was extracted from a random position inside the speech signal. The overlapping was performed with 5 different Signal-to-Noise Ratios (SNRs): 20 dB, 10dB, 0dB, -10dB. -20 dB, where speech is considered as "noise". Although the database with overlapped AEs is generated in an artificial way, it has some advantages:

a) The behaviour of the system can be analyzed for different levels of overlap.

b) The existing databases of isolated AEs with high number of instances can be used for evaluation.

Table 4.2.1. Number of occurrences per acoustic event class

Event Type	Label	Number of Occurrences			
		Audio	Audio-visual		
		UPC iso audio	CHIL seminars	UPC iso multimodal	UPC sponta-neously generated
Door knock	[kn]	50	235	79	27
Door open/slam	[ds]	120	149	256	82
Steps	[st]	73	570	206	153
Chair moving	[cm]	76	464	245	183
Spoon/cup jingle	[cl]	64	56	96	48
Paper work	[pw]	84	218	91	146
Key jingle	[kj]	65	54	82	41
Keyboard typing	[kt]	66	177	89	81
Phone ring	[pr]	116	46	101	29
Applause	[ap]	60	21	83	9
Cough	[co]	65	90	90	24
Laugh	[la]	64	191	-	-
Speech	[sp]	-	2463	74	255
Unknown	[un]	126	860	-	-
Silence	[si]	Not annotated explicitly			

4.2.2 Metric

In support of CHIL evaluation campaign a specific metric for AED technology evaluation has been defined. The metric referred to AED-ACC (4.2.1) is employed to assess the accuracy AED systems. This metric is defined as the F-score (the harmonic mean between precision and recall):

$$AED - ACC = \frac{2 * Precision * Recall}{Precision + Recall}, \qquad (4.2.1)$$

where

$$Precision = \frac{number\ of\ correct\ system\ output\ AEs}{number\ of\ all\ system\ output\ AEs},$$

$$Recall = \frac{number\ of\ correctly\ detected\ reference\ AEs}{number\ of\ all\ reference\ AEs}.$$

A system output AE is considered correct if at least one of two conditions is met: 1) There exists at least one reference AE whose temporal centre is situated between the timestamps of the system output AE, and the labels of the system output AE and the reference AE are the same. 2) Its temporal centre lies between the timestamps of at least one reference AE, and the labels of both the system output AE and the reference AE are the same. Similarly, a reference AE is considered correctly detected if at least one of two conditions is met: 1) There exists at least one system output AE whose temporal centre is situated between the timestamps of the reference AE, and the labels of both the system output AE and the reference AE are the same. 2) Its temporal centre lies between the timestamps of at least one system output AE, and the labels of the system output AE and the reference AE are the same.

The AED-ACC metric was used in the last CLEAR’2007 [CLE07] international evaluation, supported by the European Integrated project CHIL [WS09] and the US National Institute of Standards and Technology (NIST).

4.3 Detection approaches

4.3.1 Joint segmentation and classification using HMMs

Acoustic event detection requires both segmentation of the audio stream, and classification of the segments. We perform simultaneous segmentation and classification using similar to state-of-the-art methods for continuous speech recognition [NO99].

The goal of AED can be formulated as follows: find the event sequence $\Omega = (c_1, c_{1, \ldots,} c_M)$ that maximizes the posterior probability given the observation vector $O = (o_1, o_2, \ldots, o_T)$:

$$\hat{\Omega} = \arg\max_{\Omega} P(\Omega \mid O) = \arg\max_{\Omega} P(O \mid \Omega) P(\Omega) \qquad (4.3.1)$$

The acoustic model P(O|Ω) is one HMM for each AE, that has several emitting states connected with ergodic or left-to-right transitions. P(Ω) is a prior probability of AE sequence Ω. In order to avoid the dependence of AE sequence to the particular recording scenario we assume that all sequences of AEs are equally probable. The observation distributions of the states are incrementally-trained Gaussian mixtures with continuous densities. Each HMM is trained with the signal segments belonging to the corresponding event class from development data, using the standard Baum–Welch training algorithm [RJ93]. The HTK toolkit [YEK02] is used for training and testing the HMM–GMM system. The HMM topology for each AE is determined during a cross-validation procedure on the development data. The number of emitting states and Gaussian mixtures per state depends much on the amount of available training data. Usually the number of emitting states for each meeting-room AE ranges from 1 to 5 and the number of Gaussian mixtures ranges from 2 to 16. Note the two parameters, number of emitting states and number of Gaussians, usually compensate each other: with increasing number of states the number of Gaussian mixtures needed to model each state decreases. For testing, the Viterbi algorithm is used to find the sequence of states with highest probability, resulting in a sequence of detected AEs.

Although the multi-class segmentation using HMMs is usually performed within a one single pass, in our work we exploit the parallel structure of the binary detectors depicted in Figure 4.3.1. Firstly, the input signal is processed by each binary detector independently (the total number of detectors is equal to the number of AE classes N), thus segmenting the input signal in intervals either as “Class” or “non-Class”. Using the training approach known as one-against-all method [RK04], all the classes different from “Class” are used to train the “non-Class” model. Secondly, the sequences of decisions from each binary detector are superimposed together to get the final decision.

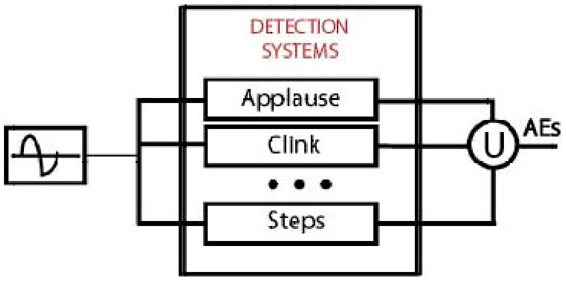


Figure 4.3.1. A set of binary detectors working in parallel

The proposed architecture with *N* separate HMM-based binary detectors working in parallel has several advantages:

1. For each particular AE, the best set of features is used. The features which are useful for detecting one class are not necessarily useful for other classes. In our case, the video features are used only for detecting some particular classes.

2. The trade-off between the number of misses and false alarms can be optimized for each particular AE class.

3. In the case of overlapped AEs, the proposed system can provide multiple decisions for the same audio segment.

However, this architecture requires *N* binary detectors that makes the detection process more complex in the case of a large number of AE classes.

4.3.2 Detection-by-classification using Support Vector Machines (SVMs)

The SVM-based AED system used in the present work is the one that was also used for the AED evaluations in CLEAR 2007 [TNB08] with slight modifications. Note this system was ranked as the second best system in the international CLEAR'07 [CLE07] AED evaluation.

The scheme of the AED system is shown in Figure 4.3.2. For each frame, a set of spectral parameters has been extracted. It consists of the concatenation of two types of parameters: 1) 16 Frequency-Filtered (FF) log filter-bank energies, along with the first and the second time derivatives; and 2) a set of the following parameters: zero-crossing rate, short time energy, 4 sub-band energies, spectral flux, calculated for each of the defined sub-bands, spectral centroid, and spectral bandwidth. In total, a vector of 60 components is built to represent each frame. The mean and the standard deviation parameters have been computed over all frames in a 0.5 sec window with a 100 ms shift, thus forming one vector of 120 elements. A sequence of decisions made on a 0.5-second window every 100 ms is obtained. That sequence is smoothed by assigning to the current decision point the label that is most frequent in a string of five decision points around the current one. The sequence of decisions is then processed again to get the detected events. In that step, only the events that have their length equal or larger than the average event length are kept, and the number of events is forced to be lower than a number which is proportional to the length of the segment. The average length of the events is estimated from the training and development databases. Finally, if the average of the above-mentioned computed confidences in a detected event is less than a threshold, the hypothesized event is marked as "unknown"; otherwise, it maintains the assigned label.

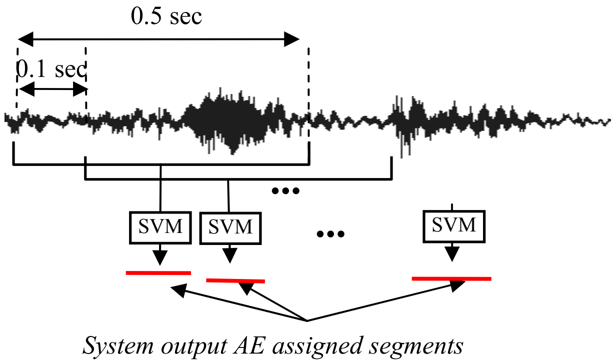


Figure 4.3.2. SVM-based AED system

The training data for each binary SVM classifier were firstly normalized anisotropicly to be in the range from –1 to 1, and the obtained normalizing template was then applied also to the testing data that are fed to that classifier. In the experiments with the SVM we used the Gaussian kernel. Leave-one-out cross validation [SS02] was applied to search for the optimal kernel parameter σ. To cope with the data imbalance we introduced different generalization parameters (C_+ and C_-) for positively and negatively labelled training samples: $C_+ = K(A_-/A_+)$, $C_-=K(A_+/A_-)$ where A_+ and A_- are the number of positive and negative training samples, respectively. K was set to value 10 for all experiments [TN09]. The MAX WINS (pair-wise majority voting) [HL02] scheme was used to extend the SVM to the task of classifying several classes. After the voting is done, the class with the highest number of winning two-class decisions (votes) is chosen.

4.4 Multimodal fusion

The information fusion can be done on data, feature, and decision levels. Data fusion is rarely found in multi-modal systems because raw data is usually not compatible among modalities. For instance, audio is represented by one-dimensional vector of samples, whereas video is organized in two-dimensional frames. Concatenating feature vectors from different modalities into one super vector is possible way for combining of audio and visual information. An alternative to feature-level fusion is to model each different feature set separately, design a specialized classifier for this feature set, and combine the classifier output scores. Each such classifier acts as an independent "expert", giving its opinion about the unknown AE. The fusion rule then combines the individual experts' match scores. This approach is referred here as decision-level fusion. In the presented work, fusion is carried out on the decision level using weighted arithmetical mean (WAM) and fuzzy integral (FI) [TMN08] [Kun04] fusion approaches. Unlike non-trainable fusion operators (mean, product), the statistical approaches WAM and FI avoid the assumption of equal importance of information sources. Moreover the FI fusion operator also takes into account the interdependences among modalities.

4.4.1 Feature-level fusion approach

In this work the feature-level fusion is implemented by concatenating the feature sets *Xs* from *S* different modalities in one super-vector

$$\boldsymbol{Z} = \boldsymbol{X_1} \cup \boldsymbol{X_2} \cup \ldots \cup \boldsymbol{X_S} \qquad (4.4.1)$$

In the case of using HMM-GMM approach, the likelihood of that observation super-vector at state *j* and time *t* is calculated as:

$$b_Z(t) = \sum_m p_m N(\boldsymbol{Z_t}; \boldsymbol{\mu_m}; \boldsymbol{\Sigma_m}) \qquad (4.4.2)$$

where $N(.;\mu;\Sigma)$ is a multivariate Gaussian pdf with mean vector μ and covariance matrix Σ, and p_m are the mixture weights.

Feature-level fusion becomes a difficult task when some features are missing. Although the audio spectro-temporal features can be extracted at every time instance, the feature that corresponds to the localization of acoustic source has an undefined value in the absence of any acoustic activity. In the experiments we substitute the missing features (*x, y, z* coordinates) with a predefined "synthetic" value (we use -1 value in our experiments). In this case we explicitly assign the 3D "position" of the silence event to have the value (-1, -1, -1).

4.4.2 Decision-level fusion approach

Multimodal decision fusion can be viewed from a broader perspective as a way of combining multiple classifiers corresponding to each modality. The main motivation here is to compensate possible misclassification errors of a certain classifier with other available classifiers and to end up with a more reliable overall decision.

In the presented experiments decision level fusion is carried out using weighted arithmetical mean (WAM) and fuzzy integral (FI) fusion approaches. Unlike non-trainable fusion operators (mean, product), the statistical approaches WAM and FI avoid the assumption of equal importance of information sources.

We are searching for a suitable fusion operator to combine a finite set of information sources $Z = \{1,...,z\}$. Let $D = \{D_1, D_2,..., D_z\}$ be a set of trained classification systems and $\Omega = \{c_1, c_2,..., c_N\}$ be a set of class labels. Each classification system takes as input a data point $x \in \Re^n$ and assigns it to a class label from Ω. Alternatively, each classifier output can be formed as an *N*-dimensional vector that represents the degree of support of a classification system to each of *N* classes. It is convenient to organize the output of all classification systems in a decision profile:

$$DP(x) = \begin{bmatrix} d_{1,1}(x)... \; d_{1,n}(x)... \; d_{1,N}(x) \\ ... \\ d_{j,1}(x)... \; d_{j,n}(x)... \; d_{j,N}(x) \\ ... \\ d_{z,1}(x)... \; d_{z,n}(x)... \; d_{z,N}(x) \end{bmatrix} \qquad (4.4.3)$$

where a row is classifier output and a column is a support of all classifiers for a class. We suppose these classifier outputs are commensurable, i.e. defined on the same measurement scale (most often they are posterior probability-like).

Let's denote h_i, i=1,..,z, the output scores of z classification systems for the class c_n (the supports for class c_n, i.e. a column from decision profile) and before defining how FI combines information sources, let's look to the conventional WAM fusion operator. A final support measure for the class c_n using WAM can be defined as:

$$M_{WAM} = \sum_{i \in Z} \mu(i) h_i \qquad (4.4.4)$$

where $\sum_{i \in Z} \mu(i) = 1, \mu(i) \geq 0 \; for\; all\; i \in Z$

The WAM operator combines the score of z competent information sources through the weights of importance expressed by $\mu(i)$. For the weights in WAM operator we use uniform class noise model with the weights computed as $\mu_i = E_i^{E_i}(1 - E_i)^{1-E_i}$ where E_i is the training error of class c_i [Kun04]. The main disadvantage of the WAM operator is that it implies preferential independence of the information sources.

Let's denote with $\mu(i, j) = \mu(\{i, j\})$ the weight of importance corresponding to the couple of information sources i and j from Z. If μ is not additive, i.e. $\mu(i, j) \neq [\mu(i) + \mu(j)]$ for a given couple $\{i, j\} \subseteq Z$, we must take into account some interaction among the information sources. Therefore, we can build an aggregation operator starting from the WAM, adding the term of "second order" that involves the corrective coefficients $\mu(i, j) - [\mu(i) + \mu(j)]$, then the term of "third order", etc. Finally, we arrive to the definition of the FI: assuming the sequence h_i, i=1,..,z, is ordered in such a way that $h_1 \leq ... \leq h_z$, the Choquet *fuzzy integral* can be computed as

$$M_{FI}(\mu, h) = \sum_{i=1}^{z} [\mu(i,...,z) - \mu(i+1,...,z)] \; h_i \qquad (4.4.5)$$

where $\mu(z+1) = \mu(ø) = 0$. $\mu(S)$ can be viewed as a weight related to a subset S of the set Z of information sources. It is called *fuzzy measure* (FM) for S, $T \subseteq Z$ it has to meet the following conditions:

$\mu(ø) = 0, \mu(Z) = 1,$ Boundary

$S \subseteq T \Rightarrow \mu(S) \leq \mu(T),$ Monotonicity

For instance, as an illustrative example let's consider the case of 2 information sources with unordered system outputs h_1=0.4 and h_2=0.3, and corresponding fuzzy measures μ(1)=0.6 and μ(2)=0.8. Note that μ(0)=0 and μ(1,2)=1. In that case, the Choquet *fuzzy integral* is computed as $M_{FI}(\mu,h)$= (μ(1,2)- μ(1))h_2+ μ(1)h_1=0.36.

A large flexibility of the FI aggregation operator is due to the use of FM that can model importance and interaction among criteria. And although the FM $\mu(i)$ provides an initial view about the importance of information source *i*, all possible subsets of Z that include that information source should be analysed to give a final score. For instance, we may have $\mu(i) = 0$, suggesting that element *i*, $i \notin T$, is not important; but if, at the same time, $\mu(T \cup i) >> \mu(T)$, this actually indicates *i* is an important element for the decision. For calculating the importance of the information source *i*, the Shapley score [Gra95] is used. It is defined as:

$$\phi(\mu,i) = \sum_{T \subseteq Z \setminus i} \frac{(|Z| - |T| - 1)!|T|!}{|Z|!}[\mu(T \cup i) - \mu(T)] \qquad (4.4.6)$$

Generally, (4.4.6) calculates a weighted average value of the marginal contribution $\mu(T \cup i) - \mu(T)$ of the element *i* over all possible combinations. It can be easily shown that the information source importance sums to one.

4.5 Experimental results

4.5.1 Improving Detection of “Steps” using audio-visual decision-level fusion

Motivated by the fact that the “Steps” AE accounted for almost 35% of all acoustic events in the CLEAR’07 evaluation database, in the first set of experiments we use video 3D tracking information to improve the detection of that particular class. Detection of AEs is carried out with one video-based and two audio-based systems: HMM-GMM-based and SVM-based. The use of the three AED systems is motivated by the fact that each system performs detection in a different manner. The difference in the nature of the considered detection systems makes the fusion promising for obtaining a superior performance. The HMM-GMM-based AED system (described in Section 4.3.1) segments the acoustic signal in events by using a frame-level representation of the signal and computing the state sequence with highest likelihood. The SVM-based system (described in Section 4.3.2) does it by classifying segments resulting from consecutive sliding windows. The video-based system uses information about position of people in the room. It utilizes velocity feature and probabilistic classifier for “Steps”/“non-Steps” detection as described in following. The normalized histograms of the logarithm of velocity for “Steps” and “non-Steps” obtained from development seminars as depicted in Figure 3.4.3 (b), from which can be seen that “Steps” are more likely to appear with higher values of velocity. The jerky nature of the “Steps” hump results from a more than 10 times scarcer representation of “Steps” with respect to “non-Steps” in the development database. These two curves are approximated by two Gaussians via Expectation-Maximization algorithm (EM). During detection on testing data the final decision for “Steps”/ “non-Steps” classes is made using the maximum aposteriori estimate:

$$P(w_j \mid x) = P(x \mid w_j)P(w_j)\,, j=\{1, 2\} \qquad (4.5.1)$$

where $P(w_1)$ and $P(w_2)$ are prior probabilities for the class “Steps” and the meta-class “non-Steps” respectively, which are computed using the prior distribution of these two classes in development data and $P(x|w_j)$ are likelihoods given by the Gaussian models.

In the experiments, late fusion is performed via combining the decisions from several information sources. In our case, not all information sources give scores for all classes. Unlike SVM and HMM-GMM-based systems, which provide information about 15 AE classes, the video-based system scores are given only for the class “Steps” and the meta-class “non-Steps”. Fusion of information sources using the late fusion can be done either by transforming (extending) the score for "non-Steps" from the video-based system to the remaining 15 classes which do not include "Steps" or, vice-versa, transforming (restricting) the scores of 15 classes provided by the SVM and HMM-GMM-based systems to one score for the class "non-Steps". In the former case, the fusion is done at one stage with all the classes. In the latter, a two-stage approach is implemented, where on the first stage the 3 detection systems are used to do "Steps"/ "non-Steps" classification and on the second stage the subsequent classification of the "non-Steps" output of the first stage is done with both SVM and HMM-GMM-based systems. The one-stage and two-stage approaches are schematically shown in Figure 4.5.1

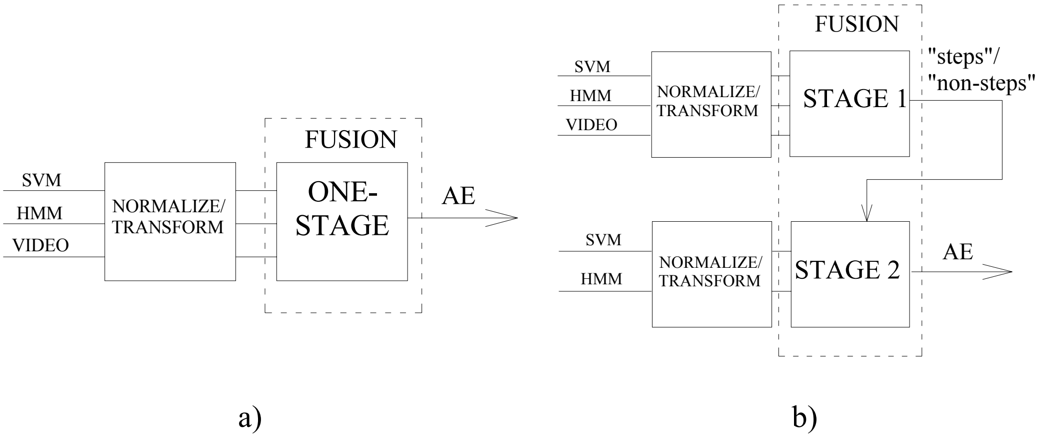


Figure 4.5.1. One-stage (a) and two-stage (b) decision-level fusion

For one-stage fusion (Figure 4.5.1 (a)) the score V of "non-Steps" of the video-based system is equally distributed among the remaining 15 classes assigning to each of them score V before applying *soft-max* normalization. At the first stage of the two-stage approach, all the classes not labelled as "Steps" form the "non-Steps" meta-class. The final score of "non-Steps" is chosen as maximum value of scores of all the classes that formed that meta-class.

The individual FMs for the fuzzy integral fusion are trained on development data in our work using the gradient descent training algorithm [Gra95]. The 5-fold cross validation on development data was used to stop the training process to avoid overtraining. The tricky point was that during training the algorithm minimizes the total error on development data. As the number of data per each class is non-uniform distributed, during the training process the number of detection mistakes for the most representative classes ("Speech", "Silence") is decreased at the expense of increasing errors on the classes with lower number of representatives. The final metric scores, however, only

12 classes which are the classes with much smaller number of representatives than e.g. “Speech”. This way, the FI with the trained FM measure tends to detect correctly the classes that are not scored by the metric. To cope with this problem, we firstly fixed the FM of the classes of no interest (“Speech”, “Unknown”, and “Silence”) to be in the equilibrium state [Gra95] and, secondly, calculate the cross-validation accuracy only for the classes of interest.

In order to fuse 3 information sources (SVM-based, HMM-GMM-based, and video-based systems), their outputs must be synchronized in time. In our case, the SVM system provides voting scores every 100ms, the video-based system every 40ms, and the HMM-GMM system gives segments of variable length which represent the best path through the recognition network. The outputs of the 3 systems were reduced to a common time step of 100ms. For that purpose the output score of the video-based system was averaged on each interval of 100ms, while for the HMM-GMM system each segment was broken into 100ms-long pieces.

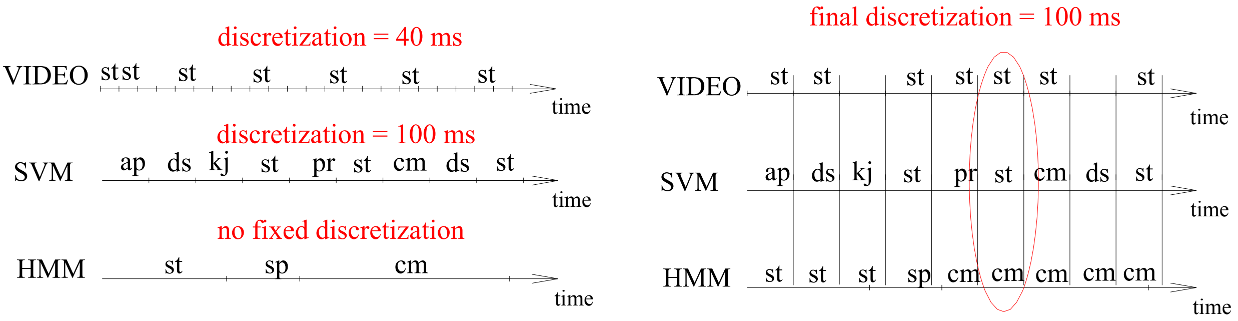


Figure 4.5.2. Synchronization of different information sources

On the other hand, to make the outputs of information sources commensurable we have to normalize them to be in the range [0 1] and their sum equal to 1. As it was mentioned in Section 4.3.2, when the SVM classification system is used alone, after voting, the class with the highest number of winning two-class decisions (votes) is chosen. In case of a subsequent fusion with other classification systems numbers of votes obtained by non-winning classes were used to get a vector of scores for the classes. For the HMM-GMM system, each hypothesis of an AE given by the optimal Viterbi segmentation of the seminar is then decoded by the trained HMM-GMM models of winning and each non-winning AE class in order to obtain the corresponding log-likelihood values which form vector of scores. In case of video-based AED system we obtain scores for the two classes “Steps” and “non-Steps” as the distance between the values of log-velocity and the decision boundary. To make the scores of video-based and HMM-GMM-based systems positive *min-max* normalization [SSK06] is used.

The *soft-max* function is then applied to the vector of scores of each detection system. This function is defined as:

$$q_i\Big|_{normalized} = \exp(k * q_i) / \sum_i \exp(k * q_i) \qquad (4.5.2)$$

where the coefficient *k* controls the distance between the components of the vector [q_1, q_2, ...,q_N]. For instance, in extreme case when *k=0*, the elements of the vector after *soft-max* normalization would have the same value *1/N*, and when *k*→∞ the elements tend to become binary. The normalization coefficients are different for each AED system, and they are obtained using the development data.

The results of first-stage fusion for "Steps"/"non-Steps" detection are presented in Figure 4.5.3. It can be seen that fusion of SVM and HMM-GMM-based systems leads to a small improvement, while in combination with video information the improvement is noticeable. It is worth to mention that 48.1% of accuracy for "Steps" detection would indicate a little worse decision than random choice if the metric scored both "non-Steps" meta-class and "Steps" class. However, in our case, only the "Steps" class is scored and thus 48.1% indicates that not only around 48.1% of "Steps" are detected (recall) but also that 48.1% of all produced decisions are correct (precision). On the first stage the FI fusion gives superior results in comparison with WAM fusion. This indicates that a certain interaction between information sources for "Steps" detection exists that can not be captured by WAM fusion operator.

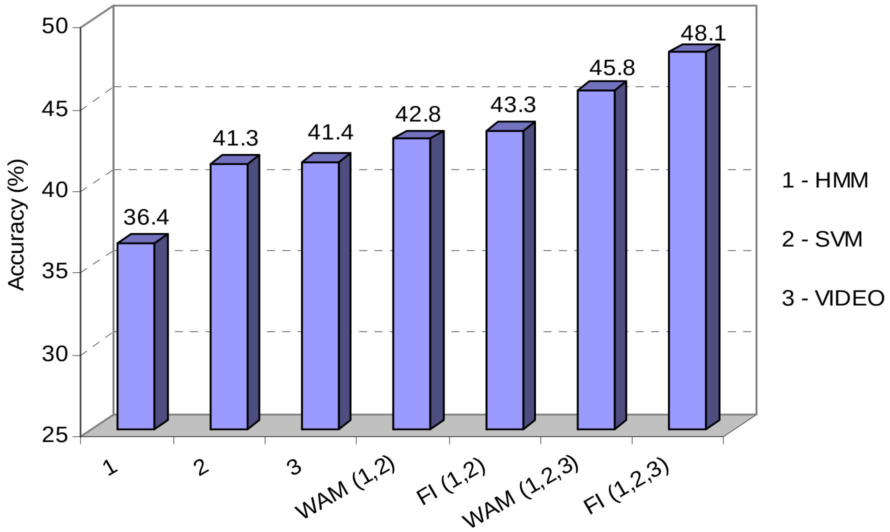


Figure 4.5.3. Accuracy of "Steps" detection on the first stage

The final results of detection of all 12 classes of AEs are presented in Figure 4.5.4. It can be seen that total system accuracy benefits from better recognition of "Steps" class. Again in this experiment the FI fusion shows better performance then WAM, resulting in a final accuracy of 40.5%.

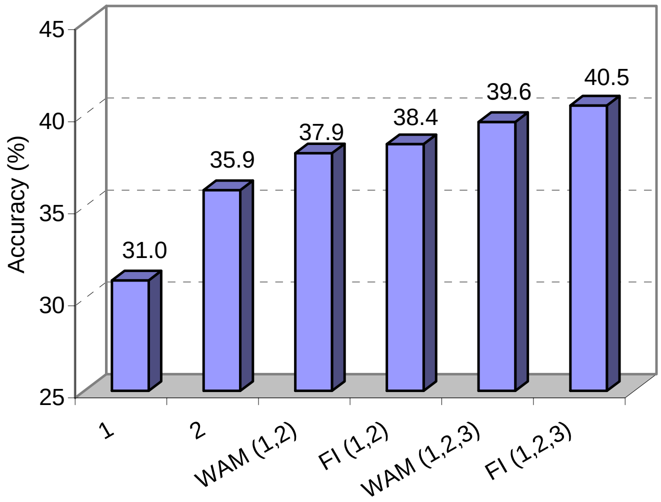


Figure 4.5.4. Overall system accuracy based on two-stage fusion

Previously explained one-stage fusion showed lower scores - only 38.8% for WAM and 39.2% with FI. This fact may indicate that in our particular case spreading no-information for classes with missing scores can be harmful and, conversely, to compress the scores of many classes to binary problems can be more beneficial.

4.5.2 Feature-level fusion of audio spectro-temporal, acoustic localization and video features

In the previous sub-Section we have shown how additional information from video 3D tracking system can improve the recognition results of "Steps" AE in CLEAR'07 evaluation database. In order to extend the multimodal AED to more classes, a new database described in Section 4.2 has been recorded which provided a close view of the subjects under study. A feature-level fusion strategy is used, and a parallel structure of binary HMM-GMM-based detectors is employed. In the experiments presented here video feature extraction is extended to 5 AE classes, and "Speech" class, is also evaluated in the final results. A statistical significance test is performed individually for each AE.

The overall diagram of the proposed system is depicted in Figure 4.5.5. Three data sources are combined together: two come from audio and one from video. The first is obtained from single channel audio processing and consists of audio spectro-temporal (AST) features. The second is obtained from microphone array processing and consists of the 3D location of the audio source. And the third is obtained from multiple cameras covering the scenario and consists of video-based features related to several AEs. The three types of features are concatenated together (feature-level fusion) and supplied to the corresponding binary detector from the set of 12 detectors that work in parallel.

