Unified data-driven approach for audio indexing, retrieval and recognition

Houssemeddine Khemiri

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0055

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Traitement du Signal et des Images »

Houssemeddine KHEMIRI

Unified Data-Driven Approach for Audio Indexing Retrieval and Recognition

Directeur de thèse : **Gérard CHOLLET**
Co-encadrement de la thèse : **Dijana PETROVSKA-DELACRÉTAZ**

Jury

Mme Geneviève BAUDOIN, Professeur, Dpt Signaux et Télécommunications , ESIEE	Rapporteur
M. Hermann NEY, Professeur , Dpt Informatique 6, RWTH Aachen	Rapporteur
M. Xavier ANGUERA, Docteur, Telefonica Research, Barcelone	Examinateur
M. Laurent BESACIER, Professeur, LIG, Université J. Fourier	Examinateur
M. Geoffroy Peeters, HDR, Analyse/Synthèse, IRCAM	Examinateur
M. Gaël RICHARD, Professeur, TSI, TELECOM ParisTech	Examinateur

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my supervisor, Dr. Dijana Petrovska-Delacrétaz, for her support and wise supervision throughout my thesis. I am especially grateful for her constructive criticism and for her confidence in my work and my ideas.

I am deeply grateful to my supervisor, Prof. Gérard Chollet, for his review and his constructive comments and suggestions that have been of great value for me.

My sincere thanks are due to Prof. Geneviève Baudoin, Prof. Hermann Ney, Dr. Xavier Anguera, Prof. Laurent Besacier, Prof. Geoffroy Peeters and Prof. Gaël Richard for being members of my PhD committee and for their valuable comments and suggestions for improving this thesis.

Financial support from the French National Research Agency (ANR) SurfOnHertz project under the contract number ARPEGE 2009 SEGI-17, the vAssist project (AAL-2010-3-106) and the FUI Arhome project, is greatly acknowledged.

My thanks are also due to Fréderic Bimbot, Jan Černocký, Asmaa El Hannani, Guido Aversano and members of the French SYMPATEX project who provided me some of the programs used for this research.

My warmest gratitude is due to the head of the department of the TSI research group in Télécom ParisTech, Prof. Yves Grenier, and all its members. I owe particular thanks to Leila, Joseph, Asmaa, Daniel, Pierre, Pierrick, Manel, Mounir, Sébastien, Sathish, Patrick, Jacques, Stephan, Jirasri and Yafei for their encouragement, reviews and interesting discussions.

Special thanks go to Prof. Bernadette Dorizzi and all the members of the EPH

department in Télécom SudParis who have been very supportive when it was necessary, especially to Sanjay, Sarra, Monia, Maher, Mouna, Nadia, Toufik and Mohamed.

I would also like to thank Hugues Sansen, Founder and CEO of SHANKAA, for his relevant comments, suggestions, help and interesting discussions.

My deep and sincere gratitude go to all my friends, especially Haythem, Borhen, Sana, Hamdi, Rachid, Kamel, Takoua, Ahlem, Azza, Meriem, Hamed, Maher, Slim, Mehdi, Marwen, Rim, Alae, Dali, Zied, Walid and Bilel for their continuous support, particularly during the difficult moments.

Finally, my most sincere gratitude goes to my wife, my lovely daughter, my parents, sister and brother for their encouragement and the tremendous support they provided me throughout my entire life, especially during my PhD period. To them I dedicate this thesis.

Abstract

The amount of available audio data, such as broadcast news archives, radio recordings, music and songs collections, podcasts or various internet media is constantly increasing. In the same time there are not a lot of audio classification and retrieval tools, which could help users to browse audio documents.

Content-based audio-retrieval is a less mature field compared to image and video retrieval. There are some existing applications such as song classification, advertisement (commercial) detection, speaker diarization and identification, with various systems being developed to automatically analyze and summarize audio content for indexing and retrieval purposes. Within these systems audio data is treated differently depending on the applications. For example, song identification systems are generally based on audio fingerprinting using energy and spectrogram peaks (as in the SHAZAM and the Philips systems). While speaker diarization and identification systems are using cepstral features and machine learning techniques such as Gaussian Mixture Models (GMMs) and/or Hidden Markov Models (HMM).

The diversity of audio indexing techniques makes unsuitable the simultaneous treatment of audio streams where different types of audio content (music, commercials, jingles, speech, laughter, etc.) coexist.

In this thesis we report our recent efforts in extending the ALISP (Automatic Language Independent Speech Processing) approach developed for speech as a generic method for audio indexing, retrieval and recognition. ALISP is a data-driven technique that was first developed for very low bit-rate speech coding, and then successfully adapted for other tasks such as speaker verification and forgery, and language identification. The particularity

of ALISP tools is that no textual transcriptions are needed during the learning step, and only raw audio data is sufficient. Any input speech data is transformed into a sequence of arbitrary symbols. These symbols can be used for indexing purposes. The main contribution of this thesis is the exploitation of the ALISP approach as a generic method for audio (and not only speech) indexing and recognition. To this end, an audio indexing system based on the ALISP technique is proposed. It is composed of the following modules:

- Automated acquisition (with unsupervised machine learning methods) and Hidden Markov Modeling (HMM) of ALISP audio models.
- Segmentation (also referred as sequencing and transcription) module that transforms the audio data into a sequence of symbols (using the previously acquired ALISP Hidden Markov Models).
- Comparison and decision module, including approximate matching algorithms inspired form the Basic Local Alignment Search (BLAST) tool widely used in bioinformatics and the Levenshtein distance, to search for a sequence of ALISP symbols of unknown audio data in the reference database (related to different audio items).