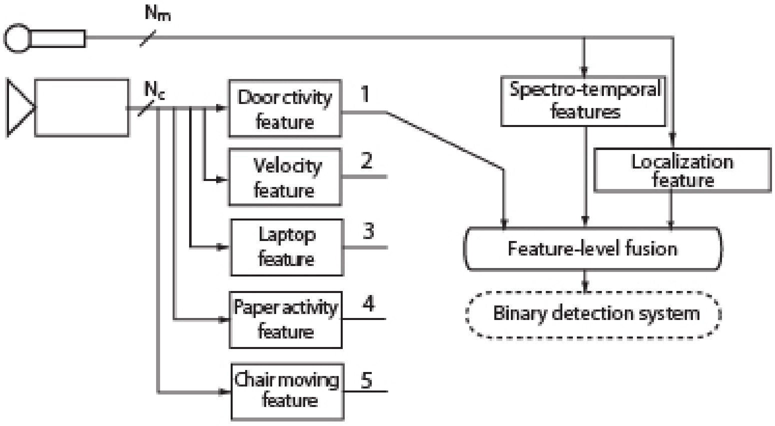


Figure 4.5.5. System flow-chart

In order to assess the performance of the proposed multimodal AED system and show the advantages of the proposed feature sets, the multimodal database of isolated AEs described in Section 4.2 was used for both training and testing: 8 sessions were randomly permuted; odd index numbers were assigned to training and even index numbers to testing. Six permutations were used in the experiments. The subset of spontaneously generated AEs was used in the final experiments in order to check the adequateness of the multimodal fusion with real world data. The HMM-GMM-based AED system described in sub-Section 4.3.1 is used for AED.

The detection results for each mono-modal detection system are presented in Table 4.5.1 (for the database of isolated AEs only). The baseline system (first column) is trained with the 32 AST features (16 FF LFBE plus the first time derivatives), while the other two systems use only one feature coming from either the video or the localization modality, respectively. As we see from the table, the baseline detection system shows high recognition rates for almost all AEs except the class "Steps" that is much better detected with the video-based AED system. The recognition results for the video-based system are presented only for those AEs for which video counterpart is taken into consideration. In the case of localization-based AED system, the results are presented only for each category rather than the particular AE class. In fact, using the localization information we are able to detect just the category but not the AE within it.

Table 4.5.1: Mono-modal recognition results

	AST (%)	Video (%)	Localization (%)
Door knock	97.20	---	82.95
Door slam	93.95	79.96	
Chair moving	94.73	77.28	83.15
Steps	60.94	75.60	
Paper work	94.10	91.42	86.31
Keyboard	95.57	81.98	
Cup clink	95.47	---	
Key jingle	89.73	---	67.70
Phone ring	89.97	---	
Applause	93.24	---	
Cough	93.19	---	
Speech	86.25	---	

The confusion matrix that corresponds to the baseline detection system is presented in Table 4.5.2, which presents the percentage of hypothesized AEs (rows) that are associated to the reference AEs (columns), so that all the numbers out of the main diagonal correspond to confusions. This table shows that some improvement may be achieved by adding localization-based features. For instance, although the “below-table” AEs (“Chair moving” and “Steps”) are mainly confused with each other, there is still some confusion among these two AEs and the AEs from other categories.

Table 4.5.2. Confusion matrix corresponding to the baseline system (in %)

	kn	**ds**	**cm**	**st**	**pw**	**kt**	**cl**	**kj**	**pr**	**ap**	**co**	**sp**
kn	98.8	0.4	0	0	0	0	0	0	0	0	0.8	0
ds	0.3	82.0	0	14.8	0.1	1.2	0.4	0.1	0.2	0.2	0.2	0
cm	0.9	0.4	93.8	4.0	0.4	0	0	0	0	0	0.1	0.3
st	0	18.1	13.8	65.4	1.2	0.5	0	0	0.2	0.4	0	0.4
pw	0	0.3	0	0.3	85.6	10.5	0	1.0	0.3	2.0	0	0
kt	0	0	0	0	0	98.9	0	0.8	0.4	0	0	0
cl	0	2.0	0	0	0	0	94.9	1.0	2.0	0	0	0
kj	0	0	0	0	5.0	0.8	0	89.5	4.7	0	0	0
pr	0	0	0	0	0	0	0	1.0	87.8	0.3	0	10.9
ap	0	0	0	0	1.2	0	1.2	0	0	97.6	0	0
co	6.9	0.4	0	0	0	0	0	0	0	0	92.4	0.4
sp	1.8	0.7	0	5.8	0	0	0	0	3.6	0	7.6	80.6

The final detection results for isolated and spontaneously generated AEs are presented in Table 4.5.3. The first column corresponds to the baseline system (that uses the 32-dimensional AST feature vector). The next columns correspond to the fusion of baseline features with the localization feature, the video feature and the combination of both of them, respectively. The last column shows the *p*-value of the statistical significance of the AST+L+V test in relation to the baseline system. If P_1 and P_2 are the accuracy measures for the baseline and the multimodal AED system, respectively, the null hypothesis H_0 is*:* $P_1 \geq P_2$; and the alternative hypothesis H_1 is*:* $P_1 < P_2$. Assuming a standard level of significance at 95%, a *p*-value that is less than *0.05* implies the rejection of the null hypothesis or, in other words, it means that the result is statistically significant.

Although the AST+L+V system improves the baseline system for most of the isolated AEs, a statistically significant improvement is only obtained for the classes "Door slam", "Door knock", and "Steps". For the data subset of spontaneously generated AEs, a significant improvement in the detection of some low energy AEs ("Steps", "Paper work", "Keyboard typing") is achieved. The best relative improvement corresponds to the "Steps" class. Other AEs have slightly improved their detection rates. In average, 15% relative error-rate reduction for isolated AEs, and 21% for spontaneously generated AEs are achieved.

Table 4.5.3. Fusion of different modalities using isolated and spontaneously generated AEs

AEs	Isolated					Spontaneously generated			
	AST	AST+L	AST+V	AST+L+V	p-value	AST	AST+L	AST+V	AST+L+V
Door knock	97.20	98.81	97.20	98.81	0.05	88.72	90.45	88.72	90.45
Door slam	93.95	95.35	97.06	96.72	0.01	75.45	82.89	85.04	87.36
Chair moving	94.73	95.18	95.24	95.93	0.09	83.89	84.32	84.12	84.82
Steps	60.94	72.51	78.09	77.25	0.04	58.56	57.12	67.12	66.58
Paper work	94.10	94.19	95.16	95.07	0.30	65.14	62.61	73.18	79.32
Keyboard	95.57	95.96	96.56	96.72	0.37	71.69	78.37	79.68	80.50
Cup clink	95.47	94.03	95.47	94.03	0.86	90.35	86.08	90.35	86.08
Key jingle	89.73	88.00	89.73	89.60	0.52	52.09	44.12	52.09	44.12
Phone	89.97	88.09	89.97	88.79	0.64	87.98	90.45	87.98	90.45
Applause	93.24	94.91	93.24	94.91	0.13	84.06	84.65	84.06	84.65
Cough	93.19	94.20	93.19	94.20	0.35	76.47	82.36	76.47	82.36
Speech	86.25	85.47	86.25	85.47	0.62	83.66	83.12	83.66	83.12
Average	**90.36**	**91.39**	**92.26**	**92.29**	**-**	**76.51**	**77.21**	**79.37**	**79.98**

As it can be observed, the video information improves the baseline results for the five classes for which video information is used, especially in the case of spontaneously generated AEs where the acoustic overlaps happen more frequently. Therefore, the recognition rate of those classes considered as “difficult” (usually affected by overlap or of low energy) increases.

Acoustic localization features improve the recognition accuracy for some AEs, but for other events, it is decreased. One of the reasons of such behaviour is the mismatch between training and testing data for spontaneously generated AEs. For instance, the “Cup clink” AE in spontaneous conditions often appears when the person is standing, which is not the case for isolated AEs. Another reason is that, for overlapped AEs, the AE with higher energy will be properly localized while the other overlapped AE will be masked. Additionally, according to the confusion matrix (Table 4.5.2), the main confusion among AEs happens inside the same category, so that the audio localization information is not able to contribute significantly.

4.5.3 Detection of overlapped with speech AEs

In previous experiments we have seen how features from additional modalities improve the baseline recognition rate of both isolated and spontaneously generated AEs. Since the baseline recognition results show high recognition accuracy (more, than 90% for most of the classes), the improvement from additional modalities is not statistical significant for many of AEs. In this Section we present new results with artificially generated database where the improvement from the video modality becomes apparent. The database consists of isolated acoustic events overlapped with speech with different SNRs as described in Section 4.2. Additionally, in the experiments we present comparison results between feature and decision level fusion approaches.

The meeting scenario adopted for the following experiments assumes that there are two simultaneous acoustic sources in the room: one is always speech and the other is a specific AE. Taking into account this assumption, our UPC's smart-room has been considered ideally subdivided in the two areas: left and right. In the left part the speaker produces speech, and in the right part the listener produces different types of AEs. This assumption allows us to analyze the left and right parts of the room independently for the extraction of acoustic source localization features.

The decision-level fusion process is schematically depicted in Figure 4.5.6. First, a HMM segmentation based on the spectro-temporal features is performed to find all non-silence segments in the input audio. Given the “Class” and “non-Class” HMM-GMM models the log-likelihood ratio (LLR) is obtained for each non-silence segment S_i and each modality separately. A high positive LLR score would mean a high confidence that the non-Silence segment belongs to the “Class”, while a low negative score would mean that the segment more likely belongs to "non-Class". A value close to zero indicates low confidence of decision. Second, the obtained scores are normalized to be in the range [0…1] and their sum equal to 1. Then the normalized values are fused together using either Weighted Arithmetical Mean (WAM) or Fuzzy Integral (FI) fusion operators.

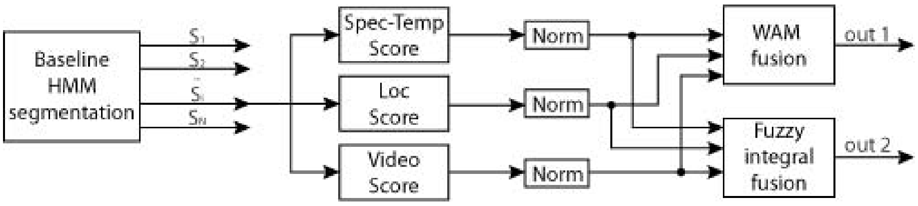


Figure 4.5.6. Flowchart of decision-level fusion

The detection results corresponding to two mono-modal AED systems based on AST and video features, respectively, are presented in Figure 4.5.7. The results for the video-based system are presented as an average accuracy score for those AEs for which the video counterpart is taken into consideration.

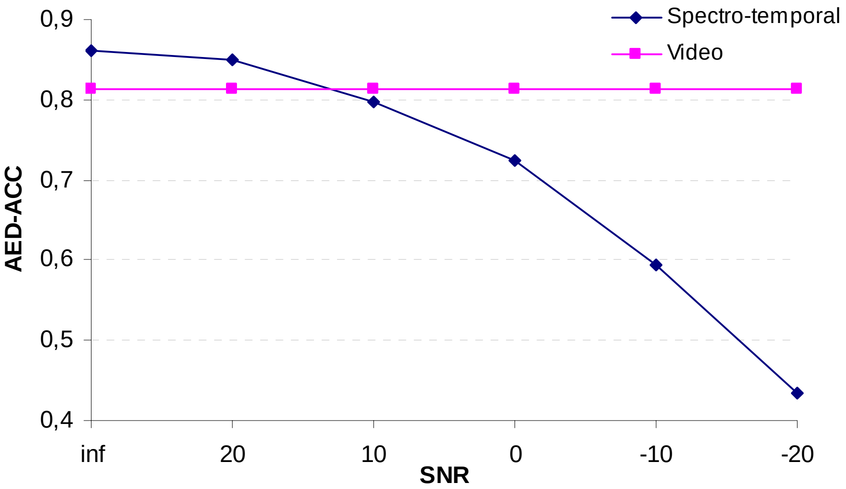


Figure 4.5.7. Mono-modal AED results

Note the recognition results do not change for different SNR conditions since the video signals are not affected by overlapped speech. We do not present results for the AED system based on localization features since the information about the position of the acoustic source enables to detect just the category but not the AE within it. As we see from Figure 4.5.7, the recognition results of the baseline system decrease significantly for low SNRs.

The average relative improvement obtained by the multimodal system with respect to the baseline system (that uses the AST features only) for different fusion techniques is displayed in Figure 4.5.8. The feature-level fusion performs better for all AEs than both WAM and FI decision-level fusion approaches, and the both decision-level fusion techniques showed similar results in our experiments.

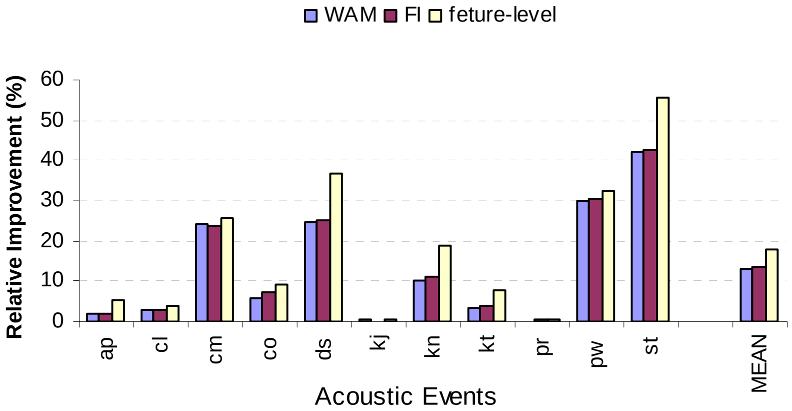


Figure 4.5.8. Average relative improvement obtained by the multimodal system

The Figure 4.5.9 summarizes the relative improvement (averaged over all AEs) obtained with the feature-level and the decision-level (using FI) fusion techniques for AED along different SNRs. Notice that the relative improvement from additional modalities at high SNRs is minimal due to the fact that in clean conditions the baseline results are already high. The same observation has been already made from Table 4.5.3. However, when the level of noise increases, the baseline results based on AST features drop down drastically yielding to increase of relative improvement coming from additional modalities.

Moreover, according to both Figure 4.5.8 and Figure 4.5.9 the fusion performed at the feature level showed better performances than both of those performed at the decision-level, highlighting that processing input data in a joint feature space is more successful. It is known fact that early integration techniques, if adequately used, are usually favoured if a couple of modalities is highly correlated. Additionally, fuzzy integral fusion technique shows higher recognition rate than WAM fusion indicating about certain interaction among information sources.

The information about the importance of each modality can be extracted applying the Shapley score to the FMs. Figure 4.5.10 shows the importance of the three information sources coming from spectro-temporal, acoustic localization and video features. The importance is calculated for those five AEs for which the video counterpart is taken into consideration. It can be observed that for

"Door slam", "Steps" and "Paper wrapping" AEs the video-based detection system demonstrate the highest importance, while for the rest two AEs, "Chair moving" and "Keyboard typing", the detection system based on acoustic spectro-temporal shows the superior importance. Note acoustic localization features exhibit minor importance for AED since the main confusion between classes occur within the same meta-class category. Another reason of low importance is the difficultness of precise estimation of localization coordinates for low-energy AEs like "Paper work", "Steps" and "Keyboard typing".

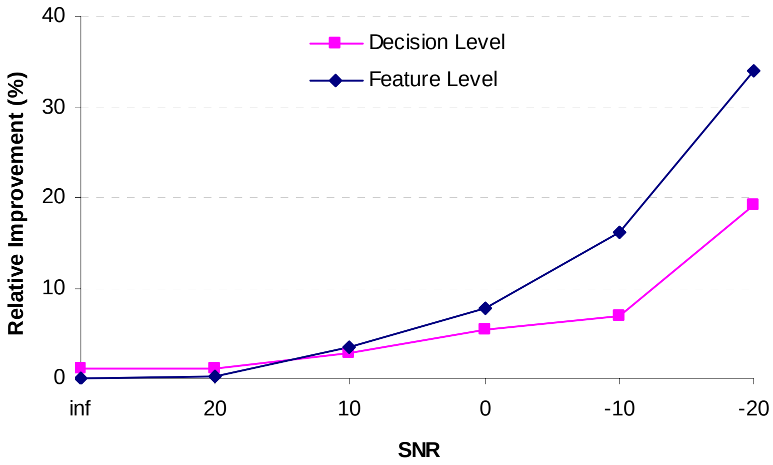


Figure 4.5.9. The relative improvement obtained from multimodal features for different SNRs

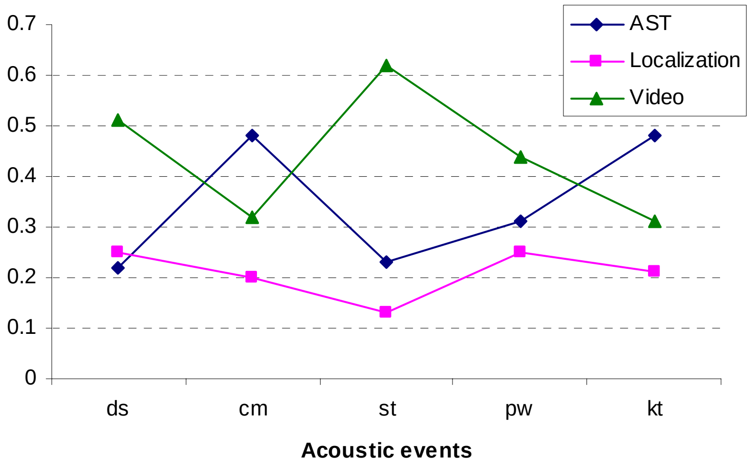


Figure 4.5.10. The importance of different modalities calculated with Shapley score

4.6 Chapter Summary

In this chapter, we have shown that video signals can be a useful additional source of information to cope with the problem of acoustic event detection. Using an algorithm for video 3D tracking, video-based features that represent the movement have been extracted, and a probabilistic classifier for "Steps"/"non-Steps" detection has been developed. The fuzzy integral and WAM techniques were used to fuse the outputs of both video-based detector and two audio-based AED systems which use either SVM or HMM classifiers. Using the CLEAR'07 evaluation database, the results showed the effectiveness of the multimodal fusion for detection of "Steps" AE.

Additionally, a multimodal system based on a feature-level fusion approach and a one-against-all detection strategy has been presented and tested with a new audiovisual database. The acoustic data is processed to obtain a set of spectro-temporal features and the localization coordinates of the sound source. A number of features are extracted from the video signals by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of several AEs. Experimental results show that information from the microphone array as well as the video cameras facilitates the task of AED for both datasets of AEs: isolated and spontaneously generated. Since in the video modality the acoustic sources do not overlap (except for occlusions), a significant error-rate reduction is obtained. The acoustic localization features also improve the results for some particular classes of AEs. The combination of all features produced higher recognition rates for most of the classes, being the improvement statistically significant for a few of them.

Using the database of isolated AEs artificially overlapped with speech we demonstrated that a significant improvement from additional modalities can be obtained in the conditions where the audio signals are strongly overlapped with speech. Additionally, the recognition results showed that fusion performed at the feature level is better than both of those performed at the decision-level, highlighting that processing input data in a joint feature space is more successful.

Chapter 5. Broadcast news audio segmentation

5.1 Chapter Overview

The recent fast growth of available audio or audiovisual content strongly demands tools for analyzing, indexing, searching and retrieving the available documents. Given an audio document, the first processing step usually is audio segmentation (AS), i.e. the partitioning of the input audio stream into acoustically homogeneous regions which are labelled according to a predefined broad set of classes like speech, music, noise, etc.

This chapter is organized as follows. Section 5.2 presents the database and metric description for the audio segmentation task. In Section 5.3 and Section 5.4 the overview of the proposed hierarchical structures of binary detectors for broadcast news AS using two databases is introduced. In Section 5.5 we present an example of how AS can be applied in speaker diarization task. Section 5.6 presents the overview of Albayzín-2010 evaluation and the main conclusions from the submitted systems and results. Inspired by the results from this evaluation, in Section 5.7 a reference audio segmentation system is constructed. Finally, in Section 5.8 we present some modifications to evaluation metric and observe how these changes may affect the evaluation results.

5.2 Databases, metric and detection approaches

5.2.1 Class definition and metric

Audio segmentation is the task of segmenting a continuous audio stream in terms of acoustically homogenous regions. The definition of acoustic classes depends much on the database and application domain. In [LHH01] the authors propose a method for robust speech, music, environment noise and silence segmentation of the audio recorded from different channels such as studio, telephone etc. In [NL05] the audio stream from the broadcast news domain is segmented into 5 different types including speech, commercials, environmental sound, physical violence and silence. The content based retrieval using TV programs is considered in [LSD01], where 7 similar classes are defined. In our work we consider 8 acoustic classes presented in Table 5.2.1.

Table 5.2.1. The acoustic classes defined for audio segmentation task

Class	Description
Speech [sp]	Clean speech from a close microphone without any sound in background
Music [mu]	Music is understood in a general sense
Speech over music [sm]	Overlapping of speech and music classes or speech with noise in background and music classes
Speech over noise [sn]	Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation)
Telephone speech [ts]	Telephonic interventions from the viewers during TV show. These interventions are mixed in the program's main audio stream
Telephone speech over music [tm]	The same as previous class but additionally there is music in the background.
Silence [si]	Absence of any acoustic activity

The metric is defined as a relative error averaged over all acoustic classes (AC):

$$Error = \underset{i}{average}(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}) \qquad (5.2.1)$$

where

dur(miss$_i$) is the total duration of all deletion errors (misses) for the *i*th AC,

dur(fa$_i$) is the total duration of all insertion errors (false alarms) for the *i*th AC,

dur(ref$_i$) is the total duration of all the *i*th AC instances according to the reference file.

An incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. A forgiveness collar of 1 sec (both + and -) is not scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an AC begins/ends.

The proposed metric is slightly different from the conventional NIST metric for speaker diarization, where only the total error time is taken into account independently of the AC. 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). This metric stimulates to detect well not only the best-represented classes, but also the minor in duration classes.

5.2.2 Databases

5.2.2.1 Àgora database

The database consists of 43h and 25m of spontaneous speech in the context of the debate TV program. Each program has been cut in two parts to exclude the commercials, and each part has duration of about 40 minutes. Àgora is a highly moderated program where around 7 different speakers discuss a wide variety of topics. The Àgora program has a fairly fixed structure, although no use of this information has been made in order to keep the system general.

The distribution of acoustic classes (ACs) in the database is presented in Table 5.2.2. Since silences are not labelled, the evaluation of “Silence” class is not included in evaluation task. Moreover, “Telephone speech” class is poorly represented in the database, so this class is not evaluated either.

Table 5.2.2. Distrbution of the acoustic classes in the Àgora database

Acoustic class	Appearance (%)
Speech	85
Speech over music	10
Telephone speech	0.02
Telephone speech over music	2
Music	2

5.2.2.2 3/24 TV channel database

The database consists of Catalan broadcast news audio from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was manually annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project. The database includes around 87 hours of annotated audio (24 files of approximately 4 hours long each). The manual annotation of the database was performed in 2 passes. A first annotation pass segmented the recordings with respect to background sounds (speech, music, noise or none), channel conditions (studio, telephone, outside and none), speakers, and speaking modes. A second annotation pass provided speech transcriptions and acoustic events (such as throat, breath, voice, laugh, artic, pause, sound, rustle or noise). For the proposed evaluation we took into account only the first pass of annotation. According to this material, a set of audio classes was defined (Table 5.2.3) which includes overlapping of speech with either music or noise.

Table 5.2.3. Distrbution of the acoustic classes in tv3/24 database

Acoustic class	Appearance (%)
Speech	37
Speech over music	15
Speech over noise	40
Music	5

The remaining 3% is referred to “Other”, any type of audio (including silence and noises) that does not correspond to the other four classes and they are not evaluated in the experiments. Although 3/24 TV is primarily a Catalan television channel, the recorded broadcasts contain a proportion of roughly 17% of Spanish speech segments. The gender conditioned distribution indicates a clear unbalance in favour of male speech data (63% versus 37%).

5.2.3 Hierarchical detection approach

The hierarchical audio segmentation architecture is a group of detectors (called modules), where each module is responsible for detection of one acoustic class of interest. As input it uses the output of the preceding module and has 2 outputs: the first corresponds to audio segments detected as the corresponding class of interest, and the other is the rest of the input stream. One of the most important decisions when using this kind of architecture is to put the modules in the best order in terms of information flow, since some modules may benefit greatly from the previous detection of certain classes. For instance, previous detection of the classes that show high confusion with subsequent classes potentially can improve the overall performance.

On the other hand, in this type of architecture, it is not necessary to have the same classifier, feature set and/or topology for different detectors. Tuning of parameters is done in each the system independently, and the two-class detection can be done in a fast and easy way. Given the modules, the detection accuracy can be computed individually and a priori. Those modules with best accuracies are then placed in the early stages to facilitate the subsequent detection of the classes with worst individual accuracies.

5.3 Hierarchical audio segmentation system for Àgora database

The flow diagram of hierarchical architecture for audio segmentation in Àgora database is presented in Figure 5.3.1. The whole detection system consists of 5 binary detectors. Although silence is not included in our evaluation task, prior segmentation of this class facilitates the detection of other classes of interest. Each detector is based on specific feature set and classification algorithm. The details of each detector are described in following.

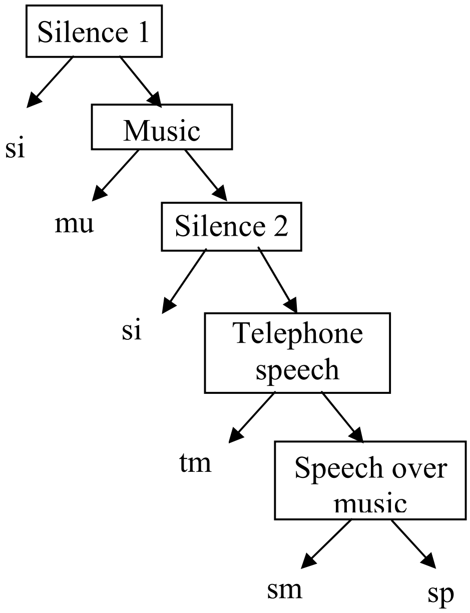


Figure 5.3.1. Flow diagram of the hierarchical architectures for Àgora database

5.3.1 Silence detectors

As depicted in Figure 5.3.1 there are two silence detectors in the proposed hierarchical structure. Since there are no references for the silences both of them are trained in unsupervised manner. The first one is intended to detect the most confident “Silence” segments. This is done to avoid confusion with silences that have musical spectra and it facilitates the detection of “Music” class in subsequent block. The detection is based on the derivative of the short time energy proposed in [LZZ02].

The second silence detector removes the rest of the silences to prepare the signal for the subsequent modules. The algorithm can be described in the following:

1. The short-time energy of the signal is transformed to the logarithmic scale, and a GMM of N Gaussians is trained. The Gaussian mixtures with a lower weight than a fixed percent of the weight of the Gaussian with the highest weight are discarded (if any).
2. This GMM is decomposed into two clusters of Gaussians (Figure 5.3.2). One cluster of Gaussians will be used to generate a GMM for "Silence" class and the other cluster for non-"Silence". In order to cluster the Gaussians they are sorted based on their mean. The N_{sil} Gaussians with the lowest mean are selected for the "Silence" class (as they represent the frames with low energy). The $N - N_{sil}$ other Gaussians are left for the "non-Silence" class. In order to do determine N_{sil} we assume that there is not a smooth transition in the energy between silence and non-silence. Based on this assumption the detector calculates the difference of the means for the sorted array of Gaussians. This lack of smooth transition makes that at some point there is a big difference between two consecutive Gaussians. The position of this gap is used to define the clusters of Gaussians. Two GMMs are formed from the clusters using the means and variances estimated previously, modifying the weights to meet the requirement of total area equal to one. The ratios between the different weights of the Gaussians inside a GMM are respected, the weights are only scaled. The Figure 5.3.2 depicts the decomposition of one GMM (red line) into two GMMs, one for silence (dashed blue line) and another one for non-silence (dashed green line), with the histogram of the feature in the background (gray color).

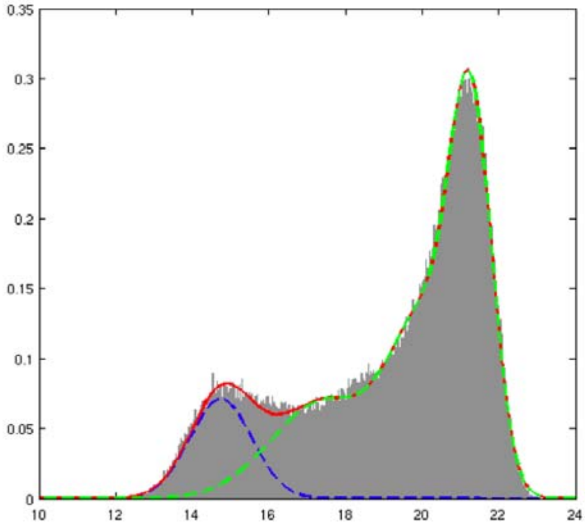


Figure 5.3.2. Two clusters of Gaussians corresponding to "Silence" and "non-Silence" classes with the histogram of the feature in the background

3. Given silence and non-silence GMMs the whole shows are evaluated frame by frame. Comparing log likelihood ratios (LLRs) with zero, each frame is classified as silence or non-silence. The decisions are smoothed using a median filter. Finally, silences shorter than the specified minimum duration are discarded.

5.3.2 Music and Speech over music detectors

Music segments usually appear at the beginning and the end of the shows or when the topic of discussion changes. Music serves as introduction to shows and it attracts attention of the audience towards its beginning. Often the discussion starts when music is still in the background. It is worth mentioning that the melody in Àgora shows does not vary much and only 2 or 3 different musical instruments could be distinguished: drums, saxophone and piano.

The differences between "Music" and "non-Music" class can be noticed in the spectral domain. The periodograms of 5 sec long "Music" and "Speech" segments are displayed in Figure 5.3.3. According to it, the spectral envelope is flatter for "Music" class while for "Speech" class. Conventional ASR features are used in both "Music" and "Speech over music" detectors: 16 FF LFBE coefficients with their first time derivatives, mean normalization is applied. Each acoustic class is modelled using HMM-GMM. Both "Music" and "Speech over music" HMMs consist of 2 emitting states with 5 Gaussians per state, while "non-Music" and "non-Speech over music" HMMs have 3 emitting states and 9 Gaussians per state. All the models have left-to-right connected state transitions.

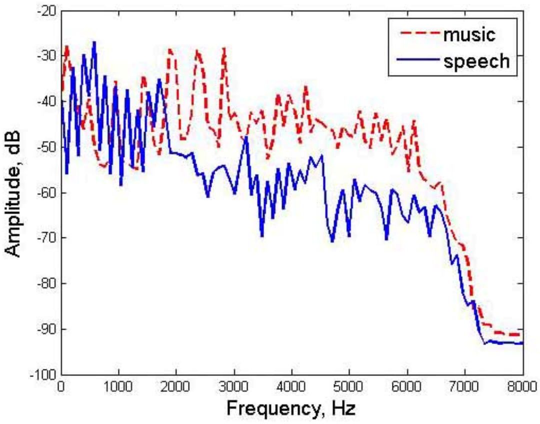


Figure 5.3.3. Periodograms corresponding to "Speech" and "Music" classes

5.3.3 Telephone speech over music detector

Some sections of the program have telephonic interventions from the viewers. These interventions are mixed in the program's main audio stream as a wide band stream.

The SVM-based AED system described in Section 4.3.2 is used. The audio signal is framed using 30 ms Hamming window and 10 ms shift. For each frame, a set of spectral parameters has been extracted. It consists of the concatenation of two types of parameters: 1) 16 FF LFBE, along with the first and the second time derivatives; and 2) a set of the following parameters: zero-crossing rate, short time energy, 4 sub-band energies, spectral flux, calculated for each of the defined sub-bands, spectral centroid, and spectral bandwidth. In total, a vector of 60 components is built to represent each frame. The mean and the standard deviation parameters have been computed over all frames in a 0.5sec window with a 100ms shift. New features, called spectral slopes described in sub-section 3.2.1.4, are concatenated leading to a feature vector of 138 components. A dataset reduction algorithm based on PSVMs [TMN07] is applied to cope with the enormous amount of data available for training.

5.3.4 Ring tone detector

Àgora database also contains such non-speech events as ring tones that usually precede the telephone conversation. In fact, all these ring tones have similar acoustical content and duration. An ad-hoc approach has been developed to detect them that consists in the following steps:

- One of the ring tones is manually isolated and saved as a waveform.

For each show:

- The input audio stream is normalized to its maximum.
- The cross-correlation $r[n]$ between the show and the isolated ring tone is computed.
- $r[n]$ is normalized to its maximum.
- A new signal is generated with pulses where $r[n] > 0.5$. The smoothing is introduced to avoid rapid changes of this signal.
- The center of each pulse is considered as the middle point of the detection.
- Taking into account the duration of the ring tone and the middle point of the detection the output hypothesis is generated.

5.3.5 Experimental results

In order to evaluate the proposed AS system the Àgora broadcast news database has been divided into three sets: training, development and evaluation. The sets have been designed to have similar acoustic characteristics. This way 27h hours were used for training, 8h of audio were used for development and the remaining 8h for evaluation.

We evaluate the improvements introduced by the hierarchical architecture by means of comparing two systems: one-step and hierarchical. In one-step system each acoustic class is modelled by HMM with 2 emitting states with 5 Gaussians per state. 16 FF LFBE coefficients with their first time derivatives together with 18 spectral slope features are used. In total, the feature vector has 50 components.

Table 5.3.1. Segmentation results per class

Class	One-step (%)	Hierarchical (%)
Speech	6.5	4.8
Music	2.5	2.4
Speech over music	5.8	4.9
Telephone speech over music	3.7	1.5
Average	**4.6**	**3.4**

From the results in Table 5.3.1, it can be observed that the use of a more flexible architecture allows developing a system that is more suited to a particular task. 26% of relative improvement in average can be obtained by using a set of detectors, which are properly combined and also tuned to the different target classes.

In table 5.3.2 we present the improvement of segmentation results for "telephone speech over music" class introduced by the proposed spectral slope features. As we can see, the inclusion of those features yield strong recognition improvement.

Table 5.3.2. "Telephone speech over music" detection results

System	Error rate (%)
without Spectral Slope	3.5
with Spectral Slope	1.5

The ring-tone detection results are presented in Table 5.3.3. We observe that the proposed ad-hoc approach demonstrate almost perfect ring-tone detection. The main source of the errors corresponds to the short mismatches at the boundaries of the detected signals.

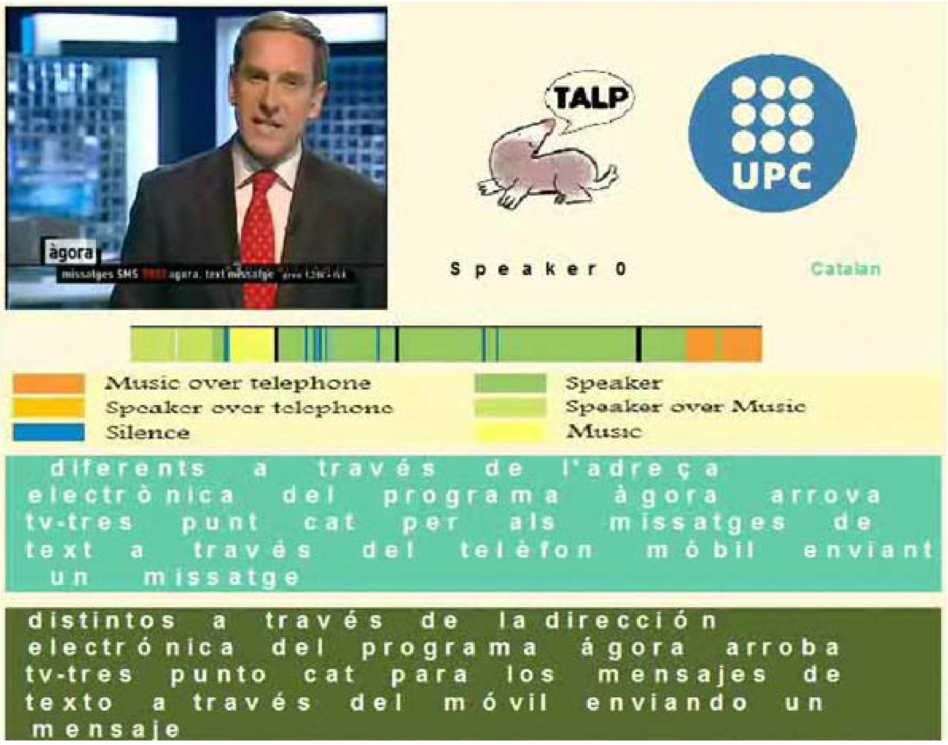
Table 5.3.3. Ring tone detection results

Scored time	49.48 sec
False alarm time	0.17 %
Missed time	2.12 %
Error rate	**2.29 %**

5.3.6 Tecnoparla application

Within the Tecnoparla research project a specific application that demonstrates the usefulness of speech and language technologies has been developed. This application performs automatic translation and subtitling of audiovisual content (Figure 5.3.4). The interest of this application is clear since a large variety of audiovisual content is available in multiple languages, either professionally created (television or radio) or created by individuals and available through the Internet. Thus the automatic translation and subtitling facilitates accessibility and dissemination of international audiovisual content.

The developed system has been designed specifically for television news or debate, and includes many speech and language technologies: audio segmentation, speaker diarization, language recognition, speech recognition, speech translation and text-to-speech synthesis. The first step in this system is audio segmentation, described in previous sub-Section. Since speech recognizer must act only when there is speech and there is no background music or it is barely perceptible, this module avoids absurd transcripts with large amount of errors. As demonstrated in Figure 5.3.4, the developed application automatically segments audio into the acoustic classes presented in Table 5.2.1.



Speaker 0

Catalan

🟧 Music over telephone
🟨 Speaker over telephone
🟦 Silence
🟩 Speaker
🟩 Speaker over Music
🟨 Music

diferents a través de l'adreça electrònica del programa àgora arrova tv-tres punt cat per als missatges de text a través del telèfon mòbil enviant un missatge

distintos a través de la dirección electrónica del programa ágora arroba tv-tres punto cat para los mensajes de texto a través del móvil enviando un mensaje

Figure 5.3.4. Automatic translation and subtitling system

5.4 Hierarchical audio segmentation system for tv3/24 database

The flow diagram of the hierarchical AS system for tv3/24 database is presented in Figure 5.4.1. The whole detection system consists of 5 binary detectors. A set of audio spectro-temporal features is extracted to describe every audio frame. It consists of 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives. In total, a 32-dimensional feature vector was initially chosen. Each binary detector is trained using the features which were selected from that 32-dimensional feature vector during the fast feature selection procedure (described in chapter 6). Each binary detector consists of 2 HMMs: "Class" and "non-Class". Both HMMs have 1 emitting state and the observation distributions are Gaussian mixtures with continuous densities. There are 16 sessions available for designing the audio segmentation system. Half of the sessions we decided to use for training/development and the other half for testing.

First we select the appropriate number of Gaussians per HMM-GMM for each binary detector. Actually, this number is a trade-off between the improvement in performance and the execution time needed to train the models with corresponding number of Gaussians. With 256 Gaussians we got the acceptable results. Note, for modelling of ACs in tv3/24 database the number of Gaussian mixtures is much higher than in the case of Àgora database due to the more complex audio content. Figure 5.4.2 demonstrates the mean error-rate obtained with increasing of the number of Gaussian mixtures per model. The average error rate was computed by means of averaging the error rates from each binary detector without combination them in hierarchical structure.

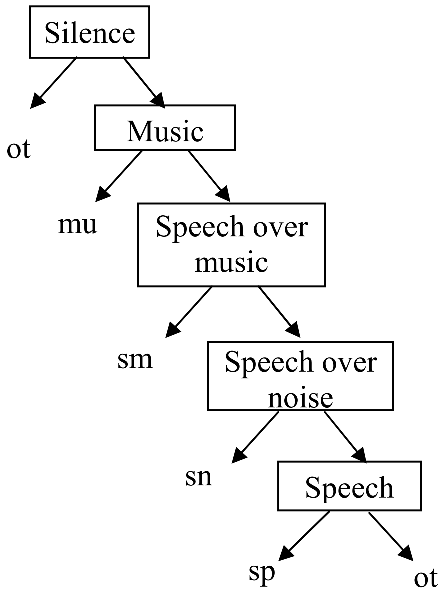


Figure 5.4.1. Flow diagram of the hierarchical architecture for tv3/24 database

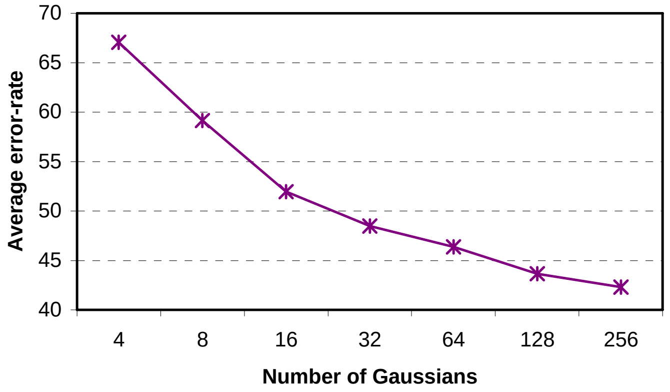


Figure 5.4.2. Relation between mean error-rate and the number of mixtures per each GMM model

In Figure 5.4.3 we compare different system architectures. The one-step multi-class system corresponds to the HMM-GMM audio segmentation performed in one step. The "Hierarchical" architecture is the one depicted in Figure 5.4.1. Finally, the system "Hierarchical + FS" is the same as previous but uses the feature selection described in chapter 6.

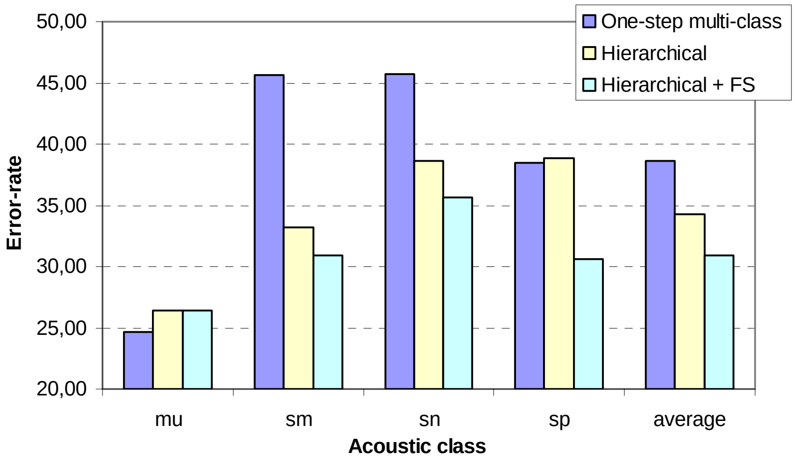


Figure 5.4.3. Comparison of different detection systems

According to results from Figure 5.4.3, the hierarchical audio segmentation system outperforms the one-step multi-class detection system (about 8% of absolute error-rate reduction in average).

The CPU time employed to perform testing is described below:

- Feature extraction: 546 sec;
- Viterbi segmentation: 3329 sec;
- Total: 3845 sec.

This processes were executed on PC with Intel Core 2 CPU, 2.13 GHz, 1Gb of RAM. This way the total CPU time, computed by adding CPU times for feature extraction and audio segmentation, falls below 1×RT (real-time factor).

5.5 Improving speaker diarization using audio segmentation[1]

5.5.1 Task definition

In the context of human language processing speaker diarization consists in segmenting and labelling an unknown set of speakers in a continuous audio stream. Speaker diarization is usually described as the task of deciding who spokes when and it can be used in large variety of applications.

The main objective of this Section is to evaluate and compare the performance of the diarization system described in [LAT08] exploiting the audio segmentation information in different ways. On the one hand, it can be used beforehand to extract speech or more condition-specific segments which are then fed to the diarization system. On the other hand, the audio segmentation hypothesis can be used with the diarization labelling to perform time masking. In our experiments audio segmentation is performed using hierarchical AS system described in Section 5.3.

5.5.2 Speaker diarization performance bounds

The results for speaker diarization experiments were obtained on two data sets: development (Devset) and evaluation set (Evalset) of Àgora database. In the first experiment we perform speaker diarization without any audio segmentation information. Thus all the audio data, silences and music included, is considered for speaker labelling. The idea is to define the lower performance bound of the system.

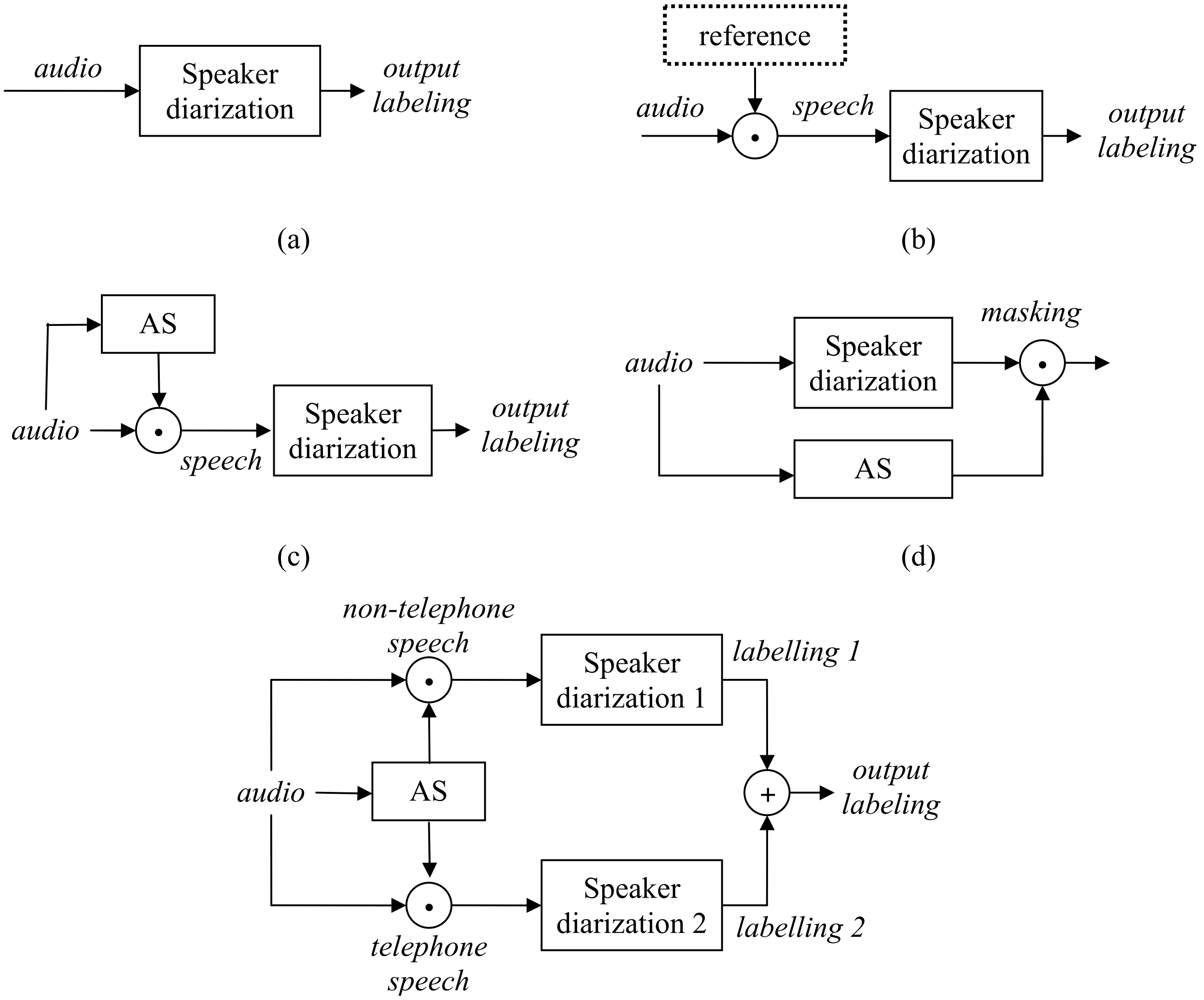
Opposite to this, we can also define the upper performance bound by using a perfect audio segmentation. A perfect segmentation is achieved by extracting speech segments according to the reference transcription. Here, the entire diarization error-rate (DER) is caused either by incorrect speaker clustering or by missed speech due to overlapped speech of multiple speakers, since speaker diarization system is assigning only one label for a segment.

The difference between the upper and lower limit corresponds to the importance of applying an audio segmentation for speech extraction. The experimental setup schemes for these two experiments are depicted in Figure 5.5.1 (a) and (b). From the numbers in Table 5.5.1 (columns “No AS” and “Perfect AS”) it is obvious that in the perfect case we can gain for development and evaluation set 1.98% and 3.00% absolute DER difference, respectively.

[1] We are grateful to Martin Zelenak for performing the experiments and system description in this sub-Section

Table5.5.1. Speaker diarization experiment results: (No AS) without any audio segmentation; (Perfect AS) speech extracted according to the reference; (SpeechExt) speech extracted according to segmentation hypothesis; (OutMask) all audio data is given to diarization and speaker labeling is masked with speech segments; (Tel + Non-Tel)

	Diarization error rate (DER), %				
	No AS	**Perfect AS**	**AS hyp. SpeechExt**	**AS hyp. OutMask**	**Tel + Non-Tel**
Devset	15.69	13.71	14.48	14.41	13.74
Evalset	13.50	10.50	12.15	12.43	12.37



(e)

Figure 5.5.1. Experiment strategies: (a) all audio is fed to the diarization system; (b) diarization over speech segments extracted according to the reference transcriptions; (c) diarization over speech segments extracted according to audio segmentation hypothesis; (d) diarization output speech masked with audio segmentation hypothesis; (e) separate diarization of telephone and non-telephone channel speech and merging of the two labelings

5.5.3 Speaker Diarization using Audio Segmentation

Speaker diarization system aims to find speaker changes and assign cluster labels to it. The audio segmentation hypothesis assists the diarization process by localizing applicable data in order to prevent labelling of non-speech segments. One approach is to extract speech only before providing the data to diarization (Figure 5.6.1 (c)) and the other is to perform a post-processing of the speaker transcription so that non-applicable time segments are discarded (Figure 5.6.1 (d)). The latter case is referred as output masking. It needs to be emphasized that the diarization labelling which is masked is obtained for the whole audio stream (including e.g. silences). The comparison of these two approaches unveils the influence of cluster purity on the performance. The difference in DER for evaluation set, as can be seen from the 3rd and 4th columns in Table 5.5.1, is not more than 2.30% relatively and for the development set the masking approach is even slightly better than the extraction approach.

After analyzing the erroneous segments, it was found that for telephone channel speakers the diarization system usually creates just one cluster with no respect to the actual number of speakers in such portions. To cope with this problem, a more tailored diarization for the telephone channel audio was applied. Audio segmentation information was used to distinguish between telephone and non-telephone speech. The structure of the TV shows guarantees that the identity of speakers in studio is different to those who are calling by telephone. This diarization strategy is schematically illustrated in Figure 5.6.1 (e). The speech data is split into two sets and separate diarization is performed for both of them in parallel. Diarization performance of this approach is presented in the last column of Table 5.5.1. The difference in DER for development and evaluation set is 1.95% and 1.13%, respectively.

5.6 Albayzín-2010 evaluation campaign

Taking into account the increasing interest in the problem of audio segmentation from the one hand, and the existence, from the other hand, of a rich variety of feature extraction approaches and classification methods, we organized an international evaluation of broadcast news audio segmentation in the context of the Albayzín-2010 campaign. The Albayzín evaluation campaign is an internationally-open set of evaluations organized by the Spanish Network of Speech Technologies (RTH) every 2 years. Actually, the quantitative comparison and evaluation of competing approaches is very important in nearly every research and engineering problem. The evaluation campaigns that independently compare systems from different research groups help to determine which directions are promising and which are not [PAL03].

The rest of this section is organized as followed: first, we describe the different feature extraction methods, the segmentation techniques, and the organization ways of the segmentation process proposed by the eight groups that submitted their results to the evaluation. Second, we compare the various segmentation systems and results, to gain an insight into the proposed solutions.

5.6.1 Participating groups and methods

Ten research groups registered for participation, but only eight submitted segmentation results: *ATVS* (Universidad Autónoma de Madrid), *CEPHIS* (Universitat Autònoma de Barcelona), *GSI* (Instituto de Telecomunicações, Universidade de Coimbra, Portugal), *GTC-VIVOLAB* (Universidad de Zaragoza), *GTH* (Universidad Politécnica de Madrid / Universidad Carlos III de Madrid), *GTM* (Universidade de Vigo), *GTTS* (Universidad del País Basco), *TALP* (Universitat Politècnica de Catalunya).

About 3 months were given to all the participants to design their own audio segmentation system. After that period, the testing data were released, and 2 weeks were given to perform testing.

In the following, the systems presented by the participant groups are briefly described. The systems are listed in the order they are ranked in the table of final results. The full description of the systems can be found in FALA 2010 conference proceedings [FAL10].

System 1

Features: segment-based. First, 15 MFCCs, the frame energy and their first and second derivatives (delta and delta-delta) are extracted. Additionally, the spectral entropy and the chroma coefficients are calculated. Second, the mean and variance of these features are computed over 1 sec interval.

Segmentation approach: HMM-based.

The acoustic modeling is performed using 5 HMMs with 3 emitting states and 256 Gaussians per state. Each HMM corresponds to one acoustic class. An 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".

System 2

Features: segment-based. First, 13 MFCCs including the zero (energy) coefficient and their first and second derivatives (delta and delta-delta) are extracted. Second, a background model based on GMM (GMM-UBM) of M mixture components is trained using data from all classes. Then, given an audio segment represented by N feature vectors of dimension D, the GMM-UBM is adapted to that audio segment using MAP adaptation. By stacking the resulting means, a supervector of dimension $M{\cdot}D$ is obtained.

Segmentation approach: detection-and-classification.

The BIC algorithm is used to detect the segment boundaries. The classification of each segment is performed using Support Vector Machines (SVMs).

System 3

Features: frame-based 7 MFCCs plus shifted delta coefficients (SDC).

Segmentation approach: HMM-based.

The acoustic modeling is performed using a five-state HMM with full connected state transitions. Each state corresponds to one acoustic class modeled by GMM with 1024 mixtures. Given a vector of observations, the Viterbi decoding algorithm is applied to obtain a sequence of HMM states. A *mode* filter (i.e. a filter that replaces a current state with *mode* of its neighboring states) is applied to avoid spurious changes between states.

System 4

Features: frame-based 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives. Mean subtraction is applied at the segment level. A wrapper-based feature selection technique is used to find the most discriminative features for each acoustic class individually.

Segmentation approach: HMM-based.

The acoustic modeling is performed using 5 HMMs with 1 emitting state and 64 Gaussians per state. Each HMM corresponds to one acoustic class. A hierarchical organization of binary HMM detectors is used. First, the audio stream is pre-segmented using a silence detector. Then non-silence portions are segmented into "Music"/"non-Music"; the "non-Music" portions are further segmented into "Speech over music"/"non-Speech over music"; the "non-Speech over music" portions are further segmented into "Speech over noise"/ "non-Speech over noise"; and, finally, the "non-Speech over noise" portions are segmented into "Speech"/"Other".

System 5

Features: frame-based 12 PLPs plus local energy and their first and second derivatives (delta and delta-delta).

Segmentation approach: HMM-based.

The acoustic modeling is performed using 5 HMMs with 1 emitting state and 64 Gaussians per state. Each HMM corresponds to one acoustic class.

System 6

Features: frame-based 16 MFCCs including zero (energy) coefficient, plus 8 perceptual coefficients (e.g. zero crossing rate, spectral centroid, spectral roll-off, etc.) and their first time derivatives.

Segmentation approach: mixed, detection-by-classification and HMM-based.

An hierarchical organization of the detection process is used. First, silence and music are located using a repetition detector system based on fingerprinting (detection-by-classification). In the proposed fingerprinting system, a 32-bit binary pattern is computed for each frame of about 200ms; spectral analysis is performed with a mel-scaled filter-bank with 32 channels, and the resulting spectrogram is binarized into a 32-bit pattern, choosing 1, essentially, when there is a spectral peak. The detection strategy consists in counting the number of matching bits between the signature and the audio binary patterns in each frame, and when this number is above a threshold, an acoustic class is detected. Second, a hybrid HMM-MLP segmentation is applied to the audio segments which are not classified as either music or silence. Each acoustic class is modeled via a 10-state HMM with left-to-right state transitions.

System 7

Features: frame-based 13 MFCCs plus their first and second derivatives (delta and delta-delta). Additionally, the mean, the variance and the skewness of the first MFCC are calculated.

Segmentation approach: detection-and-classification.

The BIC algorithm is used to detect the segment boundaries. Classification is performed with a hierarchical organization of detectors and using GMMs combined with a binary decision tree. First, the audio stream, which is pre-segmented with a silence detector, is classified into "Music"/"non-Music" segments; the "non-Music" ones are further classified into "Speech over music"/"Speech"/ "Speech over noise".

System 8

Features: frame-based 13 MFCCs including zero (energy) coefficient. Cepstral mean subtraction was not applied.

Segmentation approach: detection-by-classification.

Each class is modeled by a GMM with 1024 mixtures. For each frame, the class yielding the highest likelihood is chosen. A *mode* filter is applied to smooth the decisions along time.

5.6.2 Results

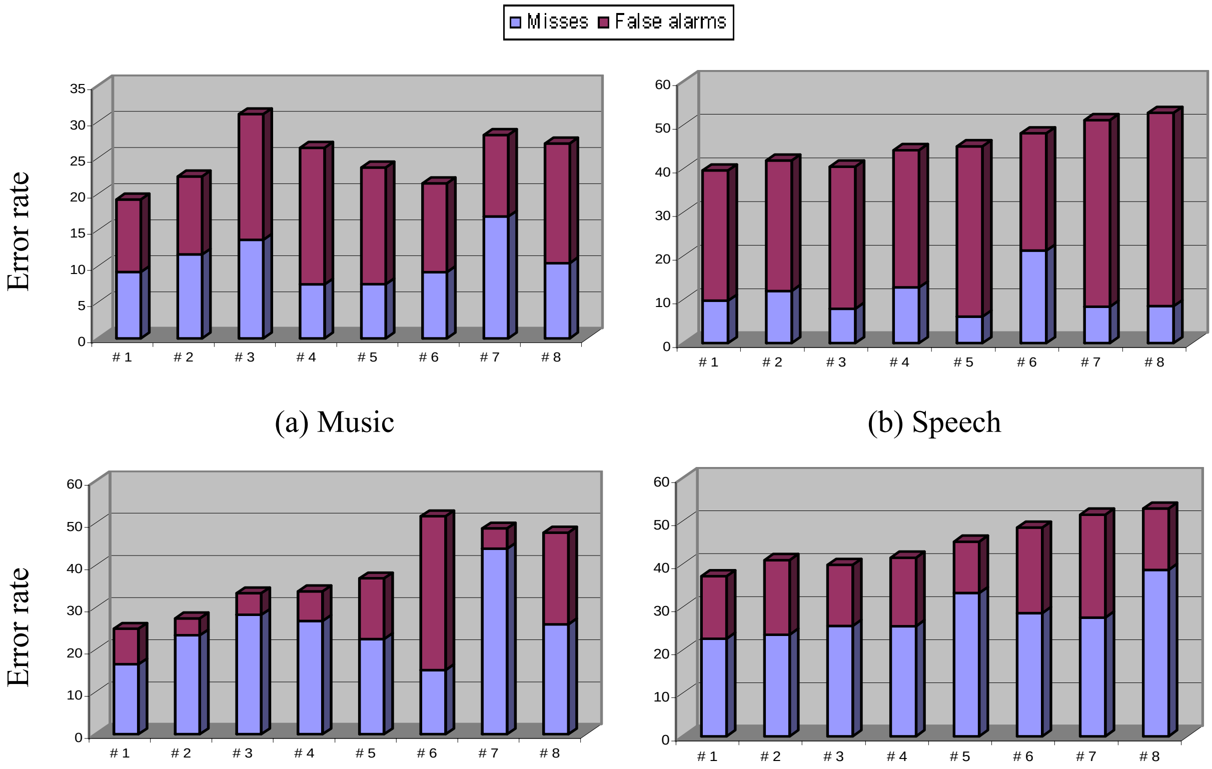
Table 5.6.1 presents the final scores from the eight systems. The error rate is presented for each evaluated class individually, together with the average score over all evaluated classes. Note, that no participant was using any additional data for training the acoustic models apart from the data provided for the evaluation.