Our main contributions in this Ph.D can be divided into three parts:

1. Improving the ALISP tools by introducing a simple method to find stable segments within the audio data. This technique, referred as spectral stability segmentation, is replacing the temporal decomposition used before for speech processing. The main advantage of this method is its computation requirements which are very low comparing to temporal decomposition.
2. Proposing an efficient technique to retrieve relevant information from ALISP sequences using BLAST algorithm and Levenshtein distance. This method speeds up the retrieval process without affecting the accuracy of the audio indexing process.
3. Proposing a generic audio indexing system, based on data-driven ALISP sequencing, for radio streams indexing. This system is applied for different fields of audio indexing to cover the majority of audio items that could be present in a radio stream:

- audio identification: detection of occurrences of a specific audio content (music, advertisements, jingles) in a radio stream;
- audio motif discovery: detection of repeating objects in audio streams (music, advertisements, and jingles);
- speaker diarization: segmentation of an input audio stream into homogenous regions according to speaker's identities in order to answer the question: "Who spoke when?";
- nonlinguistic vocalization detection: detection of nonlinguistic sounds such as laughter, sighs, cough, or hesitations;

The evaluations of the proposed systems are done on the YACAST database (a working database for the SurfOnHertz project) and other publicly available corpora. The experimental results show an excellent performance in audio identification (for advertisement and songs), audio motif discovery (for advertisement and songs), speaker diarization and laughter detection. Moreover, the ALISP-based system has obtained the best results in ETAPE 2011 (Evaluations en Traitement Automatique de la Parole) evaluation campaign for the speaker diarization task.

Glossary

Automatic speech recognition: Conversion of a speech signal into a textual representation by automated methods.

Audio fingerprint: Compact content-based signature that represents an audio recording.

Audio identification: Detection and location of occurrences of a specific audio content (music, advertisement, jingle,..) in audio streams or audio databases.

Audio indexing: Extraction of relevant information from unknown audio data.

Audio motif discovery: Detecting repeating audio objects in audio streams or audio databases.

Basic Local Alignment Search Tool (BLAST): Algorithm for comparing primary biological sequence information, such as amino-acid sequences of different proteins or the nucleotides of DNA sequences.

Data-driven approaches: Techniques that automatically learn the linguistic units and information required from representative examples of data without human expertise.

Hidden Markov Model (HMM): Statistical model used to model a process which evolves over time, where the exact state of the process is unknown, or "hidden".

High-level information: Set of information that reflects the behavioral traits such as prosody, phonetic information, pronunciation, idiolectal word usage, conversational patterns, topics of conversations, etc.

Levenshtein distance: String metrics for measuring the difference between two sequences. The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions, substitutions) required to change one word into another.

Mel-Frequency Cepstral Coefficients (MFCC): Coefficients of the cepstrum of the

short-term spectrum, downsampled and weighted according to the Mel scale that follows the sensitivity of the human ear.

Nonlinguistic vocalization: Very brief, discrete, nonverbal expressions related to human behavior.

Precision: Fraction of retrieved documents that are relevant to the search.

Recall: Fraction of the documents that are relevant to the query that are successfully retrieved.

Reference Database: Contains all the audio items to be identified by an audio identification system.

Speaker diarization: Segmenting an input audio data into homogenous regions according to speaker's identities in order to answer the question "Who spoke when?".

Speaker identification: Determining which registered speaker provides a given utterance.

Speaker verification: Accepting or rejecting the identity claim of a speaker.

Contents

List of Figures

List of Tables

Chapter 1

Résumé Long

1.1 Introduction

La quantité de données audio disponibles, telles que les enregistrements radio, la musique, les podcasts et les publicités est en augmentation constance. Par contre, il n'y a pas beaucoup d'outils de classification et d'indexation, qui permettent aux utilisateurs de naviguer et retrouver des documents audio. L'indexation audio par le contenu est un domaine moins mature que l'indexation d'images et de vidéos. Les applications existantes telles que la classification des morceaux de musique, l'identification des publicités et la segmentation et regroupement en locuteurs sont basées sur différents systèmes mis au point pour analyser et résumer automatiquement le contenu audio à des fins d'indexation et identification. Dans ces systèmes, les données audio sont traitées différemment en fonction des applications. Par exemple, les systèmes d'identification des morceaux de musique sont généralement basés sur ce qu'on appelle les empreintes audio en utilisant l'énergie ou les pics dans les spectrogrammes comme les systèmes proposés par SHAZAM et PHILIPS. Alors que les systèmes de segmentation et regroupement en locuteurs utilisent généralement les coefficients cepstraux et les techniques d'apprentissage comme les mélanges de Gaussiennes ou les modèles de Markov cachés. La diversité de ces techniques d'indexation rend inadéquat le traitement simultané de flux audio où différents types de contenu audio (musique, publicité, jingles, parole, rire, etc.) coexistent. Dans cette thèse, nous présentons nos travaux