As can be observed in Table 5.6.1, "Music" is the best detected class for all the systems. The system that obtained the best average score (30.22%), system 1, also got the highest score individually for each class.

The distribution of the miss and the false alarm errors from all systems is presented in Figure 5.6.1. This plot shows a clear unbalance between misses and false alarms for the classes "Speech" and "Speech over music".

Table 5.6.1. Results of the audio segmentation evaluation

systems	Error rate				
	mu	sp	sm	sn	**Average**
1	19.21	39.52	24.97	37.19	**30.22**
2	22.41	41.80	27.47	40.93	**33.15**
3	31.01	40.42	33.39	39.80	**36.15**
4	26.40	44.20	33.88	41.52	**36.50**
5	23.65	45.07	36.95	45.21	**37.72**
6	21.43	48.03	51.66	48.49	**42.40**
7	28.14	51.06	48.78	51.51	**44.87**
8	26.94	52.76	47.75	52.93	**45.09**



(a) Music

(b) Speech

(c) Speech over music

(d) Speech over noise

Figure 5.6.1. Distribution of errors across the eight systems and for each acoustic class

Table 5.6.2. Confusion matrix of acoustic classes

	mu	sp	sm	sn
mu	89.4	0.1	8.0	2.5
sp	0.0	70.6	2.9	26.5
sm	1.8	1.2	87.0	10.0
sn	0.3	10.2	8.3	81.2

In Table 5.6.2 we present the confusion matrix, which shows the percentage of hypothesized acoustic classes (rows) that are associated to the reference acoustic classes (columns). Data represent averages across the eight audio segmentation systems.

According to the confusion matrix, the most common errors are the confusions between "Music" and "Speech over music", between "Speech over music" and "Speech over noise", and also between "Speech" and "Speech over noise". Indeed, the two components of each of those pairs of classes have very similar acoustic content. Another interesting observation is the low proportion (almost 0%) of confusions between "Speech" and "Music". The second row of the confusion matrix indicates that 26.5% of the hypothesized speech is in fact "Speech over noise". This is the main reason of the high proportion of false alarms for the class "Speech" (Figure 5.7.1 (b)). Actually, for many "Speech over noise" audio segments the level of noise in background is extremely low so that the detection systems usually confuse "Speech over noise" with "Speech".

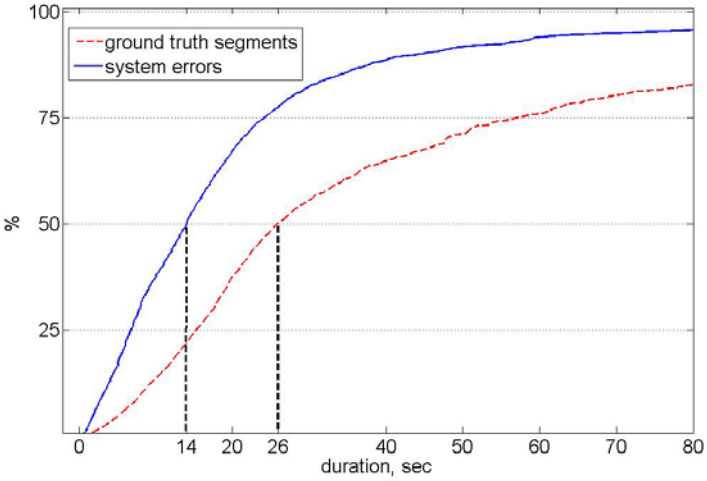
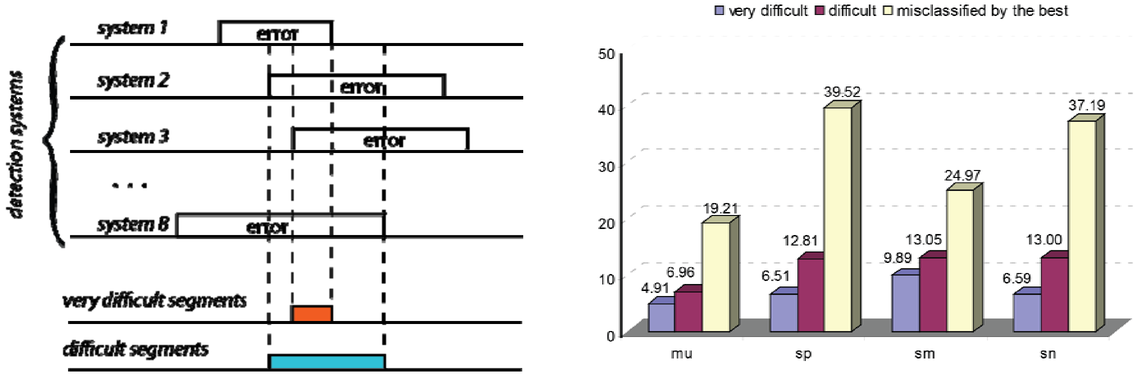


Figure 5.6.2. Cumulative distribution of segments in terms of duration

In Figure 5.6.2 we present cumulative distributions of duration of testing segments. The solid curve corresponds to the segments incorrectly detected by the audio segmentation systems for the whole set of participants. The dashed curve corresponds to the cumulative distribution of the ground truth segments. Each point (*x, y*) of this plot shows the percentage *y* of segments with duration less than *x* seconds.

According to this plot, more than 50% of the total amount of errors is shorter than 14 sec. For comparison, according to the ground truth labels, 50% of audio is represented by segments of duration less than 26 sec. So, in average, the duration of erroneous segments is almost twice shorter than that of the ground truth segments.

In Figure 5.6.3 we compare the error distribution for 3 types of segments in the testing database: *very difficult*, *difficult* and *misclassified by the best*. As illustrated in Figure 5.6.3 (a), *very difficult* are those segments which are totally included in error segments from 8 systems. *Difficult* segments are those which are included in error segments from at least 7 systems. Finally, *misclassified by the best* are those segments where the winner system in evaluation produced errors. The graphical distribution of those 3 types of segments is displayed in Figure 5.6.3 (b).



(a) (b)

Figure 5.6.3. (a) Illustration of "difficult" and "very difficult" segments; (b) Error distribution of "difficult", "very difficult", and "misclassified by the best" segments

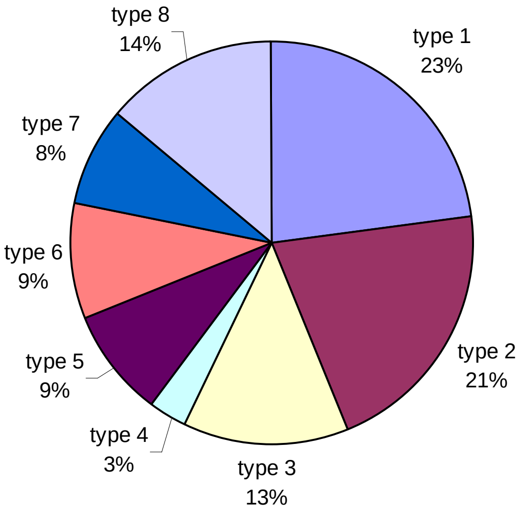


Figure 5.6.4. Percentages of distribution of the different types of shared errors

Table 5.6.3. Description of the different types of shared errors

Type of error	*Description*
1	Low level of background sound
2	Speech in background
3	The quality of music in background is low
4	Singing in background
5	Noise in background is more dominant than music for the [sm] class
6	The microphone is affected by the wind
7	Annotation mistake
8	Other

The error distribution for those segments, displayed in Figure 5.6.3, shows the degree of difficulty of the audio segmentation task. In average, only 6.98% of the segments in the testing database are *very difficult*. The rest of the segments were detected correctly at least by one detection system. Comparing this number with the final score from the winner system (30.22%) we conclude that there is still a large margin to improve the audio segmentation performance.

Table 5.6.3 shows a grouping of the errors which are shared by all the 8 segmentation systems. The groups were defined after listening to all the segments which are defined as *very difficult* and are longer than 5 seconds. Seven different types of error were distinguished, and the rest were included in *Other*.

According to the plot in Figure 5.6.4, a large percentage of shared errors was provoked by the presence of either a low level of sound in the background (23%) or overlapped speech (21%), while the annotator mistakes caused only 8% of the total amount of shared errors.

5.6.3 Discussion

By analyzing both the submitted audio segmentation systems and the corresponding segmentation results, several observations can be extracted which are outlined in the following.

1. The conventional use of automatic speech recognition features for the audio segmentation task

Historically, there have been no features specifically designed for the AS task. In the current evaluation, all systems used features that were designed for the automatic speech recognition (ASR) task, like MFCC, PLP or FF. Some systems combined the ASR features with other perceptual feature sets, but they could not report any significant improvement (for details we refer the reader to [FAL10]).

2. The systems that used segment-based features outperformed the systems with frame-based features

The best two audio segmentation systems parameterized the audio signal using segment-based features. The system 1 used the mean and variance along 1 sec segments; the system 2 used a super-vector approach to parameterize even longer segments. Note the third best system used SDC coefficients, which take into account a long audio context. Presumably, this is the main reason for their superior detection rates. It may indicate that the models trained on frame-based features do not capture the structure of the acoustic classes sufficiently.

3. The majority of the audio segmentation systems used the HMM approach

The main advantage of the HMM approach is that it performs segmentation and classification jointly. Other alternatives like *detection-and-classification* or *detection-by-classification* require two independent steps to be carried out one after the other, so that the errors produced in the first step may propagate to the next one. Additionally, more parameters for tuning are required, which makes the system task-dependent.

4. The hierarchical detection approach seems to be effective

Four research groups reported an improvement when using a hierarchical organization of the detection process. One of the most important decisions when using this kind of architecture lies in the orderings of the detection modules, since some of them may benefit greatly from the previous detection of certain classes. Those four audio segmentation systems detect the easiest classes ("Music" and silence, which is included in "Other") at the early steps, while a further discrimination among the rest of the classes is done on subsequent steps. In this type of architecture, it is not necessary to have the same classifier, feature set and/or topology for the various individual detectors.

5. The fingerprinting approach for music detection seems to be effective

Finding of repetitions with fingerprinting seems to be useful in audio segmentation of broadcast news due to the omnipresence of advertisements, jingles and even repeated programs. The system 6, which used that approach, got the second best result for the class "Music".

6. Challenge of the audio segmentation task

Only 6.98% of the audio segments were detected incorrectly by all the audio segmentation systems. The rest of audio was recognized correctly by at least one detection system. Comparing this number with the score obtained by the winner system (30.22%), we conclude that there is still a large margin for improvement of segmentation results. Taking into account that the main source of mistakes are confusions between "Music" and "Speech over music", between "Speech over music" and "Speech over noise", and also between "Speech" and "Speech over noise", future research efforts should be devoted to improved detection of background sounds.

7. Complementarity of different segmentation systems

The segmentation results from different systems are complementary up to some extend, so that the combination of them yields improvement in accuracy. A simple majority voting fusion scheme of the best three systems reduces the average score to **28.60%**, and the fusion of the best five systems, to **29.19%**. Comparing these numbers with the score obtained by the winner system (**30.22%**), we conclude that post-processing of the segmentation results from different segmentation systems is beneficial.

8. Applicability of the systems to work in real-time

Unlike many speech recognition or speaker diarization systems, whose performance drops drastically when operate in real-time, the described audio segmentation systems can work in real-time due to their relative simplicity. In fact, four participants reported timing results (systems 3, 4, 5, and 8) and the total CPU time, computed by adding CPU times for feature extraction and audio segmentation, falls below 1×RT (real-time factor).

5.7 Constructing a reference AS system

Inspired by the systems from the Albayzín-2010 evaluation and their results, we constructed a reference AS system that combines the best characteristics from those submitted systems. We use the HMMs-GMM with 1 emitting state and 256 GMM mixture components to model each individual class. The one-against-all detection strategy is employed: 5 binary detectors are organized in a hierarchical way, as depicted in Figure 5.4.1.

The following sets of features were considered:

- ASR features. 16 FF coefficients with the first time derivatives are computed in frames 30ms long with 20 shift. Then mean and variance are computed over a 1 sec window.
- 12 chroma features are computed in frames 50ms long. Then mean and variance are computed over a 1 sec window.
- 5 energy statistics features are computed over a 10 sec window with 1 sec shift. Those features, which are obtained from the amplitude histogram of the audio signal, were not used in the evaluation campaign by any of the participants. But, since we observed a high amount of confusions between the classes [sp] and [sn], we conjectured that they may improve the overall accuracy.

Using different combinations of the proposed features we found the best feature set for each detector separately. The segmentation results are presented in Table 5.7.1.

Given the results from Table 5.7.1, we selected *FF LFBE* features for detecting [sp], *FF + Chroma* features for detecting [sn], and, finally, *FF + Statistical + Chroma* features for detecting both [mu] and [sm] classes. In average, we obtained 29.21% of error-rate that is the best score among presented in Albayzìn-2010 AS evaluation.

Table 5.7.1. Segmentation results from different binary detectors using different feature sets

	Error rate			
	mu	**sp**	**sm**	**sn**
FF	17.65	**39.67**	30.71	41.93
FF + Chroma	18.34	40.52	28.35	**40.77**
FF + Stat	18.51	41.81	26.80	43.14
FF + Stat + Chroma	**17.52**	45.26	**23.00**	44.00

	Error rate				
	mu	**sp**	**sm**	**sn**	**Average**
Combined hierarchically	17.70	37.42	22.80	38.93	**29.21**

5.8 Changes in the definition and scoring of the acoustic classes

In this section we propose several changes to the definition and evaluation of the acoustic classes in tv3/24 database and see how the proposed changes affect the recognition results. The transcription of the database used for evaluations was performed according to TC-STAR European Parliament Plenary Session Transcription Guidelines [TCS]. Apart from speech transcription, the annotations include 3 different layers:

Speaker turn layer: None, Studio speaker, Outdoor speaker.

Background conditions: none, music, noise, speech, speech + music, speech + noise, noise + music, speech + noise + music.

Non-speech AEs: any short time non-speech sound like laugh, throat, knocking, etc.

Since the non-speech AEs affect just short portions of audio, the corresponding layer was discarded from the definition of the acoustic classes.

Table 5.8.1. Distribution of the acoustic classes in the database

			Speaker conditions			
			None	Studio	Outdoor	
Background conditions	None (clean conditions)	**OT**	3.85/ -	18.18/ **7.56**	18.55/ **26.34**	**SP**
	Music	**MU**	5.70/ **3.30**	11.21/ **9.99**	3.30/ **4.43**	**SM**
	Music+ Speech		0.68/ **0.00**	0.17/ **0.04**	0.30/ **0.62**	
	Music + Noise		0.05/ **0.03**	1.07/ **0.72**	0.36/ **0.60**	
	Music + speech + noise		0.00/ **0.00**	0.00/ **0.00**	0.00/ **0.00**	
	Speech	**OT**	0.37/ **0.01**	1.27/ **1.99**	2.12/ **3.60**	**SN**
	Noise		0.01/ **0.02**	10.03/ **17.60**	21.90/ **22.93**	
	Noise + speech		0.01/ **0.00**	0.20/ **0.01**	0.66/ **0.19**	

The distribution of the ACs in the database is presented in Table 5.8.1. The first number in each cell corresponds to the percentage of the corresponding AC in the whole database, the second one shows the percentage of the errors that the corresponding class provokes in the testing part of the database (the reference AS system is used to compute the error distribution). In equilibrium state those two numbers should have similar values. In the case when the first number is higher than the second one, the AC is considered easy for detection, in the opposite case it is considered difficult. From the table we observe that the classes *clean outdoor speech*, *studio speech in noise* and also *studio/outdoor speech with speech in background* are difficult for detection. In the table we also show how the classes were grouped for the Albayzín-2010 evaluation.

In the following we propose several alternatives to the initial design of the acoustic classes and the way they are evaluated, and report how these changes affect the segmentation results.

Refinement 1:

Include the *speech with speech in background* segments into the [sp] class. In fact, many times the overlapped speech appears when there is a synchronous translation. Although the amount of overlapped speech is not high (1.27 + 2.12 = 3.39% of total amount of data), the proposed refinement may partially remove confusion errors between [sp] and [sn].

Refinement 2:

We propose to make some refinement in the evaluation of the classes [sp] and [sn]. In fact, in Table 5.8.1 we see a clear unbalance between the two numbers in the *studio speech in the noise* cell, and in the *clean outdoor speech* cell. Indeed, s*tudio speech in the noise* is acoustically similar to the [sp] class, and conversely, *clean outdoor speech* is similar to the [sn] class. For the segments that are labeled as *studio speech in the noise* and *clean outdoor speech* we propose to assume that both hypothesis labels [sp] and [sn] are correct.

Refinement 3:

In the Albayzín-2010 evaluation we considered all confusion errors are equally weighted in the metric. But, for instance, it seems reasonable to weight less the confusion between [mu] and [sm] than between [mu] and [sp].

Table 5.8.2. Weights for the different types of confusion errors

Reference	mu	sm	sn	sp
sp	1	0.5	0.5	0
sn	1	0.5	0	0.5
sm	0.5	0	0.5	0.5
mu	0	0.5	1	1

hypothesis

In principle, the idea is to penalize confusion errors between ACs that have similar acoustic content less than confusion errors between classes that have very different acoustic content. The set of proposed weights are displayed in Table 5.8.2.

Refinement 4:

Although the use of single layer segments is practically convenient, we could also define the task in terms of a multiple layer segmentation. For instance, we could define the task of segmenting audio into 3 possibly overlapped ACs: “Speech”, “Music” and “Noise”. In that case the classes are acoustically different and could mutually overlap so there is no need to apply refinements 2 and 3.

Table 5.8.3. Segmentation results with the proposed refinements

	mu	sp	sm	sn	Average
Baseline	17.70	37.42	22.80	38.93	**29.21**
Refinement 1	17.31	42.32	23.40	48.20	**32.82**
Refinement 2	17.31	34.27	23.40	31.84	**26.71**
Refinement 3	13.61	22.38	14.04	23.92	**18.49**

	Speech	Music	Noise	Average
Refinement 4	6.6	69.5	82.3	**52.8**

Table 5.8.3 shows the AS results with the proposed refinements. The main conclusions are:

1. Inclusion of *speech with speech in background* into the [sp] class increases the error rate of the reference AS system. Since in our database background speech is usually bubble noise, it is more appropriate to include the overlapped speech into [sn].

2. The two modifications related to the way of evaluating the ACs indeed decrease the error rate of the reference AS system. The main benefit from the proposed modifications is obtaining a more meaningful error rate measurement. For instance, while confusion errors were counted twice in the initial metric, one as deletion and the other as insertion, since the class “Other” [ot] is not evaluated in the final tests, the confusion errors with this class were counted just once for the remaining ACs. Therefore, there were confusion errors between semantically different ACs which were implicitly weighted less than other equally important errors. Conversely, with the proposed changes, we explicitly de-weight the confusion errors between semantically similar acoustic classes.

3. The definition of the three new ACs, “Speech”, “Music” and “Noise”, that can mutually overlap, to replace the five previously defined ACs, leads to very low recognition results. Presumably, the main reason for such behavior is the high proportion of audio segments that belong simultaneously to different ACs.

5.9 Chapter Summary

In this chapter we addressed the audio segmentation task in the broadcast news domain. Two different AS systems has been developed using two broadcast news databases: the first one includes audio recordings from TV debate program Àgora from the Catalan TV3 channel, and the second one includes audio from the 3/24 Catalan TV broadcast news channel. The output of the first AS system was used in automatic translation and subtitling application developed for the Tecnoparla project. Besides, it was used to improve the robustness of the speaker diarization system. The second HMM-based AS system with feature selection got competitive results in the Albayzín 2010 audio segmentation evaluation.

From the obtained results we observed that the use of a more flexible architecture allows developing a system that is more suited to a particular task. For Àgora and TV 3/24 databases, 26% and 9% of relative improvement, respectively, is obtained by using a set of detectors, which are properly combined and also tuned to the different target classes.

Taking into account the increasing interest in the problem of audio segmentation, the Albayzín-2010 evaluation of audio segmentation systems has been organized by our research group. The evaluation setup, definition of acoustic classes and the segmentation metric were proposed. After analyzing the submitted systems and their results, several main conclusions have been outlined, and also a reference audio segmentation system has been constructed that shows a superior recognition rate.

Chapter 6. Feature selection

6.1 Chapter Overview

In this chapter we aim to improve the detection rate by means of feature selection. In previous chapters we used a set of standard ASR features together with a set of "perceptual" features for AED. In order to enhance the detection of particular sounds, new features coming not only from the audio modality but also from video are proposed. Video features improved the detection of all acoustic events (AE) while the features coming from an acoustic localization system improved accuracy only for some of them. These results mean an additional motivation for us to perform feature selection in order to find the best feature set for each particular class of interest.

We compare three feature selection approaches: one is the purely wrapper sequential backward selection (SBS) algorithm and the other two are the hybrid two-stage systems that combine filter approaches (mRMR and PCA) with wrapper. We analyse the error through the lens of bias-variance decomposition. In order to reduce the computational complexity of the conventional wrapper approach, a fast one-pass-training technique is introduced that avoids retraining of acoustic models during the evaluation of the candidate feature sets.

The chapter is organized as follows: Section 6.2 describes motivation of feature selection. Sections 6.3 and 6.4 describe the feature selection approaches applied in the tasks of the meeting-room AED and the broadcast news AS. Section 6.5 presents the obtained experimental results, and, finally, Section 6.6 provides conclusions.

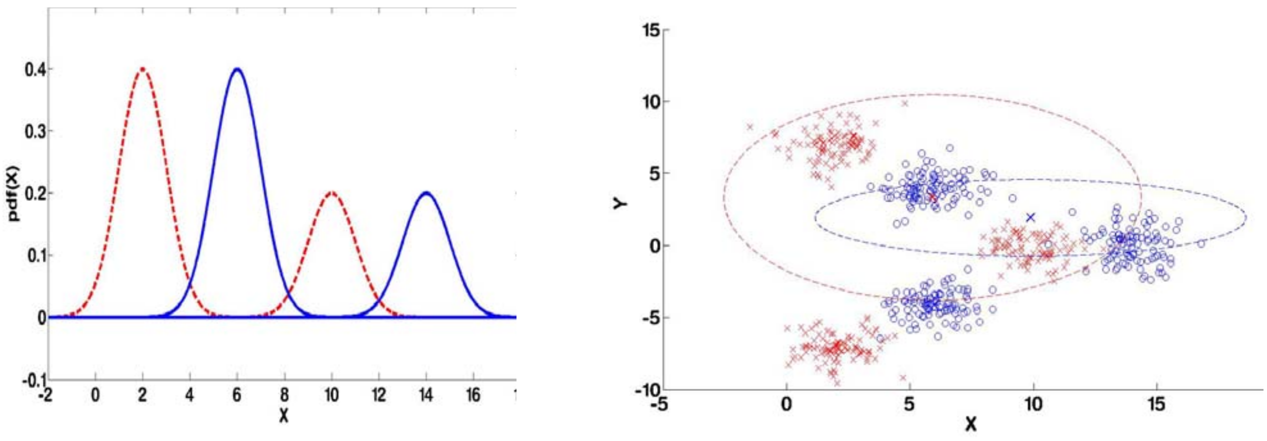
6.2 Motivation of feature selection

At least three reasons are traditionally given to motivate feature selection [GGN06]: performance improvement; general data reduction, to limit storage requirements and increase algorithm speed; and feature understanding, to gain knowledge about the data. The performance improvement is usually the primary objective in many applications and it is the main objective in the current work. Storage limitation objective is crucial, for instance, in embedded and online systems as the extraction of unimportant features may be computationally expensive and cause a time delay in classification algorithm.

In our work we aim to feature selection for two different audio recognition tasks within the widely used Gaussian Mixture Model (GMM) framework. The first one is AED in meeting-room environments. It is worth to mention that the system that got the highest accuracy in the last CLEAR 2007 [CLE07] evaluation used an automatic feature selection procedure to find the feature set that yields the highest detection rate [ZZL08]. The second task is the AS in the broadcast news domain. It has been already evaluated in the context of Albayzín-2010 international evaluation campaign [FAL10]. Noticeably, the system that got the highest score in Albayzín-2010 also performed a feature subset selection for each individual acoustic class [FAL10].

6.3 Feature selection approaches

There are many feature selection algorithms reported in the literature: some of them are effective, but very costly in computational time, and others are fast, but less effective in the feature selection task. We can distinguish two main approaches for feature selection. First, in the wrapper approach [KJ97], the actual classification algorithm is used to estimate the accuracy of feature subsets. The wrapper approach has proved to be effective but it is very slow to execute as the classification algorithm is called repeatedly. Second, the filter approach, approach, which operates independently of any classification algorithm: undesirable features are filtered out before classification starts. It is assumed that undesirable features are those that are "irrelevant", "noisy" and "redundant". However, from a theoretical point of view the increase of the number of features can never decrease the performance of the optimal (Bayes) classifier [KJ97]. This means that "irrelevant", "noisy" and "redundant" features are not necessarily harmful for a particular classifier and, conversely, potentially useful features might be harmful. Let's see a simple example of 2-class classification problem when potentially useful feature becomes harmful. The samples from both classes are generated from GMM with 3 mixture components and unit variance. The mean vectors of GMMs from the first class are: μ_1=(2, 6), μ_2=(2, -6), μ_3=(10, 0); and the mean vectors of GMMs from the second class are: δ_1=(6, 4), δ_2=(6, -4), δ_3=(14, 0). Imagine that we try to model each class using GMM with 2 components (assuming that the designer of the classification system did not select a correct number of Gaussians thus making the classifier far from optimal). In one-dimensional feature space (only one feature x is used) the classes a perfectly separated (Figure 6.3.1 (a)) but when additional feature y is added the overlap between class-conditional pdfs increases (Figure 6.3.1 (b)) thus the classification accuracy decreases. On the other hand, if the designer of the classification system selects too many Gaussian components to model the classes, the GMMs become too complex and do not generalize well. This particular example clearly shows how additional potentially useful feature y may harm the accuracy of the particular classification system. Based on this fact we are motivated to perform feature selection taking into account a particular classification algorithm (GMM-based in our case).



(a) (b)

Figure 6.3.1. (a) Pdf distribution of the two classes modeled by GMM with 2 mixture components and using only one feature x (b) The distribution of classes in 2-dimensional space (the dot ellipses show the lines of equal probability density of the Gaussians; unit variance is considered)

A search of the optimal feature set requires a state space, an initial state, a termination condition, and a search engine. The state space includes all possible combinations of features, and the search is terminated after finding a feature set with the lowest value of the evaluation (error) function *J*(.). When the initial number of features is equal to *d* the state space consists of 2^d feature subsets. In fact, it is not feasible to evaluate all possible feature combinations starting from even a low number of initial features *d*. Several feature selection approaches based on suboptimal search strategies are proposed and the ones that are used in our work are described in the following.

6.3.1 Sequential feature selection

Sequential forward selection (SFS) and sequential backward selection (SBS) are the greedy search strategies that are most popular in the wrapper-based feature selection approach. In the SFS method one starts with an empty set and progressively adds features yielding the maximum improvement of the evaluation function *J*(.). In the SBS method one starts with the full set of features and progressively eliminates the least useful ones. Both of these approaches are widely used when the initial feature set is relatively small (around 10-100 features) [GMT09]. Note that depending on the application domain, SFS and SBS approaches may lead to different results. In our work we use the SBS approach only.

6.3.2 mRMR feature selection

The minimal-redundancy-maximal-relevance (mRMR) [PLD05] is a filter approach that selects a given number of features which maximize the mutual information between the selected features and the class labels (maximal relevance) and simultaneously minimize the dependency among the selected features (minimal redundancy). In our work, this algorithm has been used to get a ranking of the features, from the most to the least significant one. Ranking of *d* features leads to *d* “nested” feature sets $S_1 \subset S_2 \subset ... \subset S_{d-1} \subset S_d$ where subset S_i is composed of the *i* most significant features according to the mRMR criterion. Using the evaluation function *J*(.) (the same one used in wrapper approach) we evaluate each feature subset S_i to choose the best one.

6.3.3 PCA feature selection

Methods that create new features based on transformations or combinations of the original feature set are called feature transformation or sometimes feature extraction algorithms. The transformed features may provide a better discriminative ability than the best subset of given features, but these new features (a linear or a nonlinear combination of given features) may not have a clear physical meaning.

The best known linear feature extractor is the principal component analysis (PCA) or Karhunen-Loève expansion, that computes the *m* largest eigenvectors of the *d x d* covariance matrix of the *n d*-dimensional patterns. The linear transformation is defined as:

$$Y = XH \qquad (6.3.1)$$

where *X* is the given *n x d* pattern matrix, *Y* is the derived *n x m* pattern matrix, and *H* is the *d x m* matrix of linear transformation whose columns are the eigenvectors. Since PCA uses the most expressive features (eigenvectors with the largest eigenvalues), it effectively approximates the data in a linear subspace. In Figure 6.3.2 we show the percentage of variance that is concentrated is the first most expressive features as a function of the number of features in the transformation domain. The number of features for classification stage is usually selected in such a way that 85-95% of variance is concentrated in first most expressive features. In our work given a feature ranking from the most expressive feature to the least one we create *d* “nested” feature sets $S_1 \subset S_2 \subset ... \subset S_{d-1} \subset S_d$ and then using the function *J*(.) we choose the best one.

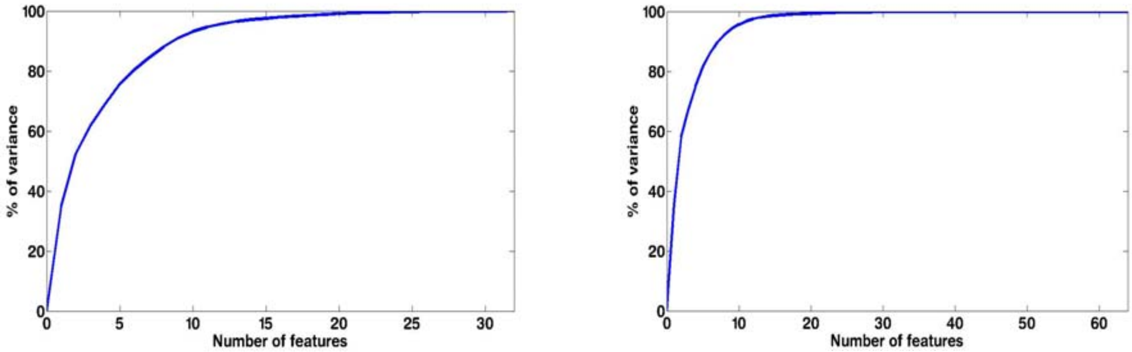


Figure 6.3.2. Percentage of variance that concentrated in the first most expressive features as a function of the number of features in the transformation domain. (a) Database of the meeting-room AEs (b) Broadcast news database

6.4 Bias-variance decomposition of the error

Two commonly used error functions *J*(.) are employed in this paper to evaluate the candidate feature sets: mean zero-one error (MZOE) and mean-square error (MSE). In the following we describe briefly these error functions and its decomposition into bias and variance terms. The bias-variance decomposition of the error allows to study the performance of the recognition system for different candidate feature sets.

Consider *n*-dimensional feature space R^n. Given a training set a classification algorithm produces a model *f*. Given a test example *x*, this model produces a prediction $y = f(x)$. Let *t* be a true value of the predicted variable for the test example *x*. A loss function $L(t; y)$ measures the cost of predicting *y* when the true value is *t*. Squared loss is defined as $L_{SL}(t, y) = (t - y)^2$ and zero-one loss $L_{0/1}(t, y) = 0$ if $y = t$, $L_{0/1}(t, y) = 1$ otherwise. In our framework, the goal of feature selection is stated as searching for feature set that produces a model with the smallest possible loss; i.e., a model that minimizes the average loss $L(t; y)$ over all examples.

Since the same classification algorithm will in general produce different models *f* for different training sets, $L(t; y)$ will be a function of the training set. Thus we are interested in expected loss $E_D[L(t; y)]$ over different training sets *D* of a given size. Bias-variance decompositions decompose the expected loss into the weighted sum of 3 terms: noise, bias, and variance, all non-negative. The noise term is the intrinsic error / uncertainty for correct prediction of *x*, regardless of the classification algorithm. Bias term measures how closely the classification algorithm average prediction (considering all possible training sets of a fixed size) matches the optimal prediction y_* (the Bayes rate prediction). In practice, we cannot know y_* for real data so we follow previous authors [MC09] in using $y_* = t$. As a result, the bias and noise cannot be separated and are combined in one term. Finally, the variance term shows how much the algorithm prediction fluctuates over different possible training sets of a given size. Domingos [Dom00] suggests the following decomposition for expected loss $\overline{L}_{SL}$ and $\overline{L}_{0/1}$:

$$
\begin{aligned}
\overline{L}_{SL} &= E_D[L_{SL}(t; y)] = B(x) + V(x) \\
\overline{L}_{0/1} &= E_D[L_{0/1}(t; y)] = B(x) + kV(x)
\end{aligned}
\qquad (6.4.1)
$$

where $B(X) = L(y_m; t)]$, $V(X) = E_D[L(y_m; y)]$ and y_m is the main prediction, i.e. the one that minimizes $E_D[L(y, y_m)]$. For squared loss y_m is the mean prediction of the classification across possible training data sets, and for zero-one loss y_m is the *mode* (the most frequent value) of predictions; and the constant k = 1 if $B(x)$=0 and k = -1 if $B(x)$=1. The error functions MSE and the MZOE are computed by averaging over multiple test examples the $\overline{L}_{SL}$ and the $\overline{L}_{0/1}$, respectively.

6.5 Feature selection for AED

In Section 6.3 we presented three search strategies to perform feature selection. There are also two alternative error functions for the evaluation of the candidate feature sets described in Section 6.4: MSE and MZOE. Usually both of them show similar results [Dom00] and could be used interchangeably. Note MZOE is the percentage of predictions that do not predict the correct class. It is often simply called the error rate for a classification model and it is tightly related to the evaluation metric in the AS task.

The process of computing MZOE for broadcast news AS task is described in the following. Given a feature set Ω*,* the "Class" and "non-Class" models (Θ_{class} and $\Theta_{non\text{-}class}$) are obtained using training data. For each audio segment *i* (we fixed the maximum length to 20 sec) in the development database the problem of detecting "Class" in that interval *i* with observation sequence X_i can be formulated as a hypothesis test, being H_0 "Class" in the interval *i* is not detected, and H_1 "Class" in this interval is detected. The necessary condition for detecting "Class" (hypothesis H_1) is the log-likelihood ratio (LLR) ΔL_i of "Class" and "non-Class" models exceeds a given non-negative threshold value *P,* i.e.

$$\Delta L_i = Q(X_i|\Theta_{class}) - Q(X_i|\Theta_{non-class}) > P \qquad (6.5.1)$$

where $Q(X|\Theta)$ denotes the log-likelihood function given a feature vector *X* and an acoustic model Θ. The predicted label *y* for the *i*th segment is considered 1 if $\Delta L_i > P$ and 0 otherwise. Given the predicted label $y = \{0, 1\}$ and the reference label $t = \{0, 1\}$ for each segment *i* in the development database, a bias-variance decomposition is obtained as described in Section 6.4. A feature set that corresponds to the lowest value of the MZOE is selected.

However, if the number of instances in the development database is not high enough (like in the case of the multimodal database of the meeting-room acoustic events), it may happen that the MZOE function is not sensitive enough to small changes in the candidate feature set Ω. Let's see an example in Table 6.5.1, where the relation between the number of selected features and the number of errors for the "Chair moving" AE is presented. Notice that MZOE is not useful to decide about selecting the features ordered as 5th, 6th and 7th since the number of errors does not change when those three features are included in the feature vector.

Table 6.5.1. Relation between the number of features and the number of errors for the "Chair moving" AE

Number of features	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**...**	**31**	**32**
Number of errors	*25*	*9*	*4*	*2*	*2*	*2*	*2*	...	*8*	*8*

To avoid that problem, the MSE loss function is used instead of MZOE. The predicted value y_i for each audio segment i in the development database is obtained as described in the following. First, the LLR is computed for all "Class" and "non-Class" AE instances in development database. As an example, in Figure 6.5.1 we display the LLRs for all AEs in the labelled development database. Squares correspond to the "Class" AE instances ("Chair moving" in our case) while crosses correspond to "non-Class" instances. Negative values of the LLR are substituted by 0. As we can see from that plot, most of the "Class" instances have higher LLR values than the "non-Class" ones. We consider the parameter P as a threshold (the horizontal line in Figure 6.5.1), and we selected P=270 for illustrative purposes. The ith AE instance is detected as "Class" if its LLR ΔL_i is above P, otherwise it is detected as "non-Class". Thus, all "Class" instances (squares in Figure 6.5.1) below the P line are misses, and all "non-Class" instances (crosses) above the P line are false alarms. In our experiments, P is selected in such a way that the numbers of misses and false alarms are equal (equal error rate).

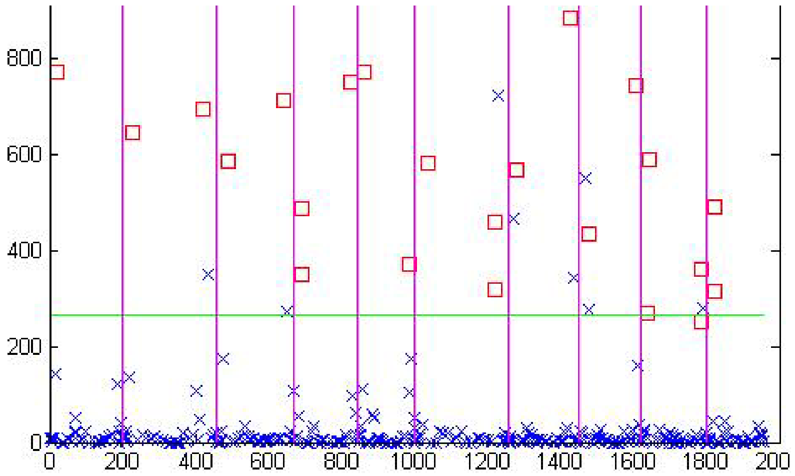


Figure 6.5.1. The LLRs corresponding to "Chair moving" class

Second, the predicted value y_i for *i*th AE instance lies in the range [0 1] and it is computed from the LLR ΔL_i with the expression:

$$y_i = g(\Delta L_i, P) = \begin{cases} 1 - \dfrac{P}{2\Delta L_i}, & \text{if } \Delta L_i > P \\ \dfrac{\Delta L_i}{2P}, & \text{otherwise} \end{cases} \tag{6.5.2}$$

The function *g*(.) is displayed graphically in Figure 6.5.2.

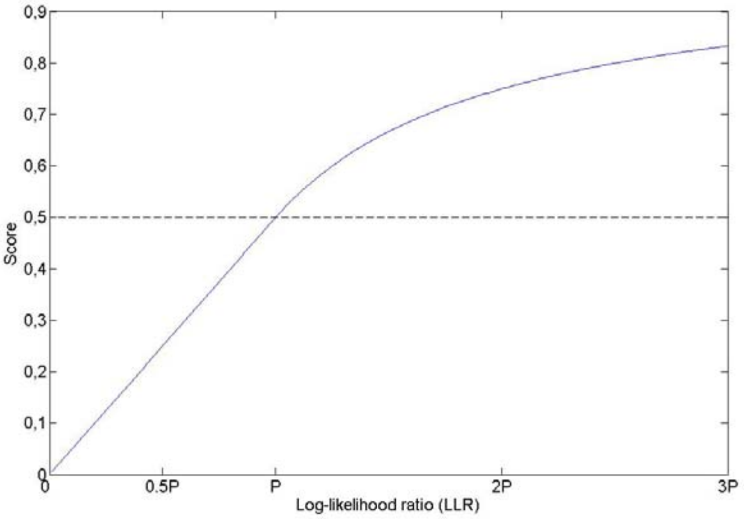


Figure 6.5.2. The normalization function g(.) used for MSE loss computation

Given the predicted and reference labels for each AE instance *i* the bias-variance decomposition of the MSE is obtained as described in Section 6.4.

6.6 Fast feature selection

Although the SBS searching strategy requires much less computation load than the exhaustive search, it is still impractical to use it even for moderate number of initial features *d.* The most time consuming operation during feature selection is the re-estimation of GMM acoustic models using expectation-maximization (EM) algorithm that is called repeatedly for each candidate feature set. Our objective is to make the sequential feature selection approach practical for large number of initial features. In our work, the GMM acoustic models Θ_i for each *i*th acoustic class are obtained only once at the beginning of feature selection process using the whole set of initial features Ω. Using a standard marginalization technique [Pap91], the same acoustic models Θ are used to evaluate any feature subset $R \subset \Omega$ by means of marginalization of components that are not included in *R*.

Further advantage in speed can be achieved by using the single dominant component in the GMM probability computation. The log-likelihood of any pattern *x* given GMM acoustic model Θ can be approximated by expression:

$$Q(x|\Theta) = \log(\sum_i \alpha_i N(\overline{x_i}, \Sigma_i)) \approx \widetilde{Q}(x|\Theta) = \log(\alpha_k N(\overline{x_k}, \Sigma_k)) \qquad (6.5.3)$$

$$k = \arg\max_i(\alpha_i N(\overline{x_i}, \Sigma_i)) \qquad (6.5.4)$$

$$N(\overline{x_i}, \Sigma_i) = \frac{1}{2\pi^{d/2} \mid \Sigma_i \mid^{1/2}} e^{-\frac{1}{2}(x-\overline{x_i})^T \Sigma_i^{-1}(x-\overline{x_i})} \qquad (6.5.5)$$

where $\overline{x_i}$, Σ_i and α_i are the mean vector, the covariance matrix and the mixture weight of the *i*th mixture component, respectively; *d* is the total number of features.

To validate the adequateness of the approximation (6.5.3) in our framework, in Figure 6.6.1 we show the approximation error as average relative difference between LLR values ΔL_1, ΔL_2 over all AE instances in the development database. The values ΔL_1 and ΔL_2 are obtained using left and right parts of the expression (6.5.3), respectively, and the average value $(\Delta L_1 - \Delta L_2)/|\Delta L_1|$ (in %) is depicted along the vertical axis as a function of the number of features. Note the approximation error is less than 5% provided that more than 6 features are used.

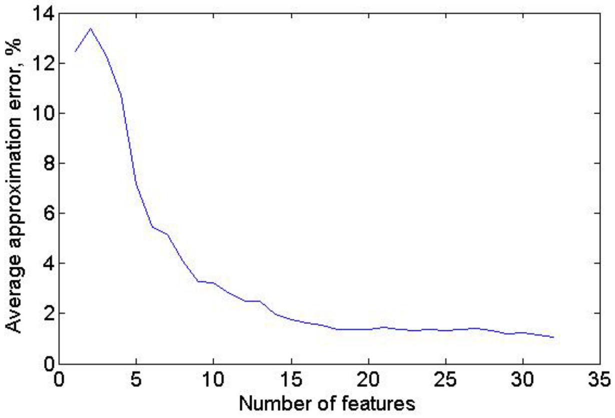


Figure 6.6.1. Average approximation error computed from the left and right parts of (6.5.3)

Assuming diagonal covariance matrix *Σ_k,* we further obtain:

$$\widetilde{Q}(x|\Theta) = \log(\alpha_k \frac{1}{2\pi^{d/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\bar{x}_k)^T \Sigma_k^{-1} (x-\overline{x_k})}) = \log\frac{1}{2\pi^{d/2}} + \log\alpha_k + \sum_{i=1}^{d} \log\frac{1}{\sigma_i} e^{-\frac{(x-\overline{x_i})^2}{2\sigma_i^2}}$$

(6.5.6)

where σ_i^2 are the diagonal elements of the covariance matrix Σ_k. Thus the log-likelihood of each example *x* can be decomposed into the sum of the following terms: a constant term, the logarithm of the mixture weight, and the sum of the components that can be considered as the contribution coming from each feature. These contributions are computed at the beginning of the feature selection process using the initial acoustic model Θ. These contributions are further used for LLR computation given any subset of initial features.

6.7 Experimental results

In the following experiments, first, we present the comparison results between the proposed fast feature selection technique that avoids retraining of acoustic models during the evaluation of the candidate feature set and the conventional approach. Second, we assess the performance of the proposed fast feature selection technique in 2 steps: in the first step the optimal set of features for each class is found using the development database. At the second step we analyze the improvement obtained by the system based on the selected features with respect to the system that uses the whole feature vector using testing database. Three different feature selection approaches are employed in our experiments: a purely wrapper SBS approach and the two hybrid approaches, PCA and mRMR, that combine filter for feature set ordering and wrapper for feature set evaluation. We analyze feature selection results for both the meeting-room AE database and the tv3/24 broadcast news database using HMM-GMM detection approaches described in chapter 4 and chapter 5, respectively.

6.7.1 Fast one-pass training feature selection approach

The multimodal database of meeting-room AEs is employed to compare fast feature selection technique with the conventional one. We start with an initial feature set that consists of 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives (in total, 32 features). We also added features coming from the acoustic source localization system (1 additional feature for all AE classes) and video signals (1 additional feature for 5 classes of interest).

Figure 6.7.1 summarizes the mean relative improvement obtained by the system based on selected features with respect to the system that uses the whole set of features. The results correspond to the testing part of the database (unseen data that is not used during the feature selection process). The dashed line corresponds to the results obtained with the baseline conventional approach, and the solid curve corresponds to the one-pass-training technique. The standard deviation is plotted with vertical lines. It has been calculated from 8 scores, which were obtained by using different combinations of partitions of the database for training, development and testing. According to them, the detection rate increases for all classes, except “Cough” and “Paper work”, by using any of the two feature selection techniques. We observe that the fast feature selection technique shows similar results compared to the conventional feature selection approach. The average of the mean relative improvement across the AEs (horizontal axis) equals to 4.5% for the conventional and 5.0% for the one-pass-training approach.

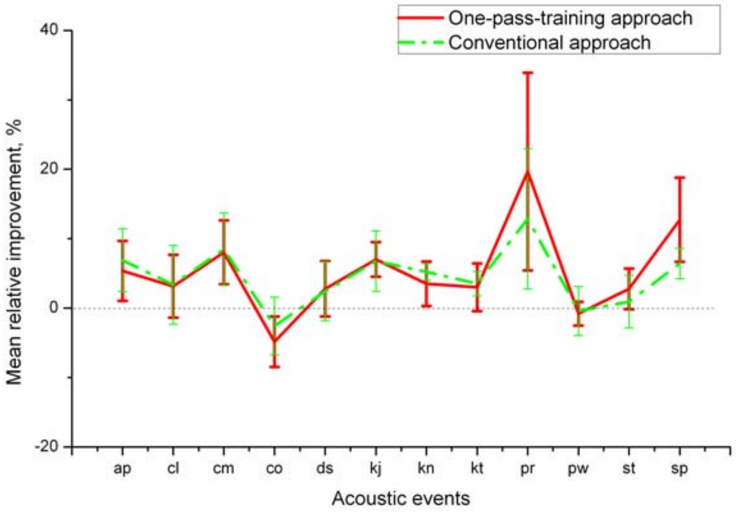
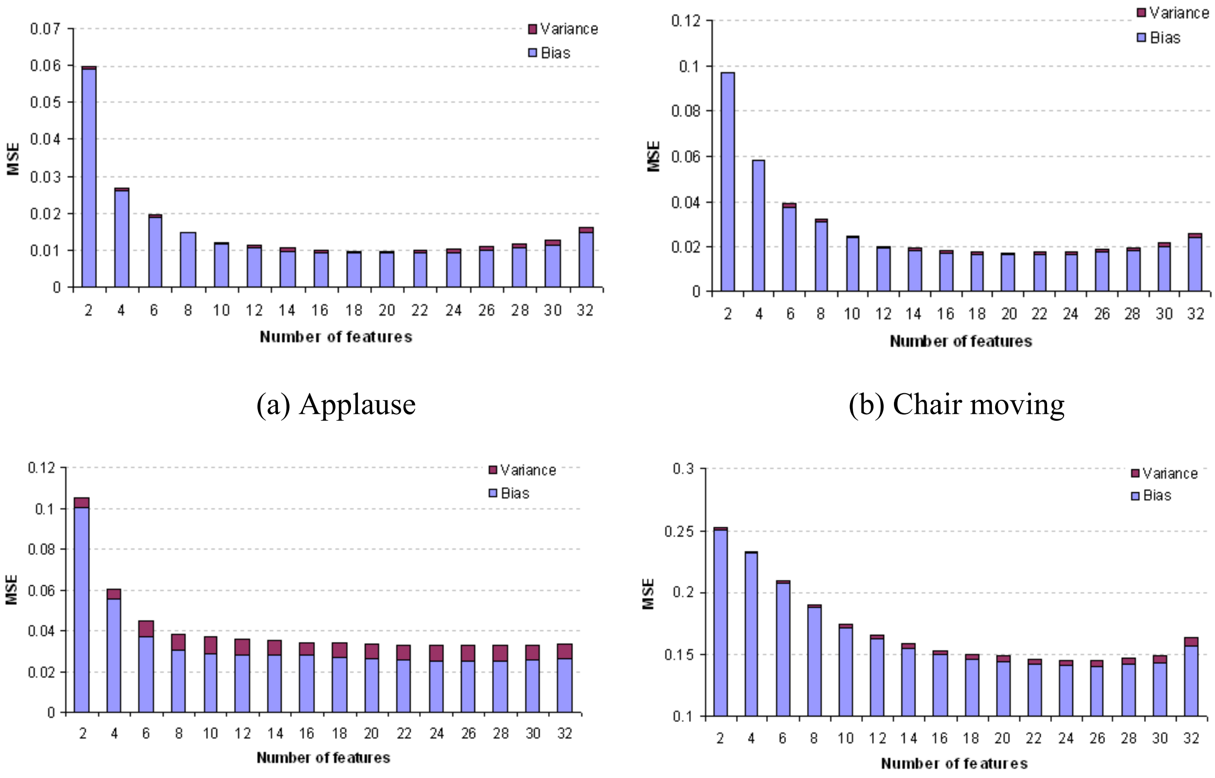


Figure 6.7.1. Comparison between the conventional and the one-pass-training features selection techniques

6.7.2 Feature selection for meeting-room acoustic event detection

Similarly to the experimental results in previous sub-Section, we start with an initial feature set that consists of 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives (in total, 32 features). We also added features coming from the acoustic source localization system (1 additional feature for all classes) and video signals (1 additional feature for 5 classes).The multimodal database has been divided into 2 parts: training (5 sessions) and development (3 sessions). The training part of the database has been further splitted into *D*=9 portions (each partition consists of 2 sessions, with 50% overlap between partitions).

Figure 6.7.2 shows an example of bias-variance decomposition of MSE error for four AE classes: “Applause”, “Chair Moving”, “Door Slam”, “Steps”; the feature ordering was done using SBS approach. The total height of each bar is the MSE error for the number of features on the *x*-axis. Each bar is subdivided into portions that are due to bias and variance of the detection algorithm.



(a) Applause

(b) Chair moving

(c) Door slam

(d) Steps

Figure 6.7.2. Bias-variance decomposition of squared error for several AE classes using development database

As we see from Figure 6.7.2, the variance term of the error of the detection algorithm increases with increasing the number of features, while the bias term of the first decreases and then, after reaching some point, increases. For each AE we select the optimal number of features with the MSE bias-variance trade-off for the learning algorithm (18, 20, 29, 26 features for each AE in Figure 6.7.2, respectively). For some AEs (like "door slam") the feature selection does not contribute significantly to MSE reduction.

Once the number of features per each class is selected, the next step is to evaluate the performance of the two AED systems: the first one is based on the selected features and the second one uses the whole 34-dimensional (for five classes of AEs) or 33-dimensional (for the rest of the classes) feature vector. The evaluation experiment was repeated across 8-folds and the 8 estimates averaged. In order to have the results consistent with the experiments in Section 6.7.3, we present the results in terms of (1 – AED-ACC) score, so that better results correspond to lower value.

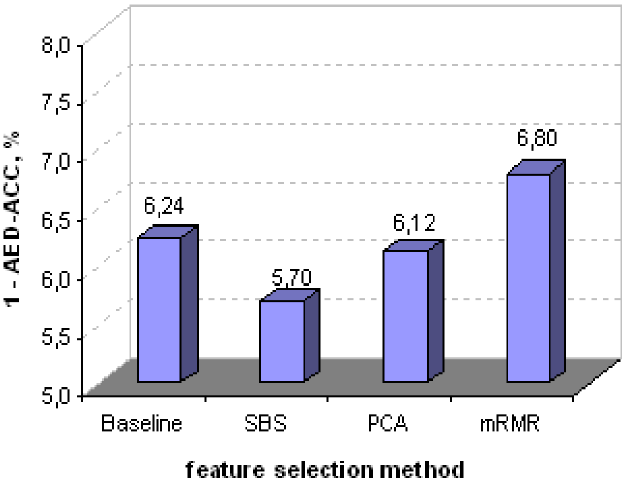


Figure 6.7.3. Comparison of the baseline recognition results with the results obtained along different AEs (a) using different features selection approaches

In Figure 6.7.3 we compare the baseline (1 - AED-ACC) score averaged over different AE classes with the results obtained by the systems based on selected features using three different feature selection techniques: SBS, PCA and mRMR. The best result corresponds to the SBS approach, achieving 8.7% of relative error reduction. With PCA and mRMR aproach we do not report significant improvement.

In Table 6.7.1, the number of selected features for different categories of features and feature selection methods is displayed. We decompose the 32 FF features into 16 static (S) and 16 dynamic (D) parameters. The next columns correspond to the number of selected features coming from localization and video, respectively. For PCA technique such kind of decomposition is not possible since the selected features do not have a clear physical meaning. In the case of SBS as well as mRMR technique, both static and dynamic FF-based features contribute to the final accuracy. The video features are an important additional source of information for detection of those five AEs for which video features are extracted. In the case of SBS technique, the acoustic localization feature was selected for eight AEs, but not for the other four. One of the reasons of such behaviour is high variability of the estimated localization coordinates for those four AEs. For instance, “Key jingle”, “Phone ring” and “Speech” do not have clearly associated *z*-coordinate in the room. Regarding low-energy AEs like “Paper work” the reliability of localization feature estimate is low compared to other AEs.

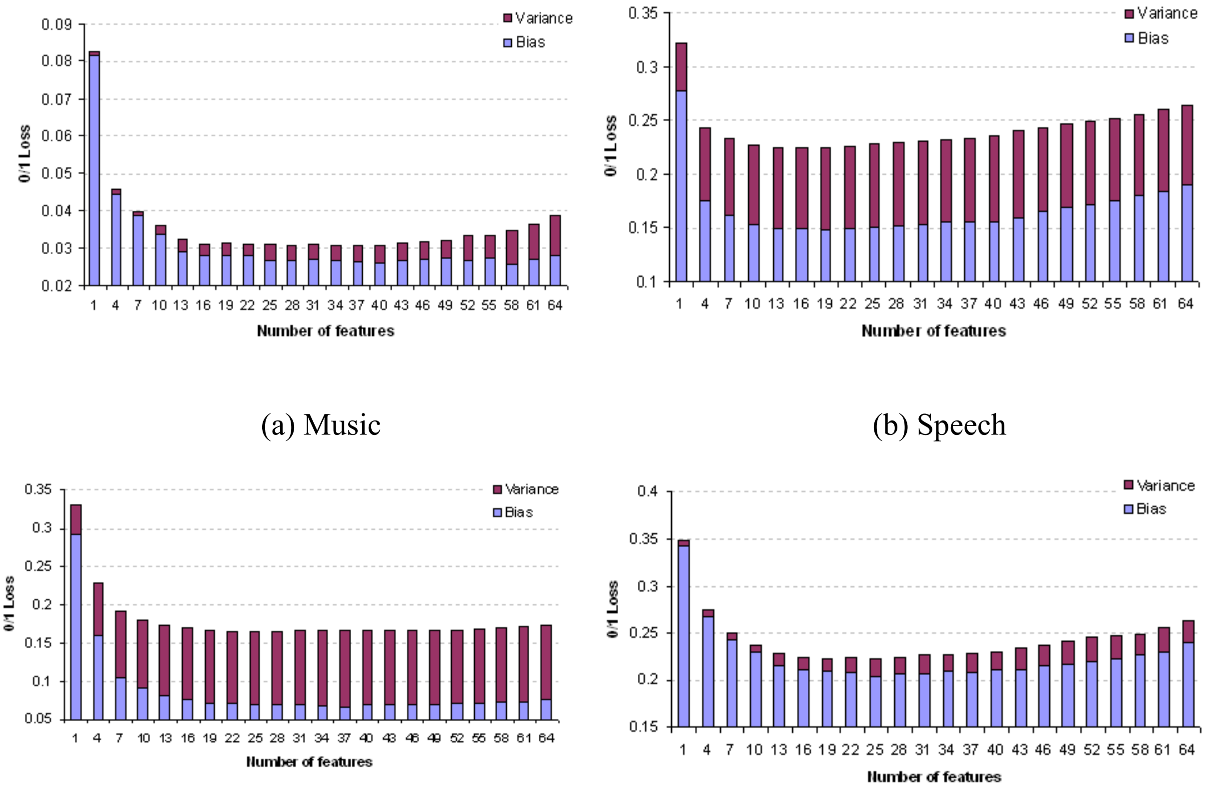
Table 6.7.1. Number of selected features for each AE

Es	Feature selection method										
	SBS					*mRMR*					*PCA*
	AST		Loc	Video	Total	AST		Loc	Video	Total	Total
	S	D				S	D				
Door knock	13	8	1	---	**22**	15	16	1	---	**32**	**14**
Door open/slam	14	13	1	1	**29**	13	15	1	1	**30**	**24**
Steps	13	11	1	1	**26**	15	16	1	1	**31**	**33**
Chair moving	10	8	1	1	**20**	16	15	1	1	**33**	**32**
Cup clink	8	7	1	---	**16**	16	15	1	---	**32**	**18**
Paper work	14	10	1	1	**26**	13	15	1	1	**30**	**16**
Key jingle	8	9	0	---	**17**	7	8	0	---	**15**	**32**
Keyboard typing	12	10	0	1	**23**	14	16	1	1	**32**	**33**
Phone ring	5	7	0	---	**12**	2	1	0	---	**3**	**32**
Applause	8	9	1	---	**18**	4	5	1	---	**10**	**23**
Cough	8	11	1	---	**20**	15	15	1	---	**31**	**33**
Speech	8	10	0	---	**18**	13	12	0	---	**25**	**18**
Average					**21**					**25**	**26**

6.7.3 Feature selection for broadcast news audio segmentation

The database that consists of 24 sessions has been divided into training (10 sessions), development (6 sessions) and testing (8 sessions) parts. The training part of the database has been further splitted into *D*=5 partitions (each partition consists of 5 sessions, with 40% overlap). The initial feature set consists of 16 frequency-filtered log filter-bank energies (FF LFBE) with their first time derivatives (in total, 32 features). The mean and the standard deviation of the features have been computed along a 1 sec window, thus forming a feature vector of 64 elements every 1 sec.

Figure 6.7.4 shows the bias-variance decomposition of the MZOE for 4 acoustic classes: “Music”, “Speech”, “Speech over music”, “Speech over noise”; the feature ordering was done using the SBS technique. Note that for the classes “Speech” and “Speech over music” the variance term of the error of the detection algorithm is much larger than for the other two classes. Similarly to the MSE, the variance term of the MZOE increases with increasing of the number of features.



(a) Music

(b) Speech

(c) Speech over music

(d) Speech over noise

Figure 6.7.4. Bias-variance decomposition of zero-one error for different acoustic classes using SBS approach

In Figure 6.7.5 we compare the AS results obtained by the systems based on the selected features (using SBS, PCA and mRMR techniques) with the system that uses the whole feature vector (baseline). When using the SBS technique the error rate decreases for all classes; using the PCA technique we got an improvement for 3 classes and, finally, using mRMR we got an improvement just for 2 classes. The overall detection results (when combined in the hierarchical way) are very similar in the case of the SBS and the PCA techniques (34.77% and 35.07%, respectively), achieving around 2% of absolute error reduction.