sur l'extension de l'approche ALISP (Automatic Speech Processing Language Independent) [28] (Chollet et al., 1999), développé initialement pour la parole, comme une méthode générique pour l'indexation et l'identification audio. ALISP est une approche non supervisée qui a été initialement développée pour le codage de la parole à très bas débit [26] (Cernoky, 1998) [96] (Padellini et al., 2005), puis exploitée avec succès pour d'autres tâches telles que la vérification du locuteur [40] (ElHannani et al., 2009) [39] (ElHannani, 2007) [102], la conversion de la voix [99] (Perrot et al., 2005) et l'identification de la langue (Petrovska-Delacrétaz et al., 2000). La particularité des outils ALISP est qu'aucune transcription textuelle ou annotation manuelle est nécessaire lors de l'étape d'apprentissage. Le principe de cet outil est de transformer les données audio en une séquence de symboles. Ces symboles peuvent être utilisés à des fins d'indexation. La principale contribution de cette thèse est l'exploitation de l'approche ALISP comme une méthode générique pour l'indexation et identification audio. De ce fait, un système d'indexation audio basé sur l'approche ALISP est proposé. Il est composé des modules suivants:

- Acquisition et modélisation des unités ALISP d'une manière non supervisée
- Segmentation (aussi appelée transcription) ALISP, qui transforme les données audio en une séquence de symboles (en utilisant les modèles de Markov cachés ALISP précédemment acquis).
- Comparaison et décision qui utilisent les algorithmes de recherche approximative des séquences de symboles, inspirées de la technique BLAST (Basic Local Alignment Search) [3] (Altschul et al., 1990) et la distance de Levenshtein [76] (Levenshtein , 1966).

Les principales contributions de cette thèse peuvent être divisées en trois parties:

1. Améliorer les outils ALISP en introduisant une méthode simple pour segmenter les données d'apprentissage en segments stables. Cette technique, appelée segmentation par stabilité spectrale, remplace la décomposition temporelle utilisée auparavant dans les outils ALISP. Le principal avantage de cette méthode est l'accélération du processus d'apprentissage non supervisé des modèles HMM ALISP.

2. Proposer une technique efficace pour la recherche et comparaison des séquences ALISP utilisant l'algorithme BLAST et la distance de Levenshtein. Cette méthode accélère le processus de la recherche approximative des séquences de symboles sans affecter les performances du système d'indexation audio

3. Proposer un système générique pour l'indexation audio pour les flux radiophoniques basé sur la segmentation ALISP. Ce système est appliqué dans différents domaines d'indexation audio pour couvrir la majorité des documents audio qui pourraient être présents dans un flux radio:

   - identification audio: détection d'occurrences d'un contenu audio spécifique (musique, publicité) dans un flux radio;
   - découverte des motifs audio récurrents: détection des répétitions des documents audio dans un flux radio (musique, publicité);
   - segmentation et regroupement en locuteurs: segmentation d'un flux audio en régions homogènes en fonction de l'identité des locuteurs afin de répondre à la question : "Qui parle quand?";
   - détection de vocalisation non linguistiques: détection de sons non linguistiques tels que les rires, soupirs, toux ou hésitations;

Les évaluations du système proposé pour les différentes applications sont effectuées avec la base de données YACAST (une base de données acquis dans le cadre du projet SurfOnHertz) et avec d'autres corpus disponibles publiquement. Les résultats expérimentaux montrent d'excellentes performances pour l'identification audio (pour la publicité et la musique), pour la découverte de motifs récurrents (pour la publicité et la musique), pour la segmentation et regroupement en locuteurs et pour la détection de rire. En outre, le système proposé basé sur ALISP, a obtenu les meilleurs résultats dans la campagne d'évaluation ETAPE 2011 (évaluations en Traitement Automatique de la Parole) pour la tâche de segmentation et regroupement en locuteurs.

Ce résumé est structuré de la façon suivante : la section 2 présente un état de l'art des principales méthodes de l'indexation audio par extraction d'empreintes. La section 3

décrit les principales contributions de nos travaux de thèse. Les évaluations du système proposé pour les tâches d'identification audio, découverte des motifs audio, segmentation et regroupement en locuteur et la détection du rire sont décrites dans la section 4.

1.2 État de l'Art des Systèmes d'Indexation Audio par Extraction d'Empreinte

L'indexation audio par extraction d'empreinte est composée de deux modules : un module d'extraction d'empreinte et un module de comparaison. La première étape dans un système d'indexation audio par extraction d'empreinte (appelé aussi l'identification audio par extraction d'empreinte) est la création d'une base d'empreintes à partir d'une base de références. La base de références contient les documents audio (musique, publicités, jingles) que le système pourrait identifier. Dans la deuxième étape un extrait audio inconnu est identifié en comparant son empreinte avec celles de la base de références. L'identification audio par extraction d'empreinte a été très étudiée durant les dix dernières années. Ainsi, l'état de l'art est relativement fourni, avec des propositions d'approches très diverses pour aborder le problème. Le principal défi de ces systèmes est de calculer une empreinte audio robuste aux différents types de distorsions et de proposer une méthode rapide de comparaison qui peut satisfaire les contraintes temps-réel quelle que soit la taille de la base de références.

Plusieurs méthodes d'indexation audio par extraction d'empreinte ont été proposées [25] (Cano et al., 2005). Nous avons choisi de présenter ces systèmes selon l'approche utilisée pour l'extraction d'empreinte. A travers les articles publiés sur le sujet, trois grandes familles se dégagent en ce qui concerne la technique d'extraction d'empreinte.