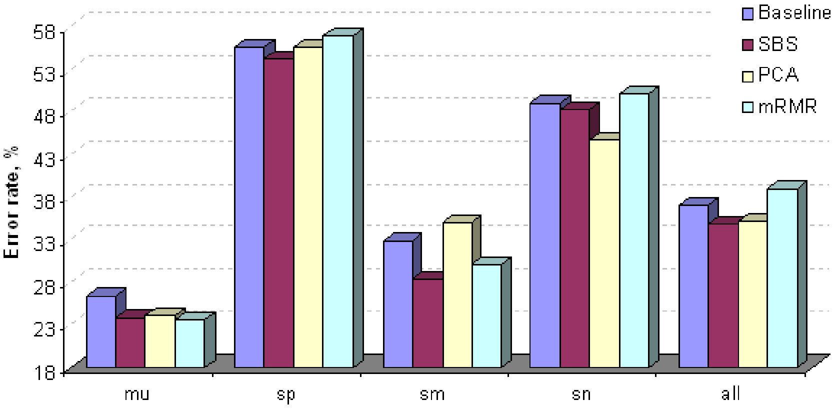


Figure 6.7.5. Comparison of the detection results along different acoustic classes using different feature selection approaches

In Table 6.7.2, the number of selected features for different feature selection techniques and different acoustic classes is displayed. Using the SBS approach we achieved the largest feature reduction ratio: 2.6, being 4.6 for “Speech” class and 1.8 for “Music” class. Noticeably, that for the classes which include overlapping of speech with either music or noise the feature reduction ratio is in between of these values.

Table 6.7.2. Number of selected features for each AC using different feature selection techniques

	Feature selection technique		
	SBS	mRMR	PCA
Music	36	62	47
Speech	14	16	16
Speech over music	27	61	59
Speech over noise	18	20	48
Average	24	40	43

The CPU time (in hours) required to perform feature selection for one acoustic class is summarized in Table 6.7.3. The processes were executed using Intel Xeon with 4 cores and 3 GHz; 32 Gb of RAM. The whole feature selection process is decomposed in 4 steps: feature ordering, initial training of models using the whole set of features and the evaluation of the candidate feature sets. In the case of SBS wrapper approach, $d(d\text{-}1)/2$ feature subsets should be evaluated; in the case of filter approaches only d nested feature subsets are evaluated, where d is the total number of features. We have presented the timing results for both fast and conventional techniques.

Table 6.7.3. Breakdown timing for feature selection algorithms (in hours)

Feature selection operation	Meeting-room AED			Broadcast news AS		
	SBS	mRMR	PCA	SBS	mRMR	PCA
Feature ordering using filter approach	-	0.13	0.002	-	0.24	0.002
Initial training of models	0.4			0.39		
Wrappers-based subset selection (fast technique)	56.1	3.4		80.6	2.6	
Wrappers-based subset selection (conventional technique)*	280.5	17		866.9	27.5	

* Not evaluated, estimated value

As we expected, the filter approaches are much faster than wrapper one. The fast wrapper-based technique is about 5 and 10 times computationally efficient than conventional one for database of isolated AEs and for broadcast news AS database, respectively. The gain in execution time is more evident when the initial number of features increases.

6.8 Chapter Summary

Taking into account a large computation load of the conventional wrapper approaches, a fast one-pass-training technique has been proposed that does not require the re-estimation of acoustic models during the evaluation of the candidate feature sets. The speed benefit of this technique becomes extremely important when the initial feature set is large and the training of acoustic models is time consuming. The experimental results show that the fast technique can be successfully applied in our AED tasks.

Three different feature selection approaches are compared in the framework of multimodal meeting-room acoustic event detection and broadcast news audio segmentation. The wrapper-based sequential backward selection approach showed the best results in terms of both accuracy and feature reduction ratio at the cost of high computational complexity. The hybrid two-stage PCA algorithm show slightly worse results in terms of accuracy, but much better performance in terms of speed. Note, for the PCA feature selection the extraction of the whole set of initial features is required so that it may delay the testing phase of detection/segmentation algorithm. This issue could be crucial in embedded and online systems as the extraction of unimportant features may be computationally expensive. Besides that, the physical meaning of the selected features is lost so that the PCA approach does not give any feedback about the suitability of a particular feature for the detection task. With the hybrid two-stage mRMR approach the recognition results are improved only for some classes but the average error along all acoustic classes increases for both databases employed in the paper. The possible reason of that is the difficulty of estimation of mutual information between features if they have continuous distribution.

The performance characteristics of the learning algorithms have been analyzed using the bias-variance decomposition of the classification loss. For mean-square error and zero-one loss this decomposition has very simple form: the error is composed of two non-negative terms, one comes from classifier imperfectness as well as intrinsic error of domain and the other comes from fluctuation of classifier prediction over different possible training sets. We observed that for all acoustic classes the bias of GMM-based classifier increases after reaching some point, while in general, the variance of the learner monotonically increases with increasing number of features. The most accurate feature set corresponds to the best bias-variance trade-off point for the learning algorithm.

Using the multimodal database of acoustic events we found that both static and dynamic spectro-temporal features contribute to the final accuracy. According to the obtained results, the video features are an important additional source of information for detection of the five AEs for which video features are extracted. The acoustic localization feature was selected only for some AEs.

Chapter 7. Online 2-source AED implementation

7.1 Chapter Overview

In this chapter we describe the 2-source acoustic event detection (AED) and acoustic source localization (ASL) systems implemented in real-time in the UPC's smart-room using HMM approach.

The remaining sections are organized as follows. In Section 7.2, the previous activities implemented in UPC's smart-room are briefly outlined. Section 7.3 describes the proposed 2-source AED/ASL system. Section 7.4 gives the implementation overview of the of the AED/ASL components using smart-audio++ package. Finally, in Section 7.5 we present some comparison results between online and offline implementations of AED.

7.2 Smart-room activities

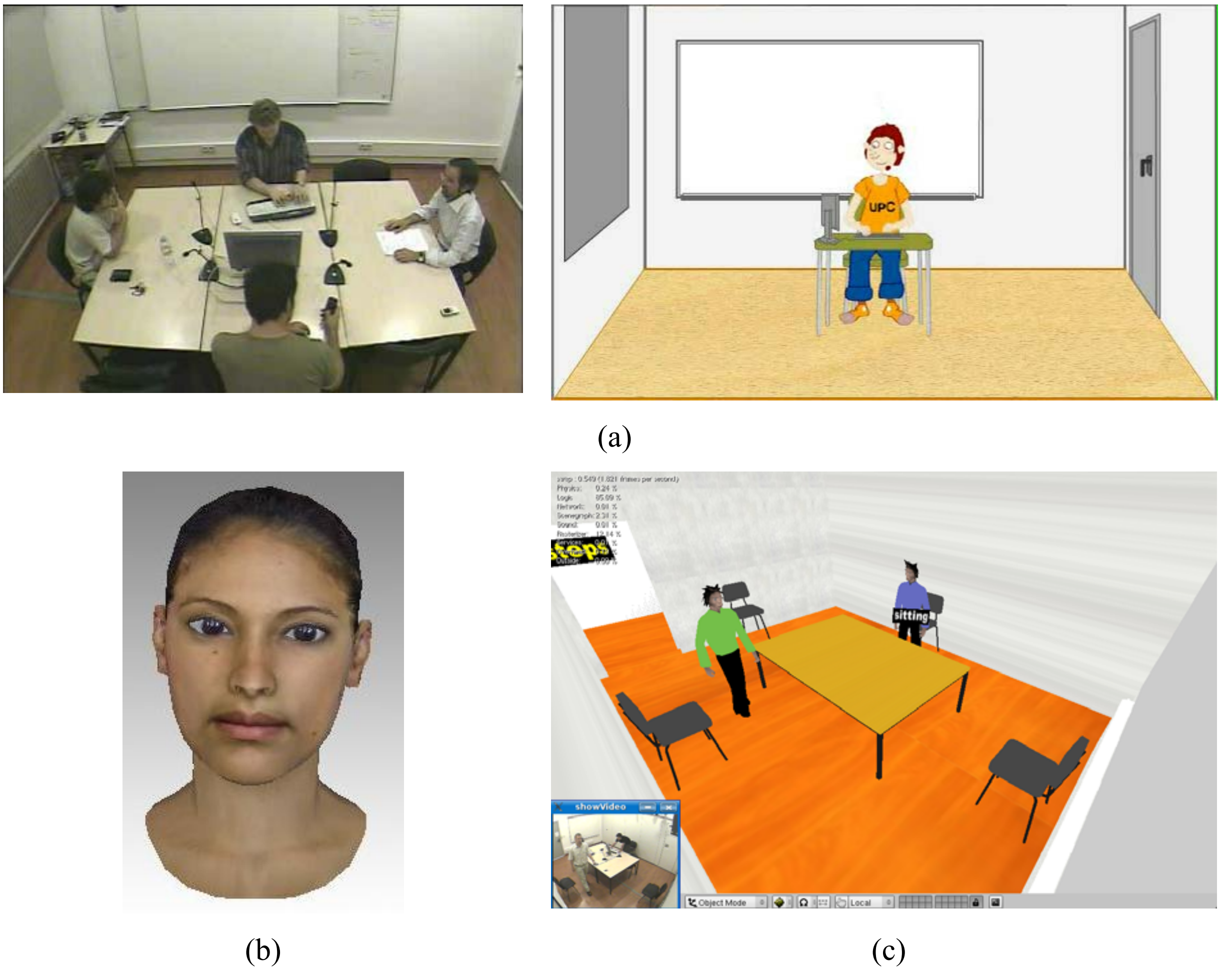
One example of new challenging multimodal research efforts is the development of smart-rooms. A smart-room is a closed space equipped with multiple microphones and cameras, which are designed to assist and complement human activities. The room serves two purposes: first, it is an experimentation environment, where researchers can test online the multimodal analysis and synthesis developments in the area of human–computer interfaces; second, it doubles as a data collection facility for research purposes, providing data for technology development and evaluation. The configuration of the smart-room created in UPC can be found in Figure 3.3.1 (a). Among others, there are several audio-visual sensors (cameras and microphones), synchronization and acquisition equipments, working computers, and a video projector. Several technologies have already been implemented in the UPC smart-room that work in real-time.

The first demo that combines both AED and ASL technologies was created by our group in the context of European CHIL project [WS09] in 2007. The one-source AED was implemented using Support Vector Machines (SVMs) [TN09]. In the SVM realization, the AED is performed by means of sequential classification of 1 sec sliding windows with 0.2 sec shift. The snapshot of the video demonstrating one-source AED/ASL is presented in Figure 7.2.1 (a). As it shown in the Figure, there are two screens in the GUI output: one corresponds to the real video captured from one of the cameras installed in the UPC's smart-room, and the other is a graphical representation of the output of the AED and ASL technologies. In this demonstration three persons produce several acoustic events in isolated way without signal overlaps.

The second demo is so called talking head demo (Figure 7.2.1 (b)) that demonstrates the context awareness given by perceptual components. According to the scenario, the talking head "interacts" with the journalists in natural way: it not only informs the journalists about available resources, and points out events such as the arrival of a latecomer or news being contributed by remote colleagues, but also facilitates information requests from the journalists in a human-like interface based on automatic speech recognition technologies. Interactive behaviour of the talking head is also achieved by means of detection of some acoustic events, like an utterance "Don't forget your keys!" when "key jingle" is detected or exclamation "Great! Well done!" when "applause" is detected. Five AEs were chosen for detection: door knock, door opening/closing, speech, applause, and key jingles.

The third demo developed in our UPC's research group demonstrates the UPC's smart-room remotely. The functionalities that are involved in the demonstration are the 3D person tracking, ASL, and AED. The virtual 3D scene is reconstructed from the images of multiple cameras in the smart-room. Figure 7.2.1 (c) shows the developed 3D visualizer. Specifically, it shows one detected person sitting, and the other passing in the room. “Steps” are detected with the AED and localized in space with the ASL. The text label “Steps” is assigned to the place where the event happens. Additionally, the small screen at the lower left corner shows the corresponding real video.

All the previously described demonstrations of the AED technology working online are focused on the detection of single AE without signal overlaps. In fact, in real scenarios the meeting-room AEs are usually overlapped with other sounds, mainly with speech. In our work we extended the online implementation of AED to two sources, one of which is always speech.



(a)

(b)

(c)

Figure 7.2.1. Previous smart-room demos that make use of AED technology. (a) One-source AED/ASL demo (b) The talking head demonstrating the context awareness given by perceptual components (c) Virtual UPC’s smart-room demo

7.3 From one-source to two-source detection

While the detection of the single acoustic source is performed with relatively high accuracy, a big challenge is dealing with two simultaneous acoustic events produced from different persons in the room. According to the last international evaluation campaign the overlapping segments account for more than 70% of errors produced by the systems. In our scenario we assume that one acoustic source is always speech and the other is a particular meeting-room AE. In our implementation we propose an alternative algorithm for AED based on HMMs, where the acoustic analysis is performed on a frame-by-frame basis, using Viterbi segmentation algorithm for recognition which allows obtain the recognition result with very low latency. Besides, the acoustic source localization algorithm is extended to 2 sources.

In our work, the overlap problem is faced at the level of models. 24 HMMs have been trained, one for each isolated acoustic event (12 acoustic models) and one for each AE overlapped with speech with different SNR (the other 12 acoustic models).

The flowchart of the proposed online AED/ASL system is depicted in Figure 7.3.1. In consists of 4 main blocks: audio acquisition, 2-source AED, 2-source ASL and visualization.

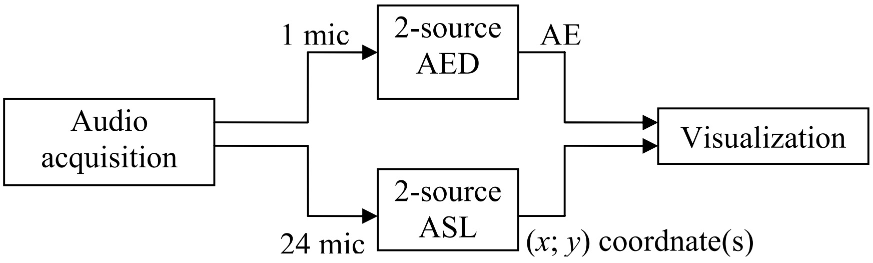


Figure 7.3.1. The flowchart of the proposed AED and ASL system working online

The audio captured from the set of microphones is used for subsequent feature extraction and recognition in 2-source AED block; the 2-source ASL block estimates the position(s) of acoustic source(s). The output from 2-source AED block is either an isolated AE (if one source is detected) or an AE overlapped with speech (if 2 sources are detected). The output from 2-source ASL block is either one or two (*x, y*) coordinates of acoustic source(s). Both outputs are combined together and visualized by graphical user interface (GUI). An ambiguity happens when the numbers of acoustic sources detected by AED and ASL blocks are different. In this case the number of the displayed acoustic sources equals to the number of acoustic sources detected by the AED module. Note that

AEs are displayed in default positions if the number of localization coordinates is less than the number of detected AEs.

7.3.1 Audio acquisition

In the audio acquisition block, the audio signals are captured simultaneously from 24 microphones from T-shape clusters located on the walls of the room. The audio signal from the microphone #18 is downsampled from 44.1 kHz to 16 kHz and subsequently used for feature extraction and recognition in 2-source AED block. Actually, the microphone #18 is the nearest one to the table and it allows the most reliable recognition of AEs. The 2-source ASL block uses the audio from the whole set of 24 microphones sampled at 44.1 kHz to estimate one or two positions of acoustic sources.

7.3.2 Two-source AED system

The first step in our 2-source AED system is feature extraction. A set of audio spectro-temporal features is extracted to describe every audio signal frame. There exist several alternative ways of parametrically representing the spectrum envelope of audio signals. Similarly to offline tests, we employ frequency-filtered (FF) log filter-bank energies (LFBE) parameters. In our experiments, the frame length is 30 ms with 20 ms shift, and a Hamming window is applied. A 32-dimensional feature vector is computed every 20 ms.

Each acoustic class is modelled using HMMs where GMMs are used to compute the state emission probability as described in chapter 4. The HTK toolkit [YEK02] is used for training the HMM–GMM models. The audio part of the multimodal database described in Appendix A is used to create those models. There is one HMM for each AE, with only one emitting state. The observation distribution of this state is Gaussian mixture with continuous density, and consists of 64 components with diagonal covariance matrix. Each HMM is trained with the signal segments belonging to the corresponding event class using the standard Baum–Welch training algorithm [RJ93]. The overlap problem has been handled at the level of models. In total, 24 HMMs is trained, one for each isolated AE class and one for each AE class overlapped with speech. For testing, the Viterbi algorithm is used. For online system implementation we use ATK, an API designed to facilitate building experimental applications for HTK.

To build an application using ATK three main resource files have to be prepared off-line. The first one is a dictionary that contains all possible acoustic events for detection. In our case it is a set of 12 isolated AEs: “ap”, “cl”, “cm”, “co”, “ds”, “kj”, “kn”, “kt”, “pr”, “pw”, “st”, “sp” and 12 AEs overlapped with speech: “oa”, “oj”, “om”, “oh”, “od”, “ok”, “on”, “ot”, “or”, “ow”, “os” and “oo”.

The second resource file is the grammar file that defines a possible sequence of AEs as illustrated in Figure 7.3.2. In HTK SLF format, this would be represented by a text file containing a list of grammar nodes (the AEs) and the links needed to combine them together. The third required resource is a set of HMMs. These would be prepared off-line using HTK, and then stored in HTK supported format. Each of the three required resources can be defined as entries in a configuration file which is loaded at start-up time.

In ATK, *ARec* component provides similar functionality to the standard HTK Viterbi decoder. It is supplied with a resource group containing dictionary, a grammar, and then decodes incoming feature vectors accordingly.

In the proposed implementation, the *ARec* recogniser always operates in one of three possible states as indicated by the state diagram shown in Figure 7.3.3. Initially, when there is no acoustic activity in the room, the recognizer operates in FLUSH state and discards the input audio packets. Audio packet is a chunk of information that is used for transmitting between asynchronously executing components. When the energy of the *N* consecutive packets exceeds the predefined threshold, the recognizer starts operating in RUN state. Similarly, the recognizer goes back to FLUSH state if the energy of *N* consecutive packets are below this threshold. In RUN state the recognizer continuously performs the Viterbi decoding of the waveform from the time interval [t_0 t_i], where t_0 is the time instance when the first non-silence packet received and t_i is the current time stamp. The grammar is defied as illustrated in Figure 7.3.2, that allows only one AE to be detected on this interval. In ANS state the recognizer sends the current decision obtained in RUN state to visualization block. There are 2 possible conditions when the recognizer goes to ANS state:

- The confidence of the current decision exceeds a predefined threshold. In this case the corresponding AE label is sent to visualization block. Very often a confident decision is obtained with just a few input packets. In this case the time delay between the AE production and its visualization is small allowing obtaining recognition result with low latency.
- During 1 sec operation in RUN state the decision with enough confidence is not obtained. In this case the “unknown” AE is sent to visualization block. Note, that different thresholds are experimentally tuned for each AE. It allows controlling the number of misses and false alarms for each AE individually. In fact, false alarms are the most annoying kind of errors in output from the AED process.

The recognizer goes back to RUN state immediately when the output label is sent to visualization block.

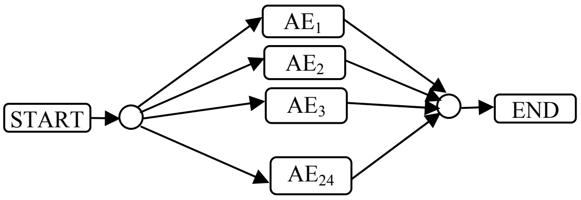


Figure 7.3.2. AED grammar

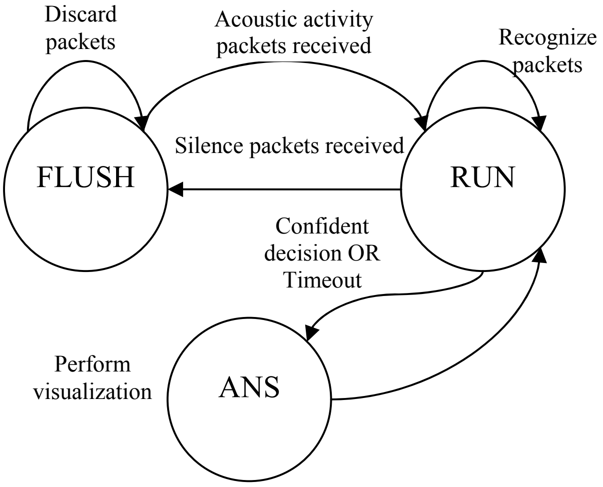


Figure 7.3.3. Finite-state machine

ARec supports a simple method of confidence computation. Every frame, the acoustic log likelihood of the best matching model state and the best matching background model state is saved. Since we do not use any background model, the average score across all models is used instead. When the AE is recognised, these *best-state* and *background state* scores are summed to form a best-possible-acoustic score *bs* and a background score *bg* over the segment of the waveform for which the AE is being hypothesised.

A raw confidence score in the range -1 to 1 is computed as:

$$rawconf = ac - \frac{bs + bg}{2}$$

where *ac* is the actual acoustic log likelihood of the AE. The confidence for that AE is then computed as

$$conf = \frac{e^x}{e^x - e^{-x}}$$

where *x* is scaled *rawconf* score

$$x = \alpha\, rawconf - \beta$$

The constant α sets the slope of the confidence curve and β sets the operating point. Their values are set with default values of *0.15* and *0.0*.

7.3.3 Two-source acoustic source localization system

The acoustic localization system used in this work is based on the SRP-PHAT localization method described in chapter 3. The contribution of the cross-correlation of every microphone pair is accumulated for each exploration region in the room. In this way, we obtain a sound map at every time instant of 50 ms, as depicted in Figure 7.3.4, where in red colour the regions with high cross-correlation contribution from all microphone pairs are highlighted. The estimated locations of acoustic sources are the positions of the quantized space that maximize that contribution.

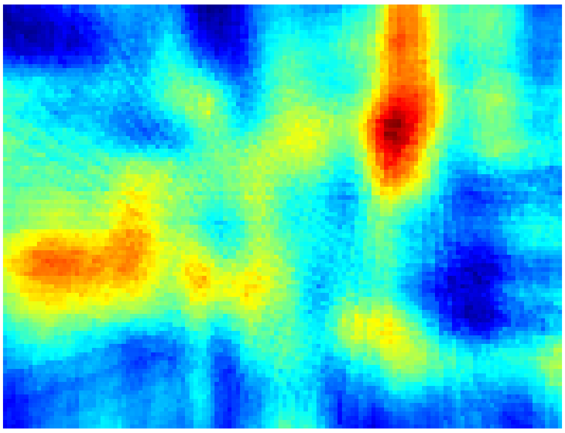
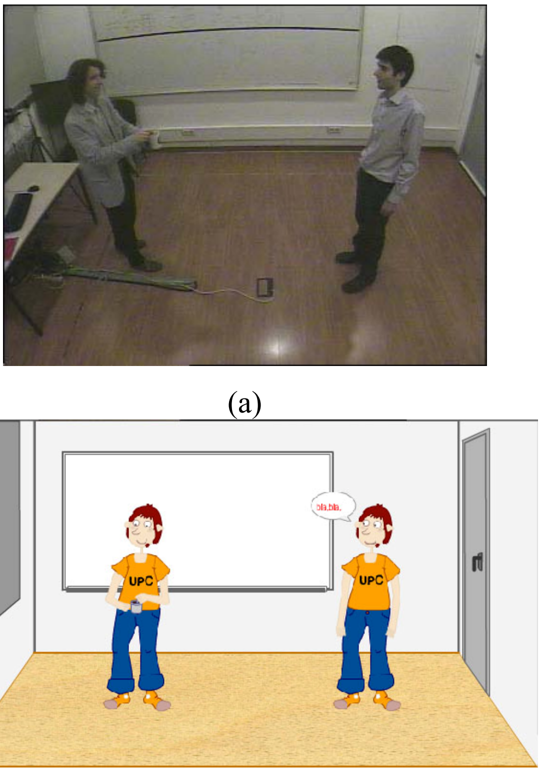


Figure 7.3.4.– Example of the Sound Map obtained with the SRP PHAT process

There is no constraint on the number of acoustic sources that the algorithm has to detect. We employ the method that dynamically estimates the number of sources based on a birth/death system [SAH07]. The ASL system uses a spatial segmentation algorithm to group locations that are close to each other in space and time. When a minimum number of locations N_b are found in a space region over a defined time window T_b, the system decides whether it is a new acoustic source. Similarly, if the previously detected acoustic source does not have any measurements that fall within its acceptance region for a given amount of time T_d, then it is dropped. The ratio between T_b and N_b used in the detection module is a design parameter. It must be high enough to filter out noises and outliers, but also not too high in order to be able to detect sporadic acoustic events. In our experiments N_b is set to 4, T_b is 460 ms and T_d is also 460 ms.

7.3.4 Visualization

The developed graphical interface (GUI) fully describes the acoustic activity in a smart room, and allows the observers to evaluate the system performance in a very convenient way. The GUI application is based on the QT Trolltech toolkit [QT], an open-source library widely used for the development of GUI programs. There are two screens in the GUI output, as shown in Figure 7.3.5. One corresponds to the real video captured from one of the cameras installed in the UPC's smart-room, and the other is a graphical representation of the output of the AED and ASL technologies.



(a)

(b)

Figure 7.3.5. The two screens of the GUI: (a) real-time video and (b) the graphical representation of the AED and ASL functionalities (''Cup clink'' overlapped with "Speech" is being produced)

When there is acoustic activity in the room, the GUI displays the animated puppets in the positions provided by the ASL system. The number of puppets depends on whether the acoustic event is produced in isolated manner (one puppet) or it is overlapped with speech (two puppets, one of which is always producing speech, as depicted in Figure 7.3.5).

7.4 Acoustic Event Detection Component Implementation

The implementation of the AED components was done using a software package called SmatAudio++ and KSC socket messaging system developed under Linux platform. SmartFlow is a set of tools that allows components from various developers to interoperate in smart-spaces and deal with the data flows coming from that variety of sources. The use of a KSC message server and a KSC client library allows sending results of data analysis in an asynchronous way.

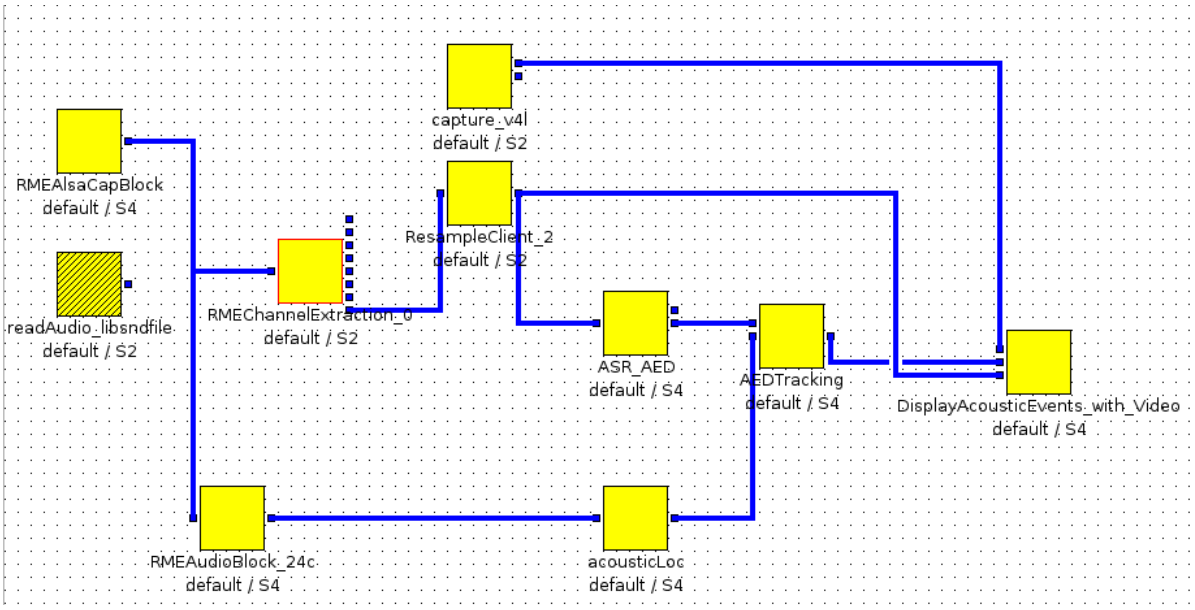


Figure 7.4.1. SmartAudio map that corresponds to the 2-source AED/ASL system

Figure 7.4.1 shows the SmartAudio map that corresponds to 2-source AED and ASL systems. The map shows the needed Smartflow clients and the interconnections among them. Each of the squares is a client designed for the execution of a specific stage of the AED/ASL task. The lines between the squares represent the connections between each client, i.e the data flows between the different AED/ASL stages. Different clients, that is, different computations required by the AED/ASL task are implemented on distinct servers in order to optimize the usage of the CPUs of the machines and splitting the computational burden among several computers. In fact, it reduces the elaboration time that is a crucial requirement for online tasks. The clients can be executed on any of the available server, whose choice has only to be guided by the criterion of the splitting of computational burden among different CPUs. Only the signal capture processes have to be implemented on the computer where the corresponding acquisition hardware actually is. This is the reason why the *RMEAlsaCapBlock* client, designed for the audio channels acquisition, and the *capture* client, designed for the acquisition of the video stream showing the real Smart-Room scene, have to be implemented, respectively, on server *s4* and *s2* (as it can be inferred from the captions of the related clients illustrated in the map). As can be seen from the SmartFlow map, the data output flow from the *RMEAlsaCapBlock* client reaches two different clients: the *RMEAudioBlock_24c* is designed for the pre-processing of the audio data successively used by *acousticLoc* for the execution of the Acoustic Source Localization (ASL) algorithm. The other client receiving the *RMEAlsaCapBlock* output is the *RMEChannelExtraction* client: as the audio capture client returns a single data flow that multiplexes the audio from all the microphones, the audio channels extraction performed by *RMEChannelExtraction* is needed in order to demultiplex the flow in its distinct audio channels, which then can be analysed separately. As it already metntioned, for AED only one audio channel related to microphone #18 of the smart-room is exploited. That is why only eighteenth output (illustrated in the SmartFlow map as small blue rectangle) of the *RMEChannelExtraction* client is fed as input to the successive stage: the *ResampleClient*. This stage is responsible for downsampling of the input audio signal from the sampling frequency of 44.1 kHz applied by the Smart-Room audio capture devices to 16 kHz. The downsampled signal is forwarded to the *ASR_AED* client that performs the extraction of the set of audio spectro-temporal features. Audio data is taken from the continuous audio using equal-sized chunks, each one 1024 audio samples long. For every chunk the features are extracted from 30-*ms*-long frames, that is, each frame includes 480 audio samples of the 16 *kHz* input audio signal. The overlapping between adjacent frames is the same as in the offline feature extraction, that is, 20 *ms* long or, equivalently, 160 audio samples. Once the features are extracted, it is possible to perform Viterbi decoding. The HMM models, are loaded only once, at the initialization of the *ASR_AED* client, and the detection process is performed by the same client.