La première famille opère directement sur la représentation spectrale du signal pour extraire les empreintes. Ce type d'empreinte est généralement facile à extraire et ne requiert pas des ressources de calcul importantes. La deuxième famille fait appel aux techniques utilisées dans le domaine de la vision par ordinateur, l'idée principale étant de traiter le spectrogramme de chaque document audio comme une image 2-D et de trans-

former l'identification audio en un problème de traitement d'images. La dernière famille inclut les approches basées sur la quantification vectorielle et l'apprentissage automatique, ces systèmes proposent un modèle d'empreinte qui imite les techniques utilisées dans le traitement de la parole.

1.2.1 Techniques Basées sur la Représentation Spectrale

Ces techniques sont les plus couramment utilisées vue la simplicité d'extraction d'empreinte. Plusieurs systèmes ont utilisé directement la représentation spectrale du signal pour construire l'empreinte.

Haitsma et al. [57] ont développé un système d'identification audio pour la reconnaissance des morceaux de musique. Ils ont utilisé une échelle Bark pour réduire le nombre de bandes fréquentielles par l'intermédiaire de 33 bandes logarithmiques couvrant l'intervalle de 300Hz à 2 kHz. Le signe de la différence d'énergie des bandes adjacentes est calculé et stocké sous forme binaire. Le résultat de ce processus de quantification est une empreinte de 32 bits par trame. La méthode de recherche adoptée par PHILIPS consiste à indexer chaque trame de référence dans une table de correspondances (lookup table). Si le nombre de sous-bandes utilisées est N_b, alors chaque trame sera représentée par un vecteur de $(N_b - 1)$ bits et on retrouvera dans le "lookup table" 2^{N_b} entrées. Chaque trame binaire de l'empreinte sert de clé dans le lookup table, toutes les empreintes de références possédant une même trame binaire qu'une empreinte à identifier sont considérées comme candidates à l'identification. Haitsma suppose donc qu'il existe au moins une trame binaire de l'empreinte à identifier non distordue par rapport à la référence qui lui correspond. Cette technique a donné lieu à des études diverses. Une amélioration de la méthode d'extraction d'empreinte de façon à rendre plus robuste le système face aux distorsions comme l'étirement temporel (pitching) a été proposée [58] (Haitsma and Kalker, 2003). Dans [78] (Liu et al., 2009) ont modifié l'algorithme pour contourner l'hypothèse de présence d'une trame binaire non distordue.

Un autre système commercial (SHAZAM) qui se base sur la représentation spectrale du système a été proposé par Wang [133] pour l'identification d'un extrait audio inconnu capturé par un téléphone mobile. Cette technique binarise le spectrogramme en ne gardant

que des maxima locaux. Il s'agit alors d'extraire des pics de ce spectrogramme en prenant soin de choisir des points d'énergie maximale localement et en s'assurant une densité de pics homogène au sein du spectrogramme. L'auteur propose alors d'indexer les empreintes des références en utilisant la localisation des pics comme index. Cependant, un index s'appuyant sur la localisation de chaque point isolément se révèle peu sélectif. Par conséquent, Wang propose d'utiliser des paires de pics en tant que index, chaque pic est combiné avec ses plus proches voisins. Cette technique est utilisée pour identifier un morceau de musique dans un milieu bruité. Cependant pour les objets de courte durée (une publicité ou un jingle), elle s'avère inefficace vue le nombre insuffisant de pics extraits. De plus Fenet et al. [44] ont montré que ce système n'est pas robuste à l'étirement temporel et ont proposé une version différente de cet algorithme en se basant sur la transformée à Q constant (Constant Q Transform-CQT).

1.2.2 Techniques Basées sur la Vision par Ordinateur

Il y a eu plusieurs expériences de l'utilisation des techniques de vision par ordinateur pour l'identification audio par extraction d'empreinte. L'idée principale est de traiter le spectrogramme de chaque document audio comme une image 2-D.

Baluja et al. [12] ont exploité l'applicabilité des ondelettes dans la recherche des images dans des larges bases de données pour développer un système d'identification audio par extraction d'empreinte. Cette technique consiste à générer un spectrogramme à partir d'un signal audio avec les mêmes procédures que [57] (Haitsma and Kalker, 2002), ce qui donne 32 bandes d'énergie logarithmique entre 318 Hz et 2 kHz pour chaque trame. Ensuite, une image spectrale est extraite à partir de la combinaison des bandes énergétiques sur un certain nombre de trames et la décomposition en ondelettes, utilisant les ondelettes de Haar, est appliquée sur les images obtenues. Les signes du premiers 200 amplitudes des ondelettes sont exploité pour construire une empreinte binaire. Enfin, une table de hachage est utilisée pour trouver les meilleures empreintes et la distance de Hamming est calculée entre les empreintes candidates de morceaux de musique et les empreintes de la requête.

Ke et al. [68] ont proposé un système d'identification de morceaux de musique basé

sur l'algorithme de Viola-Jones [132] (Viola and Jones, 2001). Un algorithme de 'boosting' est utilisé sur un ensemble de descripteurs de Viola-Jones pour apprendre des descripteurs locaux et discriminants. Durant la phase de recherche, une liste des candidats est déterminée à partir des descripteurs appris auparavant. Pour chaque candidat, l'algorithme RANSAC [45] (Fishler and Bolles, 1987) est appliqué pour aligner le candidat avec la requête et une mesure de vraisemblance est calculée entre les deux morceaux.