The flow containing the decisions from *ASR_AED* is the input of the *AEDTracking* client, which is designed for the fusion of acoustic event detection and acoustic localization data. The output flow from the *AEDTracking* client, containing data about the detected AE(s) and estimated position(s) inside the room is finally fed to the *DisplayAcousticEvents_with_Video* client showing the real scene shot by one of the smart-room cameras. The video stream comes from *capture* client, that is one of the inputs of the *DisplayAcousticEvents_with_Video* client. *DisplayAcousticEvents* client also shows the real-time display of the detected acoustic events by means of animated drawings that can be intuitively related to the corresponding AE. The additional functionality that provides this client is recording of a video file with animation and the associated audio stream that made possible to record 2 min demo. The whole SmartFlow AED system can also work offline due to the *readAudio* client, shown in Figure 7.4.1 disconnected from any flows. In order to run the system offline, it is sufficient to connect the *readAudio* client (instead of the *RMEAlsaCapBlock*) to the *RMEChannelExtraction* client and define an input audio file for each of the audio channels to be used.

7.5 Experimental Results

In order to prove the adequateness of the proposed approach, a series of experiments have been conducted to compare the implemented AED system working online with the baseline offline system and the results are presented in Table 7.5.1. In both online and offline tests the first column corresponds to the detection accuracy of the isolated acoustic events and the second one corresponds to AEs overlapped with speech.

In our experiments we used 8 sessions of isolated acoustic events from the database described in Appendix A. Additionally, these sessions were artificially overlapped with speech with different SNRs: -10 dB, 0 dB and +10 dB. For both offline and online tests, seven sessions (from 2 to 8) were used for training, and the remaining session 1 for testing.

Table 7.5.1. Comparison of the recognition results between offline and online AED systems.

AEs	Offline system		Online system	
	Iso, %	Overl, %	Iso, %	Overl, %
ap	100	100	92	84
cl	100	100	85	89
cm	97	97	64	65
co	67	95	87	75
ds	83	100	84	84
kj	100	100	97	93
kn	100	95	52	72
kt	67	100	85	86
pr	100	96	92	97
pw	64	86	74	73
st	80	82	75	70
Average	91.3 %		80.6%	

The main difference between the online and the offline tests is in the way of processing of the input waveform. During the offline tests the entire session is available for Viterbi segmentation. In this case the only parameter for tuning is the word insertion penalty parameter (*p*-value) that is the kind of trade-off between misses and false alarms. In our experiments p = -200. In online tests the recognition is performed on a frame-by-frame basis using the additional technologies described in sub-Section 7.3.2: silence detector, finite state machine, etc. In that case more parameters have to be tuned: the silence threshold, the number of silence frames, the confidence thresholds for each AE, etc. Note that in online tests the output hypothesis labels are those that displayed by the visualization block.

As can be seen from Table 7.3.2, almost all AEs are well detected in offline simulations. Relatively low detection rate corresponds to low-energy AEs, such as “Keyboard typing”, “Paper work” and “Steps”; additionally, the AE “Cough” is often confused with speech. In online simulations the best detection rate is achieved for such AEs as “Applause”, “Cup clink”, “Key jingle” and “Phone ring”, but AEs as “Door knock” and “Chair moving” showed relatively low detection rates. Actually, these AEs have the shortest duration in the testing database employed in the experiments. The obtained absolute difference between online and offline recognition results is less than 11% in terms of conventional AED-ACC metric.

7.6 Chapter Summary

In this chapter we described a 2-source acoustic event detection and localization system running in real-time in the UPC's smart-room. The detection of AEs is performed using our HMM-GMM approach, which allows analyzing the input waveform on a frame-by-frame basis with low latency. The AED and ASL systems are visually monitored by a GUI application which shows the output of AED and ASL technologies in real-time. Using this application, a video recording has been captured that contains the output of the GUI during a session lasting about 2 min, where three people in the room speak, interact with each other in natural way producing AEs which may overlap with speech. The implementation of AED components using Smart-audio++ package has been reviewed. The experimental results show the absolute difference between online and offline recognition results is less than 11% in terms of conventional AED-ACC metric.

Chapter 8. Conclusions and Future Work

8.1 Summary of Conclusions

This thesis presents the work done by the author in the area of Acoustic Event Detection (AED) focusing on: 1) multimodal techniques to deal with the difficult problem of signal overlaps present in meeting-room acoustic signals, and 2) audio segmentation techniques in the broadcast news domain. The HMM-GMM classifier is chosen as the basic detection technique to perform both offline experiments and implementation in real-time of a 2-source acoustic event detection system.

There are several contributions of this thesis. Regarding feature extraction, firstly, the use of video features, which are new for the meeting-room AED task. A number of features were extracted from video recordings by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of 5 classes of AEs. Since the video modality is not affected by acoustic noise, the proposed features improved AED in both isolated and spontaneous scenario recordings. Secondly, the inclusion of acoustic localization features, which, in combination with the usual spectro-temporal audio features, yielded further improvements in recognition rate. The following meta-classes were defined based on the acoustic localization information: “near door” and “far door”, related to the distance of the acoustic source to the door; and “below table”, “on table” and “above table”, related to the *z*-coordinate of the detected AE.

Two different strategies for fusion of audio and video modalities have been employed in this thesis: feature-level fusion is based on concatenating feature vectors from different modalities into one super vector; and decision-level fusion, where each modality acts as an independent “expert”, giving its opinion about the unknown acoustic event (AE). Decision-level fusion is carried out using weighted arithmetical mean (WAM) and fuzzy integral (FI) approaches. Unlike non-trainable fusion operators (mean, product), the statistical WAM and FI approaches avoid the assumption of equal importance of information sources. We demonstrated that the FI fusion operator can capture interactions among the various modalities. Additionally, the fuzzy measure, which is associated with the fuzzy integral, can be used to measure the importance for each information source for detecting particular AEs.

Taking into account that the task AED in meeting room environments is relatively new, there is a lack of annotated multimodal data, in particular data with temporal overlaps of sounds which is needed for training and testing the proposed technologies. In total, about 3 hours of new data with AEs were recorded in the UPC multimodal room from 5 video cameras and 24 microphones. The database includes two kinds of datasets: recordings of isolated AEs (2 hours), where several participants performed each AE several times, and a more spontaneously generated dataset (1 hours) which consists of 9 scenes of about 5 min long with 2 participants that interact with each other in a natural way: drink coffee, speak on the mobile phone, etc. Manual annotation of the data has been done to get a reliable performance evaluation. In order to encourage other researchers to work on this multimodal AED field, these datasets are made publicly available.

Another contribution of this thesis is in the area of feature selection. Taking into account the large computational load of the conventional wrapper approaches, a fast one-pass-training technique has been proposed that does not require the re-estimation of acoustic models during the evaluation of the candidate feature sets. The benefit in terms of computational speed of this technique becomes extremely important when the initial feature set is large and the training of acoustic models is time consuming. The experimental results show that the fast technique can be successfully applied in our multimodal audio recognition tasks. Three different feature selection approaches have been compared in the framework of multimodal acoustic event detection and broadcast news audio segmentation. The wrapper-based sequential backward selection approach showed the best results in terms of both accuracy and feature reduction ratio at the cost of a high computational complexity. The hybrid two-stage PCA algorithm show slightly worse results in terms of accuracy, but much better performance in terms of speed. Note that for PCA feature selection the extraction of the whole set of initial features is required, which may produce a remarkable latency during the testing phase of the detection/segmentation algorithm. This issue could be crucial in embedded and online systems as the extraction of unimportant features may be computationally expensive. Besides that, the physical meaning of the selected features is lost so that the PCA approach does not give any feedback about the suitability of a particular feature to the detection/segmentation task. The performance characteristics of the learning algorithms have been analyzed using the bias-variance decomposition of the classification loss. For mean-square error and zero-one loss this decomposition has very simple form: the error is composed of two non-negative terms, one comes from classifier imperfectness as well as intrinsic error of domain, and the other comes from fluctuation of classifier prediction over different possible training sets. We observed that for all acoustic classes the bias of the GMM-based classifier increases after reaching some point, while, in general, the variance of the learner monotonically increases with increasing number of features. The most accurate feature set corresponds to the best bias-variance trade-off point for the learning algorithm. Using the multimodal database of acoustic events we found that both static and dynamic spectro-temporal features contribute to the final accuracy. According to the obtained results, the video features are an important additional source of information for detection of the five AEs for which video features are extracted. The acoustic localization feature was selected only for some AEs.

The problem of audio segmentation in broadcast news domain has been also tackled in this work, and two hierarchical AS systems have been developed. The hierarchical system architecture is a group of detectors (called modules), where each module is responsible for detection of one acoustic class of interest. As input it uses the output of the preceding module and has 2 outputs: the first corresponds to audio segments detected as the corresponding class of interest, and the other is the rest of the input stream. In this type of architecture, it is not necessary to have the same classifier, feature set and/or topology for different detectors. Two different AS systems has been developed using two broadcast news databases: the first one includes audio recordings from the TV program Àgora from the Catalan TV3 channel and the second one includes audio from the 3/24 Catalan TV channel. The output of the first AS system was used in automatic translation and subtitling application developed for Tecnoparla project, that also demonstrates other speech and language technologies: speaker diarization, language recognition, speech recognition, speech translation and text-to-speech syntesis. The first pre-processing step in the presented system is audio segmentation. The second HMM-GMM-based AS system with feature selection got competitive results in the Albayzín-2010 audio segmentation evaluation.

Taking into account the increasing interest in the problem of audio segmentation from the one hand, and the existence, from the other hand, of a rich variety of feature extraction approaches and classification methods, the Albayzín 2010 evaluation of audio segmentation systems was organized. The Albayzín evaluation campaign is an internationally-open set of evaluations organized by the Spanish Network of Speech Technologies (RTH) every 2 years. Actually, the quantitative comparison and evaluation of competing approaches is very important in nearly every research and engineering problem. The evaluation campaigns that independently compare systems from different research groups help to determine which directions are promising and which are not. The evaluation setup, including the database, definition of acoustic classes and the segmentation metric were proposed in the framework of this thesis. The results from 8 participants from Spanish and Portuguese universities were compared and reported. After analyzing both systems and results, some main conclusions have been outlined.

Real-time processing is a requirement for many practical signal processing applications. In this thesis we implemented online 2-source AED and acoustic source localization (ASL) algorithms in a smart-room, a closed space equipped with multiple microphones. The AED and ASL systems are visually monitored by a GUI application which shows the output of AED and ASL technologies in real-time. Acoustic event detection is based on HMM-GMMs, that enable to process the input audio signal with low latency. The experimental results from online tests show promising recognition accuracy for most of AEs both isolated and overlapped with speech. A graphical user interface that shows 2-source AED and ASL functionalities working together is currently running in real time in the smart-room. Using this application, a video recording has been captured that contains the output of the GUI during a session lasting about 2 min, where three people in the room speak, interact with each other or produce one of the 12 isolated as well as overlapped with speech meeting-room AEs.

8.2 Future Work

The following list contains the most important points requiring improvements as well as a few directions for future work.

8.2.1 Detection of novel objects

Machine learning approaches in audio recognition create acoustic models using training data sampled from the application domain as well as prior knowledge about the problem. During testing these trained models are applied to new data in order to estimate the acoustic class. An implied assumption is that the future is stochastically similar to the past. This approach fails when the system is confronted with situations that are not anticipated from the past experience. Novelty detection refers to the recognition of unknown (or novel) data, i.e. data which differ considerably from the ones that the system processed during training. It is a fundamental requirement for a good machine learning system to automatically identify data from regions not covered by the training data since in this case no reasonable decision can be made. The recently introduced theory of incongruence [WHZ08] allows for detection of unexpected events in observations via disagreement of classifiers at different levels of classifier hierarchy. Several possibilities of how incongruence can appear from the point of view of class-membership hierarchy are discussed.

Many previous works exploit novelty detection for monitoring and surveillance applications to identify hazardous events. However in the meeting-room domain the problem is not sufficiently addressed. It is common experience that in meeting-room environments out of the dictionary acoustic events appear. The examples are occasionally falling objects during presentations, human scream from outside the room, giving a flick by some participants etc. Note the online 2-source AED system implemented in UPC's smart-room presented in chapter 7 includes "unknown" output symbolized with "?". It appears when the AED algorithm does not have enough confidence to assign a detected non-silent event to one of defined 12 classes. However the "uknown" acoustic event is not evaluated in final experiments due to the fact that it is not feasible to obtain a wide variety of data, which are representatives of "unknown" events. This issue can be addressed in future work.

8.2.2 Multi-microphone approach to deal with the problem of signal overlaps

Detection of acoustic events (AED) that take place in meeting-rooms environment becomes a difficult task when signals show a large proportion of temporal overlap of sounds, like in seminar-type data, where the AEs often occur simultaneously with speech. Several evaluation campaigns report low detection accuracies in such environments. In the presented thesis we proposed different strategies to deal with this problem. First is using additional video modality that is less sensitive to the overlap phenomena present in the audio signal to improve the baseline recognition rate. Second, the overlap problem has been dealt at feature level: several features coming from acoustic localization system showed to be useful in meeting-room scenario AED task. Third, the overlapping problem has been addressed at the level of models by modelling the possible combinations of sounds by classifiers in the online 2-source AED system.

A possible improvement can be also achieved at the signal level using source separation techniques like independent component analysis (ICA). Note ICA was originally developed to deal with problems that are closely related to the cocktail-party problem [WB06]. Using multiple channels is also motivated by the fact that in most biological systems perform sound analysis using several binaural cues: interaural time differences, interaural intensity densities and interaural spectral densities.

Moreover, at the level of decisions different weights can be assigned within multi-microphone system architecture to particular microphones to improve the robustness of AED. Another approach lies in selecting the best microphone in terms of recognition accuracy, for instance, using the likelihood at the decoder output. An alternative way of approaching microphone selection is based on measures or features extracted from the signals corresponding to the various microphones using different strategies: selection based on room impulse response related measures, selection based on position and orientation and selection based on signal distortion [WN10] etc.

8.2.3 Extending multimodal AED to more classes

New video technologies can be used for further improvement of multimodal AED. In this thesis we limited the video feature extraction to 5 classes: steps, chair moving, paper wrapping, keyboard typing and door slam. However for other AEs the video counterpart can be also taken into consideration. Object detection video technology can be used for analysis of the three-dimensional regions of interest in the room. For instance, the localization of the cell-phone in the room may be useful for detection of “phone-ring”, detection of cup may be useful for “cup clink”, detection of keys for “key jingle” etc.

Another useful technology is face detection. It can be used for further analysis of human facial expressions for detecting of laugh AE. The human siluette reconstruction technology may allow reconstructing the movement of human hands that may be beneficial to detect the “applause” AE or “cough” (under the assumption of polite environment).

8.2.4 AED for Automatic Speech Attribute Transcription

Automatic Speech Attribute Transcription (ASAT) task approaches speech recognition from linguistic perspective by means of detection acoustic and auditory cues, weighting, combining and processing them until consistent speech understanding is achieved [LCD07]. This detection-based ASR approach allows incorporating expert knowledge of linguistics and acoustic phonetics into speech recognition systems and can be considered as alternative to classical ASR paradigm. The ASAT front-end processing assumes a bank of detectors of useful and meaningful attributes of speech signal. The design of these detectors is critical problem for the detection-based paradigm. We expect that the proposed HMM-GMM and SVM detection approaches developed for the meeting-room AED as well as statistical fusion techniques can be successfully applied in ASAT task.

8.2.5 Cross-site event detection

In the presented experiments the recognition accuracy for most of isolated AEs is higher than 90% being 98% for some of them. This creates a certain barrier for further development and improvement of AED technologies since the actual baseline recognition rate is high. One possibility for further research is working with signal overlap problem and under noisy conditions. However creating the database with signal overlaps produced in natural way is difficult problem. Moreover, creating the database with artificially overlapped signals (using superposition of different signals recorded separately) may be far away from real audio.

A possible direction of further research could be the cross-site event detection, i.e. the case when acoustic models are created using the database from one site and testing is performed using the database from another site. This is natural requirement for many practical applications to work equally well in different conditions and environments. In fact, within CHIL European project different databases with AEs have been recorded from UPC, UKA, IBM, AIT and ITC sites [WS09] that can be used in experiment. Preliminary experiments performed in mismatched conditions show that recognition rate diminishes for current approaches.

8.2.6 Factor analysis for audio segmentation

The recent Albayzín audio segmentation evaluation showed that there is still a large margin for improvement of the segmentation results. Only 23% of errors produced by the best AS system were also produced by all the other AS systems. Since the main source of mistakes are confusions between "Music" and "Speech over music", between "Speech over music" and "Speech over noise", and also between "Speech" and "Speech over noise", future research efforts should be devoted to improved detection of acoustic classes with different acoustic environment in background.

Taking into account the recent success of joint factor analysis (JFA) [KBD05] and total variability [DDK09] approaches for language/speaker identification tasks, those approaches can be applied for the problem of audio segmentation. JFA approach tries to determine low dimension subspaces of the high-dimensional feature space that cover most of the inter-session variance and most of the inter-class variance. Once these sub-spaces are identified the acoustic class can be detected using the information in the inter-class variability sub-space.

Total-variability [DDK09] approach does not try to segregate inter-class and inter-session variability sub-spaces but finds a sub-space that covers most of the variability (both acoustic class and inter-session) by means of Principal Component Analysis (PCA). The audio segmentation problem can be addressed using feature vectors from this new subspace and several normalization/compensation techniques.

Appendix A. UPC-TALP Multimodal Database of Isolated and Spontaneous Acoustic Events

Introduction

This database contains 2 types of multimodal recordings (hereafter *S-recordings* and *T-recordings*) of AEs that occur in a meeting room environment. The *S-recordings* correspond to isolated sounds that do not have temporal overlaps; within the same class the AE instances have approximately the same length and only one person per session was acting during recordings. The *T-recordings* correspond to the set of spontaneous AEs that occur in more realistic conditions. These recorded sounds may contain temporal overlaps; they have different length and were produced by 2 interacting persons per session. The database can be used as a training material for AED task as well as for testing AED algorithms in quite and noisy environments with or without temporal sound overlaps.

Description of the acoustic events

For recording, we used the same list of sounds that was defined in CHIL [WS09] with conventional labels[2]:

Acoustic event	Label
o Knock (door, table)	kn
o Door slam	ds
o Steps	st
o Chair moving	cm
o Spoon (cup jingle)	cl
o Paper work (listing, wrapping)	pw
o Key jingle	kj
o Keyboard typing	kt
o Phone ringing/Music	pr
o Applause	ap
o Cough	co
o Speech	sp

Description of the recording setup

The whole database was recorded using the following audio equipment: six T-shape clusters (4 microphones per cluster). In total 6*4=24 microphones. For video recordings 5 video cameras in

[2] The class “laugh” was substituted by class “speech”.

different positions of the room were used. The positions of the microphones and video cameras in the UPC Smart-Room are described in Figure A.1. Figure A.2 describes the configuration of the T-shaped clusters. Audio data was recorded at 44.1kHz, 24-bit precision, and then packed in *.wv format (WavPack format that provides lossless compression). All the channels were synchronized.

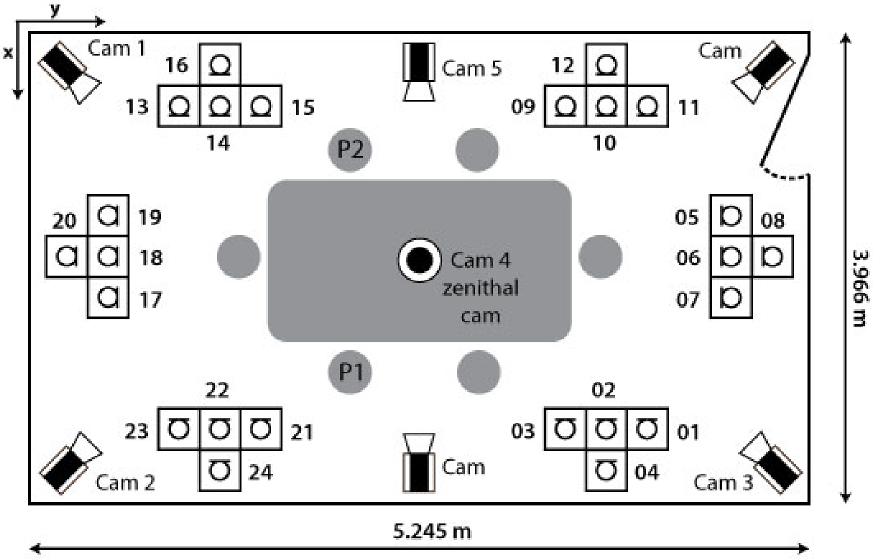


Figure A.1. Microphone and camera positioning in the UPC smart-room

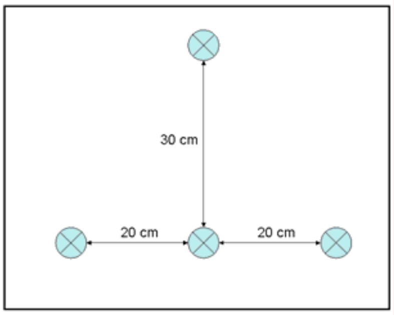


Figure A.2. Configuration & orientation of the T-shaped microphone clusters

Description of the recorded database

The number of sounds per each sound class is about 100-400 instances. 6 people participated in recordings. During each *S-recording*, the participant took a position P1 (Figure A.1). In the case of

T-recording the participants took position P1 and P2. Both positions are marked in Figure A.1. The exact coordinates of these positions are given in Table A.1.

During each *S-recording* a person had to produce a complete set of sounds 10 times. A script indicating the order of events to be produced was given to each participant. The participant was allowed to change the order of AEs. Almost each event was followed and preceded by a pause of several seconds. All sounds were produced individually. During *T-recording* 2 persons had to make a short meeting (around 5 minutes), discussing certain subject. The approximate script indicating the scenario of meeting was given to read to each participant before recording. The participants were allowed to make improvisations.

Table A.1. X, Y coordinates of the positions of participants

Position	**X (meters)**	**Y(meters)**
P1	1.28	2.40
P2	3.03	2.40

Annotation of the database

The annotation was done manually by listening at signals from a single channel (the 7th channel). The following criterion was used during the annotation. If an event of class *X* includes a pause of minimum 300ms and both parts of the event, the one before the pause and the one after the pause, can be (subjectively) assigned a label *X*, then the event is annotated as two separated events of class *X*. If either the pause length is less than 300ms or the first/second part of the event is not recognizable without hearing the other part, the whole event is marked as only one event of class *X*.

Content of the distributed database

The database that is distributed in 6 DVDs contains signals corresponding to 24 audio channels, 5 video channels and the corresponding labels. The splitting of the data for purposes of training and testing can be based on either participants, sessions or type of recordings (*S-recordings* or *T-recordings*). To produce the DVDs, we distributed the sessions in the most compact way in order to make the minimum number of DVDs (in our case 6). Table A.2 shows the distribution of the audio and video material among the sessions. The distribution sessions among DVDs is indicated in Table A.3.

Table A.2. Number of annotated acoustic events in each session

Event type	S01	S02	S03	S04	S05	S06	S07	S08	T01	T02	T03	T04	T05	T06	T07	T08	T09	TOTAL
Knock (door, table), **<kn>**	9	8	10	10	10	8	11	13	2	3	2	3	3	4	2	5	3	**106**
Door slam, **<ds>**	17	15	19	20	40	37	56	52	8	11	8	9	10	8	10	10	8	**338**
Steps, **<st>**	10	10	8	23	43	34	28	50	15	17	12	18	20	21	16	17	17	**359**
Chair moving, **<cm>**	19	37	32	22	23	38	34	40	17	21	15	20	22	24	15	23	26	**428**
Spoon (cup jingle), **<cl>**	10	11	13	11	10	15	11	15	5	3	8	4	4	6	2	11	5	**144**
Paper work (listing, wrapping), **<pw>**	9	11	10	8	17	12	12	12	7	6	9	18	10	18	17	25	36	**237**
Key jingle, **<kj>**	11	11	11	8	0	13	10	18	1	6	1	4	2	9	4	7	7	**123**
Keyboard typing, **<kt>**	10	10	13	12	10	13	10	11	8	9	6	9	8	12	10	11	8	**170**
Phone ring-ing/Music, **<pr>**	11	18	11	14	8	11	13	15	4	4	4	4	4	0	3	4	2	**130**
Applause, **<ap>**	9	5	9	11	12	9	14	14	1	0	1	1	1	1	2	1	1	**92**
Cough, **<co>**	10	10	12	13	9	13	11	12	7	3	2	1	4	2	1	3	1	**114**
Speech, **<sp>**	0	0	0	0	8	20	12	34	27	33	36	31	41	46	41	0	0	**329**

The name of a audio file is NAME_N.wv, where NAME is the name of session; N is the number of the microphone, The sampling frequency is 44100 Hz, the number of bits per sample is 24. The name of video file is camN.avi, where N is the number of video camera. The name of an annotation file is <session_name>.csv (e.g. S02.csv – session 2 of isolated AEs). The format of its content is analog to that of the AGTK ".csv" format, i.e. "start_ts, end_ts, event_id", where the labels start_ts, end_ts, event_id denote the starting time stamp (from he beginning of the file), the ending time stamp, and the event label, respectively. The time stamps are given in seconds from the beginning of the file.

The structure of the DVD_N (where N denotes the DVD number) is:

```
/<session_name>

    /audio                     // audio recordings from T-shape microphones
    <session_name>_01.wv
    <session_name>_02.wv
      ...
    <session_name>_24.wv
    /video                     // video recordings from 5 cameras
      cam1.avi
      cam2.avi
      cam3.avi
      cam4.avi
      cam5.avi
```

```
        <session_name>.csv          // file with annotations
/license_agreement.pdf              // license agreement
/readme.txt                         // brief introduction
```

Table A.3. Distribution of sessions among DVDs

DVD number	Sessions
1	S01, S02
2	S03, S04
3	S05, S06
4	S07, T02, T03, T04
5	S08, T01
6	T05, T06, T07, T08, T09

Script of S-recordings

Scenery: Smart-room closed; a laptop, papers, keys, a cell phone, a spoon/cup are on the chair. Nobody is in the room. Duration is about 15-20min. Table A.4 summarizes the script:

Table A.4 Script of S-recordings

Phases	**Sound producing action**	**Number of repetitions**	**Estimated duration (minutes)**
Entrance	Knock the door	10	3
	Open and close the door	10	
Stepping & Sitting	go to the different chairs loudly and sit down, stand up.	10	
		10	
Producing noises	put the spoon into the cup and stir up an imaginary sugar	10	12
	Take out the keys from a pocket and put them on the table / move the keys from one place to another on the table/ put them back into the pocket	10	
	keyboard typing	10	
	Phone rings (different melodies) & Speech	10	
	Take some papers from the table and count them in hand / even the papers by knocking them to the table	10	
	Applause (of several people)	10	
	Cough	10	

Standing up	stand up, move the chair changing sitting position	1	1
Stepping	from the chair loudly	1	
Exit	Open and close the door	1	

Script of T-recordings

Scenery: Smart-room is closed; a keyboard, papers, keys, a cell phone, a spoon/cup on the chair. Duration is about 5 min. 2 persons participate in it: person **A** and person **B**. At the beginning of session nobody is in the room. The approximate scenario of recordings is following (improvisation was welcomed):

- <Door knock> A is entering the room<door slam>.
- A moves <steps> towards the chair and takes a seat <chair moving>.
- A makes a cough <cough>. He takes a paper from the table <paper wrapping> and reads it.
- B is knocking the door <knock>.
- A says: “Yes, come in, please”.
- B enters the room <door slam>.
- A stands up <chair moving> and moves towards B <steps> to great him and says something like: “Hello, how are you? Nice to see you!”
- B: “I’m fine, thanks. …“!
- A: “Oh, it’s good that you’ve come. I have something interesting for you. Come here” <steps>, <cough>.
- B: “Did you get cold?”
- A: “Yes, a little bit. Yesterday at night I drank several bottles of cold beer” <cough>.
- A: “Please, take a seat”. <chair moving>. B during the sitting puts the keys on the table. <key jingle>.
- A: “This is a paper about Multimodal Acoustic Event detection. You could be interested in it”. <paper work>
- B: “Oh, what is the conference? Yes, yes I see. It could be helpful for me”.
- A: “Do you want coffee? I can prepare it right now…”.
- B: “Yes, with much pleasure, thanks”.
- A: “Just a moment!”. A is leaving the room, preparing coffee <steps> <door slam>.
- B: <typing something on the laptop>

- A is outside the room now. He rings to B <phone ringing>. He rings 2 times.

- A enters the room <steps><door slam> "Please, take a coffee"

- B: "Thanks!" "I was just looking for this paper through the internet".

- A and B start drinking a coffee. <cup clink>.

- B likes the coffee and says "Great coffee!" and makes greeting applauses <applause>.

- B: "Ok, I should go, see you later", B takes the keys from the table. <key jingle>

- A: "Bye" <steps>, <door slam>.

User guidelines

- All sounds should be done with 2-5 second pauses in between, e.g. you make a cough, wait 2-5 seconds, then make another cough and wait 2-5 seconds, etc.
- All sounds should be done in various manners, i.e. please, do not do the same sounds several times; e.g. if you knock the door, do not knock it exactly 3 times with the same speed

Entrance

- Knock the door, pause 2-5 second, open the door, enter the room and close the door. Repeat it 10 times.

Stepping, sitting down, standing up.

- Go to different chairs loudly and wait several seconds before sitting and standing up. Repeat it 10 times.

Producing noises

- Put the spoon into the cup and stir up an imaginary sugar (10 times with 2-5 second pauses in between)
- Make key jingling sounds like this: take out the keys from a pocket and put on the table - move the keys from one place to another on the table- put back into the pocket (Each of the ways is counted as one key jingle – you have to have 2 of them)
- Keyboard typing. Type on the keyboard for 2 seconds. Do it 10 times with 2-5 second pauses.
- Your cell phone is on and the volume is maximum, somebody calls your cell phone 10 times.
- You have to do some paper work - do one of the following: take some papers from the table and count them in hand / even the papers by knocking them to the table (each of it is counted as one paper work – you have to produce 2 of them with 5 second pauses inside)
- Make applause 10 times (don't forget to do it differently)
- Make cough 10 times (don't forget to do it differently)

Sitting

- Here you have to do 1 chair moving while standing up

Stepping

- Go from the chair to the door loudly and wait some seconds.