1.2.3 Techniques Basées sur la Modélisation Statistique

Cette dernière famille regroupe les techniques utilisées habituellement pour le traitement de la parole, comme la quantification vectorielle ou les modèles de Markov cachés.

Cremer et al. [31] ont proposé une approche essentiellement basé sur la quantification vectorielle. La création de l'empreinte se fait à partir des descripteurs utilisés dans la norme MPEG-7. Les descripteurs utilisés sont l'intensité, la mesure de platitude spectrale et le facteur de crête spectral. La méthodologie de l'identification consiste à extraire ces descripteurs à partir des références. Un algorithme de quantification vectorielle produit ensuite un ensemble de centroides (appelés vecteurs de codage) approximant les vecteurs des descripteurs de la référence. Lorsque le système identifie un extrait inconnu, il extrait les vecteurs descripteurs du signal, puis pour chaque référence, projette ces vecteurs sur les vecteurs de codage de la référence. La référence possédant les vecteurs de codage qui produisent l'erreur de projection minimale est considérée comme la référence à identifier.

Cano et al. [24] ont proposé un système basé sur la modélisation de Markov caché. 32 modèles HMM appelés gènes audio sont utilisées pour segmenter le signal audio en utilisant l'algorithme de Viterbi. L'empreinte audio se compose de séquences d'étiquettes (les gènes) et d'information temporelle (temps du début et de la fin de chaque gène). Durant le processus d'appariement, des séquences des gènes sont extraites à partir d'un flux radio continu et comparées avec les empreintes des références. Afin de réduire la durée du traitement, l'algorithme de recherche de l'ADN appelé FASTA [98] (Pearson and Lipman, 1988) a été utilisé. Ce système a été évalué sur la tâche de l'identification des morceaux de musique dans un flux radio.

1.2.4 Etude Comparative

Comme l'on a mentionné auparavant, les systèmes d'indexation audio par extraction d'empreintes ont pour but de calculer une empreinte audio robuste contre différents types de distorsions et de proposer une méthode de comparaison efficace et rapide qui peut satisfaire les contraintes temps-réel. Nous avons comparé les systèmes présentés dans les sections précédentes en termes des critères suivants :

- Fiabilité : le nombre d'identifications correctes, les fausses alarmes et les fausses identifications.
- Robustesse: La capacité du système à identifier correctement les documents audio en présence de différents types de distorsion (bruit, filtrage, pitching, etc.).
- Granularité: La durée minimale de l'empreinte requête nécessaire pour identifier le document audio. Par exemple, la durée moyenne des publicités varie de 5 à 30 secondes, de ce fait il est nécessaire d'avoir une granularité inférieure à 5 secondes.
- Complexité : La complexité du système détermine le coût et le temps de calcul nécessaire pour l'identification.
- Passage à l'échelle: les performances du système en présence de plus grande base de références. Ce critère est en relation directe avec la complexité et la fiabilité du système.

Le tableau 1.1 illustre les performances des systèmes d'indexation audio par extraction d'empreintes selon les critères décrit en dessus.

D'autre part, différents protocoles expérimentaux sont utilisés pour évaluer les systèmes d'indexation audio par extraction d'empreintes. Ces protocoles sont résumés dans le tableau 1.2. Les deux mesures de performance utilisées pour évaluer ces systèmes sont :

- Précision: le nombre de documents audio correctement détectées / nombre total de documents audio.

Systèmes	Fiabilité	Robustesse	Granularité	Complexité	Passage à l'échelle
Haitsma et al. [57]	+	NA	NA	+	-
Wang et al. [133]	-	-	-	+	+
Pinquier et al. [104]	+	NA	+	+	-
Baluja et al. [12]	+	+	NA	-	NA
KE et al. [68]	-	-	NA	-	NA
Cremer et al. [31]	+	NA	NA	-	-
Cano et al. [24]	+	+	NA	-	NA

Table 1.1: Performances des système d'indexation audio par extraction d'empreintes en termes de fiabilité, robustesse, granularité, complexité et passage à l'échelle.

Systèmes	Références	Test	Précision	Rappel
Haitsma et al. [57]	4 chansons	4 chansons	100%	100%
Wang et al. [133]	10,000 chansons	250 chansons	NA	80%
Pinquier et al. [104]	32 jingles	10h radiodiffusion	100%	98,5%
Baluja et al. [12]	10,000 chansons	1,000 chansons	NA	97,9%
KE et al. [68]	1,862 chansons	220 chansons	93%	80%
Cremer et al. [31]	15,000 chansons	15,000 chansons	NA	98%
Cano et al. [24]	50,000 chansons	12h radiodiffusion	100%	100%

Table 1.2: Comparaison des performances des systèmes décrits dans la section 3.5.3, les bases de référence et l'ensemble de test, précision et rappel.

- Rappel: le nombre de documents audio correctement détectées / Le nombre de documents audio qui doivent être détectées.

La plupart des systèmes d'indexation audio décrits dans les tableaux 1.1 et 1.2 sont évalués sur un type spécifique de contenu audio (musique ou jingle). De plus ces systèmes utilisent des protocoles d'évaluation privés rendant la comparaison entre eux impossible.

Dans cette section, un aperçu des méthodes d'indexation audio par extraction d'empreintes est présenté. Ces systèmes devraient répondre à certains critères comme la granularité et la précision. En outre, l'empreinte doit être robuste à différentes dégradations que le signal audio pourrait subir. Nous avons aussi montré que ces systèmes utilisent différentes techniques pour extraire l'empreinte et proposent plusieurs méthodes de recherche des empreintes dans la base de références.

Dans cette thèse nous proposons un système d'indexation audio générique capable d'identifier simultanément les morceaux de musique, publicités, tours de parole et les

rires. Ce système sera évalué sur des bases privées et publiques, lors de la campagne d'évaluation QUAERO 2010 [108] (Ramona et al., 2012) et la campagne d'évaluation ETAPE 2011 [55] (Gravier el al., 2012). Dans la section suivante, les pricipales contibutions de nos travaux sont présentées. Elles concernent le module d'acquisition et modélisation des unités ALISP, le module de la recherche et comparaison des séquences ALISP et le développement du système générique d'indexation audio.

1.3 Contributions à l'Indexation Audio Non Supervisée

Les principales contributions de cette thèse peuvent être divisées en trois parties:

1. Améliorer les outils ALISP en introduisant une méthode simple pour segmenter les données d'apprentissage en segments stables. Cette technique, appelée segmentation par stabilité spectrale, remplace la décomposition temporelle utilisée auparavant dans les outils ALISP. Le principal avantage de cette méthode est l'accélération du processus d'apprentissage non supervisé des modèles HMM ALISP.

2. Proposer une technique efficace pour la comparaison et la recherche des séquences ALISP utilisant l'algorithme BLAST et la distance de Levenshtein. Cette méthode accélère le processus de la recherche approximative des séquences de symboles sans affecter les performances du système d'indexation audio.

3. Proposer un système générique pour l'indexation audio pour les flux radiophonique basé sur la segmentation ALISP. Ce système est appliqué dans différents domaines d'indexation audio pour couvrir la majorité des documents audio qui pourraient être présents dans un flux radio.

    - identification audio: détection d'occurrences d'un contenu audio spécifique (musique, publicité) dans un flux radio;
    - découverte des motifs audio: détection des répétitions des documents audio dans un flux radio (musique, publicité);

- segmentation et regroupement en locuteurs: segmentation d'un flux audio en régions homogènes en fonction de l'identité des locuteurs afin de répondre à la question : "Qui parle quand";
- détection de vocalisation non linguistiques: détection de sons non linguistiques tels que les rires, soupirs, toux ou hésitation;

Comme l'on a souligné précédemment, les outils ALISP ont été déjà utilisés pour le codage de la parole à très bas débit, la reconnaissance du locuteur et de la langue et la conversion de voix.

L'objectif de cette thèse est d'exploiter les informations de haut niveau fournies par les unités ALISP afin de développer un système d'indexation audio générique et unsupervisée.

Notre méthode consiste à segmenter les données audio en utilisant les modèles HMM ALISP. La particularité des outils ALISP est qu'aucunes transcriptions textuelles ne sont nécessaires lors de l'étape d'apprentissage, et seules les données audio brutes sont suffisantes. De cette manière, toutes les données audio sont transformées en une séquence de symboles, appelés symboles ALISP. Ces symboles peuvent être utilisés à des fins d'indexation.

1.3.1 Amélioration des Outils ALISP

Une partie de nos travaux est liée à adapter et améliorer les outils ALISP à l'égard de la tâche et les bases de données. Les améliorations que nous avons apportées concernent la segmentation initiale faite par la décomposition temporelle. La décomposition temporelle est utilisée pour obtenir une segmentation initiale et quasi-stationnaire des données audio. Ces segments sont regroupés en utilisant la quantification vectorielle. Ensuite, ces segments ainsi que leurs étiquettes sont utilisés comme transcription initiale pour la modélisation de Markov caché.

Dans cette section, d'autres méthodes de segmentation sont explorées afin d'accélérer le processus d'apprentissage des modèles ALISP et d'étudier l'influence de la segmentation initiale sur le système d'indexation audio. Ces méthodes sont les suivantes:

- Segmentation uniforme : c'est l'approche la plus simple pour segmenter les données audio. Elle consiste à segmenter les données audio en trame de taille égale.

- Segmentation par stabilité spectrale : Le but de cette méthode est de trouver les régions stables du signal audio. Ces régions représentent les segments spectralement stables des données audio. Ce processus est effectué en utilisant la courbe de stabilité spectrale obtenue en calculant la distance euclidienne entre deux vecteurs MFCC successives comme suit:

$$d = \sqrt{\sum_{i=1}^{n} (C_{i,j} - C_{i,j+1})^2} \tag{1.1}$$

  Où $C_{i,j}$ et $C_{i,j+1}$ sont deux vecteurs MFCC successifs et n est leur taille. Les maxima locaux de cette courbe représentent les frontières des segments alors que les minima représentent les trames "stables" du signal audio.

- Segmentation phonétique : Cette méthode consiste à utiliser des modèles HMM phonétiques pour obtenir la segmentation initiale des données audio. Cette segmentation est utilisée pour déterminer si les modèles phonétiques pourraient être utilisés à des fins d'indexation audio. Les modèles HMM phonétiques sont appris avec la base de données ESTER (base de données française de radiodiffusion) [49] (Galliano et al., 2009). Comme pour les modèles ALISP, chaque phone (41 phones) est modélisé par un HMM gauche-droite ayant trois états émetteurs sans sauts. La segmentation phonétique remplace la décomposition temporelle et la quantification vectorielle. En fait, la segmentation phonétique est utilisée en tant que transcription initiale pour la modélisation de Markov caché.