Own publications

- T. Butko, C. Nadeu, "Audio segmentation of broadcast news in the Albayzín-2010 evaluation: overview, results, and discussion", *EURASIP Journal on Audio, Speech, and Music Processing*, 2011, in print.

- T. Butko, C. Canton-Ferrer, C. Segura, X. Giro, C. Nadeu, J. Hernando, J.R. Casas, "Acoustic event detection based on feature-level fusion of audio and video modalities", *EURASIP Journal on Advances in Signal Processing*, vol. 2011, 2011.

- T. Butko, C. Nadeu, "On building and evaluating a broadcast-news audio segmentation system", in Proc. Interspeech, 2011 (to appear)

- T. Butko, F. Gonzalez Pla, C. Segura, C. Nadeu, J. Hernando, "Two-source acoustic event detection and localization: online implementation in a smart-room", in Proc. European Signal Processing Conference (EUSIPCO-2011), Barcelona, Spain, 2011 (to appear)

- T. Butko, C. Nadeu, "Audio Segmentation of Broadcast News: A Hierarchical System with Feature Selection for the Albayzín-2010 Evaluation", *IEEE ICASSP*, 2011.

- T. Butko, C. Nadeu, "A fast one-pass-training feature selection technique for GMM-based acoustic event detection with audio-visual data", *Proc. Interspeech*, 2010.

- T. Butko, C. Nadeu, "On Enhancing Acoustic Event Detection by Using Feature Selection and Audiovisual Feature-Level Fusion", *Workshop on Database and Expert Systems Applications, DEXA*, Bilbao, pp.271-275, 2010.

- T. Butko, C. Nadeu, "Detection of Overlapped Acoustic Events using Fusion of Audio and Video Modalities", *VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop*, Vigo, 2010.

- T. Butko and C. Nadeu, "A Hierarchical Architecture with Feature Selection for Audio Segmentation in a Broadcast News Domain", *VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop*, Vigo, 2010.

- T. Butko, C. Nadeu, H. Schulz, "Albayzín-2010 Audio Segmentation Evaluation: Evaluation Setup and Results", *VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop*, Vigo, 2010.

- T. Butko, C. Canton-Ferrer, C. Segura, X. Giro, C. Nadeu, J. Hernando, J.R. Casas, "Improving Detection of Acoustic Events Using Audiovisual Data and Feature Level Fusion", *Proc. Interspeech*, pp. 1147-1150, 2009.

- C. Canton-Ferrer, T. Butko, C. Segura, X. Giro, C. Nadeu, J. Hernando, J.R. Casas, "Multimodal Acoustic Event Detection Towards Scene Understanding ", *Proc. of IEEE Workshop on Human Communicative Behavior Analysis within the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

- M. Aguilo, T. Butko, A. Temko, C. Nadeu, "A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task", *I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, Porto Salvo, Portugal*, 2009.

- T. Butko, A. Temko, C. Nadeu, C. Canton, "Inclusion of Video Information for Detection of Acoustic Events using the Fuzzy Integral", *Machine Learning for Multimodal Interaction*, LNCS, vol. 5237/2008, pp. 74-85, Springer, 2008.

- T. Butko, A. Temko, C. Nadeu, C. Canton, "Fusion of Audio and Video Modalities for Detection of Acoustic Events", *Proc. Interspeech,* pp. 123-126, 2008.

Bibliography

[ALM03] N. Adami, R. Leonardi, P. Migliorati, “An overview of multi-modal techniques for the characterization of sport programmes”, *Proc. of SPIE-VCIP*, pp. 1296-1306, 2003.

[AMB03] J. Ajmera, I. McCowan, H. Bourlard, “Speech/music segmentation using entropy and dynamism features in a HMM classification framework,” *Speech Communication*, vol. 40, no. 3, pp. 351–363, 2003.

[AMG02] M. Arulampalam., S. Maskell, N. Gordon, T. Clapp, “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking”, *IEEE Trans. on Signal Processing*, vol. 50, 174-188, 2002.

[AMI] http://corpus.amiproject.org/

[Ara08] G. Aradilla, *Acoustic Models for Posterior Features in Speech Recognition*, PhD Thesis, EPFL, Lausanne, 2008.

[BAK11] J.-H. Bach, J. Anemuller, B. Kollmeier, “Robust speech detection in real acoustic backgrounds with perceptually motivated features”, *Speech Communication*, 2011.

[BAL05] M. Büchler, S. Allegro, S. Launer, N. Dillier, “Sound classification in hearing aids inspired by auditory scene analysis,” *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 18, pp. 2991-3002, 2005.

[BC00] L. Bu, T.-D. Chiueh, “Perceptual speech processing and phonetic feature mapping for robust vowel recognition”, *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 105-114, 2000.

[BD99] A. Bobick, J. Davis, “The recognition of human movement using temporal templates”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, v. 23, pp.257–267, 1999.

[BFM02] A. Bugatti, A. Flammini, P. Migliorati, “Audio classification in speech and music: a comparison between a statistical and a neural approach,” *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 4, pp. 372–378, 2002.

[CAW06] C.-Y. Chen, A. Abdallah, W. Wolf, “Audiovisual gunshot event recognition”, *IEEE International Conference on Systems, Man, and Cybernetics*, 2006.

[CC77] T.M. Cover J.M. V. Campenhout, “On possible orderings in the measurement selection problem”, *IEEE Trans. On Systems, Man and Cybernetics*, vol. smc-7, no. 9, pp. 657-661, 1977.

[CEY06] H.E. Çetingül, E. Erzina, Y. Yemeza, A.M. Tekalpa, “Multimodal speaker/speech recognition using lip motion, lip texture and audio”, *Signal Processing*, vol. 86, iss. 12, pp. 3549-3558, 2006.

[CG01] W. Chou, L. Gu, “Robust singing detection in speech/music discriminator design”, *Proc. IEEE ICASSP*, vol. 2, pp. 1331-1334, 2001.

[CG98] S. S. Chen, P.S. Gopalkrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion”, *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127-132, 1998.

[Che01] T. Chen, “Audiovisual speech processing”, *Signal Processing Magazine*, 2001.

[CKC07] G. Castellano, L. Kessous, G. Caridakis, “Multimodal emotion recognition from expressive faces, body gestures and speech”, *Proc. of the Doctoral Consortium of 2nd International Conference on Affective Computing and Intelligent Interaction*, 2007.

[CLE06] CLEAR, 2006. Classification of Events, Activities and Relationships. Evaluation and Workshop. <http://isl.ira.uka.de/clear06>.

[CLE07] CLEAR, 2007. Classification of Events, Activities and Relationships. Evaluation and Workshop. <http://www.clear-evaluation.org/>.

[CLH06] R. Cai, L. Lu, A. Hanjalic, H. Zhang, and L.-H. Cai, “A flexible framework for key audio effects detection and auditory context inference”, *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 3, pp. 1026–1039, 2006.

[CNJ08] S. Chu, S. Narayanan, C.-C. Jay Kuo, “Environmental sound recognition using MP-based features”, *Proc. ICASSP*, pp. 1-4, 2008.

[CNK09] S. Chu, S. Narayanan, C.-C. J. Kuo, “Environmental sound recognition with time-frequency audio features”, *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, is. 6, 2009.

[Com94] P. Comon, “Independent component analysis, a new concept?”, *Signal Processing*, vol. 36, no. 3, pp. 287-314, 1994.

[CS02] M. Cowling, R. Sitte, “Analysis of speech recognition techniques for use in a non-speech sound recognition system”, *Proc. of the 6th International Symposium on Digital Signal Processing for Communication Systems*, pp.16-20, 2002.

[CSC08] C. Canton-Ferrer, R. Sblendido, J.R. Casas, M. Pardàs, “Particle filtering and sparse sampling for multi-person 3D tracking”, *Proc. IEEE International Conference on Image Processing*, pp. 2644-2647, 2008.

[CXZ07] I. Cohen, Q. T. Xiang, S. Zhou, T. S. Huang, “Feature selection using principal feature analysis”, *Proc. of 15th Conf. on Multimedia*, pp.301-304, 2007.

[CZH98] M. T. Chan, Y. Zhang, and T. S. Huang, “Real-time lip tracking and bi-modal continuous speech recognition”, *IEEE Workshop on Multimedia Signal Processing*, pp. 65-70, 1998.

[DDK09] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, P. Dumouchel, “Support Vector Machines versus fast scoring in the low-dimensional total variability space for speaker verification”, *Proc. Interspeech*, 2009.

[DHS00] R. Duda, P. Hart, D. Stork, *Pattern classification (2nd Edition)*, Wiley-Interscience, 2000.

[DL00] S. Dupont, J. Luettin, “Audio-visual speech modeling for continuous speech recognition”, *IEEE Trans. Multimedia*, 2000.

[DM11] C. N. Doukas, I. Maglogiannis, “Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components”, *IEEE Trans. on Information Technology in Biomedicine*, vol. 15, no. 2, 2011.

[Dom00] P. Domingos, “A unified bias-variance decomposition and its applications”, *Proc. 17th International Conf. on Machine Learning*, 2000.

[DPK96] T. Dau, D. Puschel, A. Kohlrausch, “A quantitative model of the “effective" signal processing in the auditory system: I. Model structure”, *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615-3622. 1996.

[DPR09] P. Dhanalakshmi, S. Palanivel, V. Ramalingam, “Classification of audio signals using SVM and RBFNN”, *Proc. Expert Systems with Applications*, v. 36, issue 3, part 2, pp. 6069-6075, 2009.

[DSB01] J. Dibiase, H. Silverman, M. Brandstein, “Microphone arrays. Robust localization in reverberant rooms”, *Springer*, 2001

[EGR07] M. Exposito, G. Galan, R. Reyes, V. Candeas, “Audio coding improvement using evolutionary speech/music discrimination”, *Proc. IEEE Conference on Fuzzy Systems*, pp. 1-6, 2007.

[EJF06] H.K. Ekenel, Q. Jin, M. Fischer, R. Stiefelhagen, “Multimodal person recognition for human-vehicle interaction”, *IEEE Multimedia*, vol.13 (2), pp.18–31, 2006.

[FAL10] http://fala2010.uvigo.es/images/proceedings/index.html

[FD00] R. Frischholz, U. Dieckmann, “BioID: a multimodal biometric identification system”, *Journal IEEE Computation*, vol. 33 (2), pp. 64–68, 2000

[Fuj99] T. Fujishima, “Realtime chord recognition of musical sound: a system using common lisp music”, *Proc. of the Int. Computer Music Conference (ICMC)*, pp. 464–467, 1999.

[Fuk72] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, 1972.

[GE03] I. Guyon, A. Elisseeff, “An introduction to variable and feature selection”, *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[Ger02] D. B. Gerhard, “Perceptual features for a fuzzy speech-song classification”, *Proc. ICASSP*, vol. 4, pp. 4160-4163, 2002.

[GGN06] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, “Feature extraction, foundations and applications”, *Series Studies in Fuzziness and Soft Computing*, Physica-Verlag, Springer, 2006.

[GM07] X. Giró, F. Marqués, “Composite object detection in video sequences: applications to controlled environments”, *Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services*, pp. 1–4, 2007.

[GMT09] D. Gelbart, N. Morgan, A. Tsymbal, “Hill-climbing feature selection for multi-stream ASR”, *Proc. Interspeech*, 2009.

[GP06] H. Gunes, M. Piccardi, “Bi-modal emotion recognition from expressive face and body gestures”, *Journal of Network and Computer Applications*, 2006.

[Gra95] M. Grabisch, “Fuzzy integral in multi-criteria decision-making”, *Fuzzy Sets & Systems*, vol. 69, pp. 279-298, 1995.

[GS08] L. Golipour, D. O’Shaughnessy, “An intuitive class discriminability measure for feature selection in a speech recognition system”, *Proc. Interspeech*, pp. 1345-1348, 2008.

[Gun98] S. Gunn, *Support vector machines for classification and regression*, Technical report, ISIS, May 1998.

[HAA07] E.M. Hernandez, K. Adiloglu, R. Annies, H. Purwins, K. Obermayer, “Perceptual representation for classification of everyday sounds”, *Proc. of the Conference on Interaction with Sound*, vol. II, pp. 90-95, 2007.

[HDG06] M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, D. Zhang, “Multimodal integration for meeting group action segmentation and recognition”, in *Machine Learning for Multimodal Interaction*, LNCS, vol. 3869, pp. 52-63, 2006.

[Her90] A. Hermansky, “Perceptual linear predictive (PLP) analysis of speech”, Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738-1752, 1990.

[HKL97] L. Holmström, P. Koistinen, J. Laaksonen, and E. Oja. Neural and statistical classifiers--taxonomy and two case studies. *IEEE Transactions on Neural Networks*, 8(1):5-17, 1997.

[HL02] C. Hsu, C. Lin, “A comparison of methods for multi-class support vector machines”, *IEEE Trans. Neural Networks* (2002), pp. 415–425.

[HLW99] J. Huang, Z. Liu, Y. Wang, Y. Chen, E. K. Wong, “Integration of multimodal features for video scene classification based on HMM”, *IEEE Workshop on Multimedia Signal Processing*, pp. 53-58, 1999.

[HLW05] J. Huang, Z. Liu, Y. Wang, “Joint scene classification and segmentation based on hidden Markov model”, *IEEE Trans. on Multimedia*, vol. 7, iss. 3, pp. 538-550, 2005.

[HS98] M. A. Hall, L. A. Smith, “Practical feature subset selection for machine learning”, *Proc. of the 21st Australian Computer Science Conference*, Springer, pp. 181-191, 1998.

[IMK08] T. Izumitani, R. Mukai, K. Kashino, “A background music detection method based on robust feature extraction”, *IEEE ICASSP*, pp. 13-16, 2008.

[INM06] K. Ishizuka, T. Nakatani, Y. Minami, “Speech feature extraction method using subband-based periodicity and nonperiodicity decomposition”, *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 443-452, 2006.

[JDM00] A.K. Jain, R. P. W. Duin, J. Mao, "Statistical pattern recognition: a review", *IEEE Trans. on PAMI*, vol. 22, no. 1, pp. 4-37, 2000.

[JJK05] C. Jianfeng, Z. Jianmin, A. Kam, L. Shue, "An automatic acoustic bathroom monitoring system", *Proc. IEEE International Symposium on Circuits and Systems*, 2005.

[KBD03] Z. Khan, T. Balch, F. Dellaert, "Efficient particle filter-based tracking of multiple interacting targets using an MRF-based motion model", *International Conference on Intelligent Robots and Systems*, 2003.

[KBD05] P. Kenny, G. Boulianne, P. Dumouchel, "Eigenvoice modeling with sparse training data", *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp 345-354, 2005.

[Kle02] M. Kleinschmidt, *Robust speech recognition based on spectro-temporal processing*, PhD Thesis, 2002.

[Kle03] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition", *Proc. Eurospeech*, pp. 2573-2576, 2003.

[Ke08] Y. Ke, *Volumetric features for video event detection*, PhD Thesis, 2008.

[KJ97] R. Kohavi, G. John, "Wrappers for feature subset selection", *Artificial Intelligence, Spec. Issue on Relevance*, vol. 97, pp. 273-324, 1997.

[KQG04] S. Kiranyaz, A. F. Qureshi, M. Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval", *Proc. of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, pp. 55-62, 2004.

[KR05] R. Kaliouby, P. Robinson, "Generalization of a Vision-Based Computational Model of Mind-Reading", *Proc. of First International Conference on Affective Computing and Intelligent Interfaces*, pp 582-589, 2005.

[KS96] D. Koller, M. Sahami, "Toward optimal feature selection", *Proc. ICML*, pp. 284-292, 1996.

[Lan06] O. Lanz, "Approximate Bayesian multibody tracking", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 28(9), pp. 1439-1449, 2006.

[LAT08] J. Luque, X. Anguera, A. Temko J. Hernando, "Speaker diarization for conference room: the UPC RT07s evaluation system", *Multimodal Technologies for Perception of Humans*, LNCS, vol. 4625/2008, Springer Berlin/Heidelberg, pp. 543–553, 2008.

[LCC07] A. López, C. Canton-Ferrer, J. R. Casas, "Multi-person 3D tracking with particle filters on voxels", *IEEE ICASSP*, pp. 913-916, 2007.

[LCD07] C.-H. Lee, M.A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, L. Rabiner, "An overview on automatic speech attribute transcription (ASAT)", *Proc. Interspeech*, pp. 1825-1828, 2007.

[LH08] J. Luque, J. Hernando, "Robust speaker identification for meetings: UPC CLEAR-07 meeting room evaluation system", *Multimodal Technologies for Perception of Humans*, LNCS, vol. 4625/2008, pp. 266-275, 2008.

[LHH01] L. Lie, J. Hao, Z. H. Jiang, "A robust audio classification and segmentation method", *Proc. 9th ACM conference on Multimedia*, pp. 203-211, 2001.

[Lip91] R.P.Lippmann, "A critical overview of neural network pattern classifiers", *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 266-275, 1991.

[LSD01] D. Li, I.K. Sethi, N. Dimitrova and T. McGee, "Classification of general audio data for content-based retrieval", *Pattern Recognition Letters*, vol. 22, pp. 533–544, 2001.

[LSH97] G.. Langer M. Sams, P. Heil, H. Schulze, "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: evidence from magnetoencephalography", *Journal Computational Physiology, A: Neuroethol.*, pp 665–676, 1997.

[LT05] Z. Li, Y.-P. Tan, "Event detection using multimodal feature analysis", *Proc. ISCAS*, vol. 4, pp. 3845-3848, 2005.

[LT07] O. Lartillot, P. Toiviainen, “MIR in Matlab (II): A toolbox for musical feature extraction from audio”, *Proc. International Conference on Music Information Retrieval*, 2007.

[LV10] S. Lefèvre, N. Vincent, “A two level strategy for audio segmentation”, *Digital Signal Processing*, 2010.

[LXY06] S. Liu, M. Xu, H. Yi, L.-T. Chia, D. Rajan, “Multimodal semantic analysis and annotation for basketball video”, *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–13, 2006.

[LZJ02] L. Lu, H. Zhang, H. Jiang, “Content analysis for audio classification and segmentation”, *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[LZZ02] Q. Li, J. Zheng, Q. Zhou, “Robust endpoint detection and energy normalization for real-time speech and speaker recognition”, *IEEE Trans. on Speech and Audio Processing*, vol. 10, 2002.

[MC09] A. Munson, R. Caruana, “On feature selection, bias-variance, and bagging”, *ECML PKDD '09 Proc. of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*, 2009.

[MDV01] S. Moncrieff, C. Dorai, and S. Venkatesch, “Detecting indexical signs in film audio for scene interpretation”, *Proc. of IEEE ICME*, pp. 989-992, 2001.

[MJ08] R. Munkong, B.-H. Juang, “Auditory perception and cognition”, *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 98-117, 2008.

[MKP00] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, “Speech/music discrimination for multimedia applications”, *IEEE ICASSP*, vol. 6, pp. 2445–2448, 2000.

[MMS10] N. Misdariis, A. Minard, P. Susini, G. Lemaitre, S. McAdams, E. Parizet, “Environmental sound perception: metadescription and modeling based on independent primary studies”, *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 26 pages, 2010.

[NH02] M. R. Naphade, T. S. Huang, “Extracting semantics from audiovisual content: the final frontier in multimedia retrieval”, *IEEE Trans. on Neural Networks*, vol. 13, no. 4, 2002.

[NHA02] S. Nakamura, K. Hiyane, F. Asano, Y. Kaneda, T. Yamada, T. Nishiura, T. Kobayashi, S. Ise, H. Saruwatarig, “Design and collection of acoustic sound data for hands-free speech recognition and sound scene understanding”, *Proc. Multimedia and Expo*, 2002.

[NL05] T. L. New, H. Li, “Broadcast news segmentation by audio type analysis”, *in Proc. ICASSP*, vol. 2, pp. 1065-1068, 2005.

[NMH01] C. Nadeu, D. Macho, J. Hernando, “Frequency & time filtering of filter-bank energies for robust HMM speech recognition”, *Speech Communication*, vol. 34, pp. 93-114, 2001.

[NNM03] T. Nishiura, S. Nakamura, K. Miki, K. Shikano, “Environmental sound source identification based on hidden Markov models for robust speech recognition”, *Proc. Eurospeech*, pp. 2157–2160, 2003.

[NO99] H. Ney, S. Ortmanns, “Dynamic programming search for continuous speech recognition”, *IEEE Signal Processing Magazine*, vol. 16, pp. 64–83, 1999.

[OS97] M. Omologo, P. Svaizer, “Use of the crosspower-spectrum phase in acoustic event location”, *IEEE Trans. on Speech and Audio Processing*, v. 5:3, pp. 288–292, 1997.

[PA04] J. Pinquier, R. André-Obrecht, “Jingle detection and identification in audio documents”, *IEEE ICASSP*, vol. 4, pp. 329-332, 2004.

[Pal03] D. S. Pallet, *A look at NIST’s benchmark ASR tests: past, present, and future*, Technical Report, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 2003.

[Pap91] A. Papoulis, *Probability, random variables and stochastic processes*, McGraw-Hill Companies, p.126, 1991.

[Pee03] G. Peeters, *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*, CUIDADO Project Report, 2003.

[Pfe01] S. Pfeiffer, “Pause concepts for audio segmentation at different semantic levels”, *Proc. ACM International Conference on Multimedia*, pp.187-193, 2001.

[PH96] R.D. Patterson, J. Holdsworth “A functional model of neural activity patterns and auditory images”, *Advances in Speech, Hearing and Language Processing*, vol. 3, pp. 547–563, 1996.

[Pir04] S. Piramuthu, “Evaluating feature selection methods for learning in data mining applications”, *European Journal of Operational Research*, vol. 156, no. 2, pp. 483–494, 2004.

[PHC10] M. Paleari, B. Huet, R. Chellali, “Towards multimodal emotion recognition: a new approach”, *Proc. of the ACM International Conference on Image and Video Retrieval,* 2010.

[PLD05] H. Peng, F. Long, C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy”, *IEEE Trans. on PAMI*, vol. 27, no. 8, pp. 1226-1238, 2005.

[PP08] S. Petridis, M. Pantic, “Audiovisual discrimination between laughter and speech”, *IEEE ICASSP*, pp. 5117-5120, 2008.

[PR09] F. Pachet, P Roy, “Analytical features: a knowledge-based approach to audio feature generation”, *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[PRA02] J. Pinquier, J.-L. Rouas, R. André-Obrecht, “Robust speech/music classification in audio documents”, *ICSLP*, vol. 3, pp. 2005-2008, 2002.

[QT] “QT”, http://trolltech.com/products/qt.

[RF03] D. Rocchesso, F. Fontana, *The sounding object*, Freely distributed under the GNU Free Documentation License, http://www.soundobject.org/SObBook, 2003

[RG93] L. Rabiner, B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[RGA00] Y. Rui, A. Gupta, A. Acero, “Automatically Extracting Highlights for TV Baseball Programs”, *Proc. of ACM Multimedia*, pp. 105- 115, 2000.

[RJ93] L. Rabiner, B. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[RK04] R. Rifkin, A. Klautau, “In defense of one-vs-all classification”, *Journal of Machine learning Research*, vol. 5, pp.101-141, 2004.

[RT09] NIST. (2009), *The NIST Rich Transcription evaluation project website*. http://www.itl.nist.gov/iad/mig/tests/rt/

[SAH07] C. Segura, A. Abad, J. Hernando, C. Nadeu, “Multispeaker localization and tracking in intelligent environments”, *CLEAR 2007 and RT 2007*, LNCS, v. 4625, pp.82-90, 2008.

[Sau96] J. Saunders, “Real-time discrimination of broadcast speech/music”, *IEEE ICASSP*, vol. 2, pp. 993-996, 1996.

[SG00] P. Salembier, L. Garrido, “Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval”, *IEEE Trans. on Image Processing*, v. 9:4, pp. 561–576, 2000.

[SG99] C. Stauffer, W. Grimson, “Adaptive background mixture models for real-time tracking”, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 252–259, 1999.

[SK04] S H Srinivasan, M. Kankanhalli, “Harmonicity and dynamics-based features for audio”, *IEEE ICASSP*, vol. 4, pp. 321-324, 2004.

[SLP03] M. Stäger, P. Lukowicz, N. Perera, T. Büren, G. Tröster, T. Starner, “Sound button: design of a low power wearable audio classification system”, *Proc. IEEE Int. Symp. on Wearable Computers*, pp. 12–17, 2003.

[SP03] C. Sanderson, K. K. Paliwal, “Noise compensation in a person verification system using face and multiple speech features”, *Pattern Recognition*, vol. 36, no. 2, pp. 293–302, 2003.

[SPP99] S. Srinivasan, D. Petkovic, D. Ponceleon, “Toward robust features for classifying audio in the cue video system”, *Proc. 7th ACMInternational Conference on Multimedia*, pp. 393–400, 1999.

[SS02] B. Schölkopf, A. Smola, *Learning with kernels*, MIT Press, Cambridge, MA, 2002.

[SS97] E. Scheirer, M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator”, *IEEE ICASSP*, 1997.

[SSK06] L. Shalabi, Z. Shaaban, B. Kasasbeh, “Data mining: a preprocessing engine”, *Journal of Computer Science*, vol. 2 (9), pp. 735-739, 2006.

[TCS] http://www.tc-star.org/

[Tem07] A. Temko, *Acoustic event detection and classification*, PhD Thesis, UPC, Barcelona, 2007.

[TMN07] A. Temko, D. Macho, C. Nadeu, “Enhanced SVM training for robust speech activity detection”, *IEEE ICASSP*, 2007.

[TMN08] A. Temko, D. Macho, C. Nadeu, “Fuzzy integral based information fusion for classification of highly confusable non-speech sounds”, *Pattern Recognition*, vol. 41 (5), pp.1831-1840, Elsevier, 2008.

[TN09] A. Temko, C. Nadeu, “Acoustic event detection in meeting-room environments”, *Pattern Recognition Letters*, v. 30/14, pp 1281-1288, Elsevier, 2009.

[TNB08] A. Temko, C. Nadeu, J-I. Biel, “Acoustic Event Detection: SVM-based System and Evaluation Setup in CLEAR'07”, *Multimodal Technologies for Perception of Humans*, LNCS, vol. 4625, pp.354-363, Springer, 2008.

[UKR07] K. Umapathy, S. Krishnan, R. K. Rao, “Audio signal feature extraction and classification using local discriminant bases”, *IEEE Trans. On Audio Speech And Language Processing*, vol. 15, iss. 4, pp. 1236-1246, 2007.

[Van79] N. J. Vanderveer, *Ecological acoustics: human perception of environmental sounds*, PhD dissertation, Cornell University, 1979.

[Vap98] V. N. Vapnik, “Statistical learning theory”, John Wiley & Sons, New York, 1998.

[VGB08] D. Vasquez, R. Gruhn, R. Brueckner, W. Minker, “Comparing linear feature space transformations for correlated features”, *Perception in Multimodal Dialogue Systems*, LCNS, vol. 5078/2008, pp. 176-187, Springer, 2008.

[VIB03] M. Vacher, D. Istrate, L. Besacier, E. Castelli, J. Serignat, “Smart audio sensor for telemedicina”, *Proc. Smart Object Conference*, 2003.

[WB06] D. Wang, G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley-IEEE Press, 2006.

[WBW05] S. Wrigley, G. Brown, V. Wan, S. Renals, “Speech and crosstalk detection in multichannel audio”, *IEEE Trans. Speech Audio Processing*, v. 13, 84–91, 2005.

[WCC05] K. Wickramaratna, M. Chen, S. Chen, M. Shyu, “Neural network based framework for goal event detection in soccer videos”, *Proc. of the Seventh IEEE International Symposium on Multimedia*, pp. 21-28, 2005.

[WHL99] H. L. Wang, J. Huang, Z. Liu, Y. Wang, Y. Chen, E. K. Wong, “Integration of multimodal features for video scene classification based on HMM”, *IEEE Workshop on Multimedia Signal Processing*, pp. 53-58, 1999.

[WN10] M. Wolf, C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones", *Proc. Interspeech*, pp. 574-577, 2010.

[WS01] T. Wark, S. Sridharan, “Adaptive fusion of speech and lip information for robust speaker identification”, *Digital Signal Processing*, vol. 11, no. 3, pp. 169–186, 2001.

[WS09] A. Waibel, R. Stiefelhagen, *Computers in the Human Interaction Loop*, Springer, New York, USA, 2009.

[XDX03] M. Xu, L.-Y. Duan, C.-S. Xu, Q. Tian, “A fusion scheme of visual and auditory modalities for event detection in sports video”, *Proc. ICASSP*, vol. 3, pp.189-192, 2003.

[XMZ02] G. Xu, Y.-F. Ma, H.-J. Zhang, S. Yang, “Motion based event recognition using HMM”, *Proc. ICPR*, vol. 2, pp. 831-834, 2002.

[YEK02] S.J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (for HTK version 3.2)*, Cambridge University, 2002

[ZK99] T. Zhang, C.-C. Kuo, “Hierarchical classification of audio data for archiving and retrieving”, *Proc. IEEE ICASSP*, vol. 6, pp. 3001–3004, 1999.

[ZLC07] Y. Zhang, Q. Liu, J. Cheng, H. Lu, “Multimodal based highlight detection in broadcast soccer video”, *Asia-Pacific Workshop on Visual Information Processing*, 2007.

[ZSH10] M. Zelenák, C. Segura, J. Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features", *Proc. Interspeech*, pp. 2302-2305, 2010.

[ZWR03] M. Zobl, F. Wallhoff, G. Rigoll, “Action recognition in meeting scenarios using global motion features,” *Proc. Fourth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 32–36, 2003.

[ZZH08] X. Zhuang, X. Zhou, T.S. Huang, M. Hasegawa-Johnson, “Feature analysis and selection for acoustic event detection”, *Proc. ICASSP*, pp. 17-20, 2008.

[ZZL08] X. Zhou, X. Zhuang, M. Lui, H. Tang, M. Hasgeawa-Johnson, T. Huang, “HMM-based acoustic event detection with adaboost feature selection”, *Multimodal Technologies for Perception of Humans*, LNCS, vol. 4625, Springer, 2008.