Un ensemble de modèles ALISP est appris pour chaque technique de segmentation initiale en utilisant une base de données d'apprentissage de 288 heures issues 12 radios françaises. La figure 1.1 illustre le spectrogramme d'un extrait audio et les segmentations obtenues avec chaque ensemble de modèles ALISP.

Cette figure montre que la segmentation par stabilité spectrale fournit la segmentation la plus proche à celle fournie par la décomposition temporelle. D'autre part, les segmentations phonétiques et uniformes ne sont pas appropriées pour obtenir une segmentation en région spectralement stables des données audio.

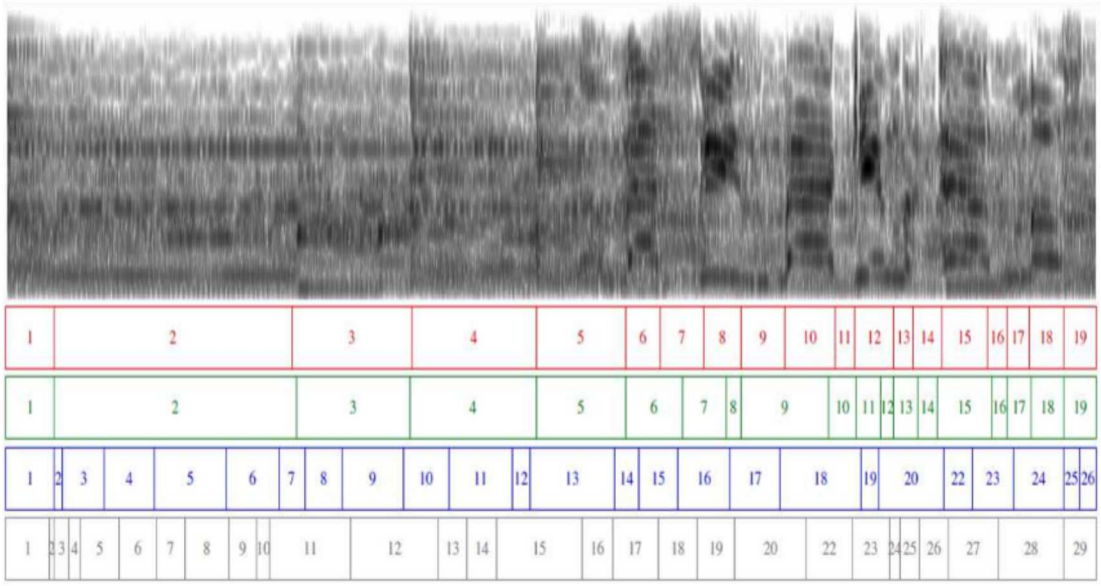


Figure 1.1: Spectrogramme d'un extrait audio et les segmentations obtenues avec chaque ensemble de modèles ALISP utilisant la décomposition temporelle (rouge), Segmentation par stabilité spectrale (vert), segmentation uniforme (bleu), segmentation phonétique (gris).

En plus, pour acquérir 32 modèles ALISP avec 288 heures de données audio, le temps de traitement de compose comme suit:

- 10 jours pour la décomposition temporelle;
- 7 jours pour la segmentation par stabilité spectrale;
- 6 jours pour la segmentation uniforme;
- 18 jours pour la segmentation phonétique.

Ce résultat montre qu'en remplaçant la décomposition temporelle par la segmentation par stabilité spectrale, le temps de traitement est diminué de 3 jours. D'autre part, l'influence des quatre méthodes de segmentation sur les performances du système d'indexation audio proposé sera étudiée dans les sections suivantes.

1.3.2 Appariement Approximatif des Séquences ALISP

Le système d'indexation audio proposé est composé de trois modules: acquisition et modélisation des unités ALISP, le module de segmentation ALISP et le modules d'appariement approximatif des séquences ALISP. Dans la section précédente, nous avons présenté nos contributions pour le premier et le deuxième module. Dans cette section, une nouvelle technique de recherche approximative de séquence de symboles ALISP est proposée. Cette technique est basée sur l'algorithme BLAST et la distance de Levenshtein.

Comme la principale exigence du système d'indexation audio est la robustesse aux plusieurs types de distorsions, les séquences de symboles ALISP extraites du signal audio n'est pas entièrement identique aux séquences qui existent dans la base de références. De ce fait, deux techniques d'appariement approximatif des séquences ALISP sont développées. La première est basée sur une recherche exhaustive (ou recherche brute), tandis que la seconde technique est inspirée de la méthode BLAST, utilisée généralement en bioinformatique.

1.3.2.1 Recherche Exhaustive

Dans cette méthode les séquences ALISP extraites du flux radio continu sont comparées contre les transcriptions ALISP stockées dans la base de référence. Tout d'abord, les transcriptions ALISP de chaque document audio de référence (ceux que nous allons chercher dans le flux radio continu) sont calculées. Ensuite, le flux radio de test est transformé en une séquence de symboles ALISP. Une fois les transcriptions ALISP de référence et de données de test sont obtenues, nous pouvons passer à l'étape d'appariement. La mesure de similarité utilisée pour comparer les transcriptions ALISP est la distance de Levenshtein. La distance de Levenshtein mesure la similarité entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.

Pour commencer, la méthode de recherche utilisée dans notre système est très élémentaire. A chaque itération on avance par une unité ALISP dans le flux radio de test et la distance de Levenshtein est calculée entre la transcription de référence et la transcription de l'extrait sélectionné dans le flux radio. Au moment où la distance de Levenshtein est inférieure à un

certain seuil, cela signifie que nous avons un chevauchement avec la référence. Puis nous continuons la comparaison en avançant par un symbole ALISP jusqu'à ce que la distance de Levenshtein augmente par rapport à sa valeur à l'itération précédente. Ce point indique l'appariement optimal, où toute la référence a été détectée.

Afin d'accélérer la phase de recherche, une méthode alternative d'appariement approximatif des séquences ALISP, basée sur BLAST et la distance de Levenshtein, est développée.

1.3.2.2 BLAST Algorithm

BLAST est un algorithme de comparaison de séquence biologique, tels que les séquences de nucléotides ou d'acides aminés. Une recherche BLAST permet de chercher une séquence requête dans une base de données, et identifier les séquences de chaînes de caractères ayant une mesure de similarité inférieur à un certain seuil.

Soit q la séquence de chaîne requête, D la base de données et w une sous-chaîne de la séquence q. La première étape de l'algorithme consiste à construire un "Lookup Table (LUT)" qui contient toutes les sous-chaînes dans D de longueur w. Chaque entrée de LUT pointe à la position de la sous-chaîne dans la base D. Dans la deuxième étape, pour chaque sous-chaîne de la séquence requête q, une liste de sous-chaînes est générée en utilisant le LUT. Cette liste contient toutes les sous-chaînes de longueur w avec un score de similarité supérieur à un certain seuil T. La dernière étape de l'algorithme consiste à étendre chaque sous-chaîne candidate pour trouver l'alignement optimal avec la séquence requête q. Un candidat est considéré comme l'alignement optimal si son score de similarité avec la requête q est supérieur à un certain seuil S. Dans notre cas, la requête est une longue séquence de symboles ALISP où des occurrences de publicités et des morceaux de musique sont recherchées. Afin de résoudre ce problème, l'algorithme BLAST a été adapté comme suit.

1.3.2.3 Méthode Proposée pour l'Appariement Approximatif

Le processus d'appariement approximatif illustré dans la figure 1.2 est proposé. Tout d'abord, un LUT est créé par toutes les séquences ALISP de longueur w mais avec un décalage de k unités qui existent dans les transcriptions ALISP de la base de références.

Cette base contient tous les documents audio que le système pourrait identifier, tels que des morceaux de musique, des publicités, des tours de parole et des motifs audio.

Chaque entrée de LUT pointe vers sa position dans le document de référence. Comme une séquence ALISP peut se produire dans plusieurs références, une séquence ALISP peut avoir plusieurs pointeurs et positions.

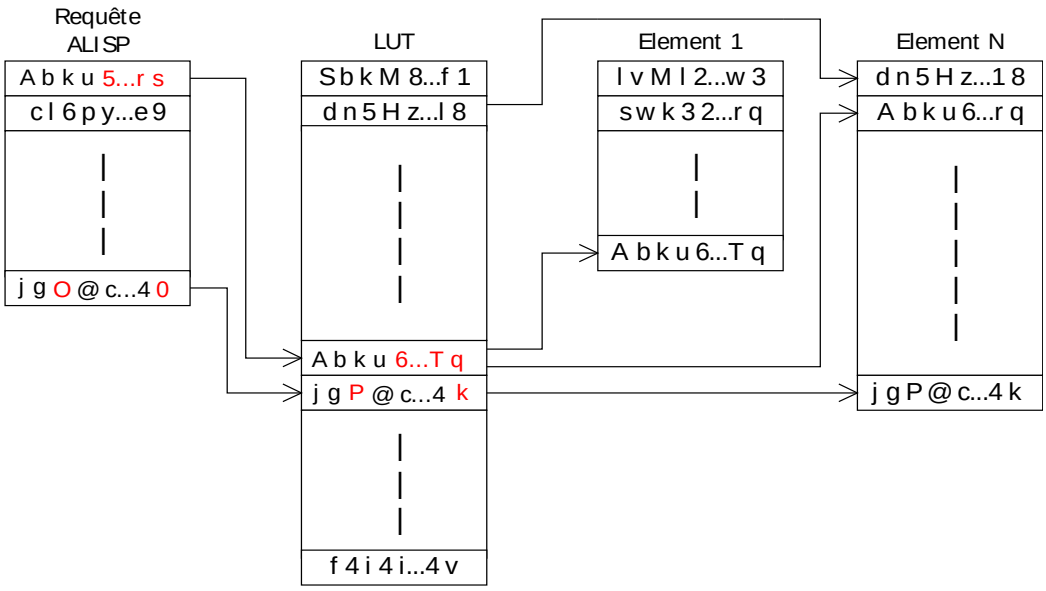


Figure 1.2: Appariement approximatif d'une requête ALISP en utilisant un Lookup Table (LUT) et une base de référence contenant N éléments.

Ensuite, la transcription ALISP de la requête est calculée, et pour chaque sous-séquence w avec un décalage de k de cette requête une liste de sous-séquences candidates est générée à l'aide du LUT. A partir de cette liste de sous-séquences, une liste de références et la position dans laquelle les sous-séquences se produisent est créée.

Comme la base de référence est formée par la transcription ALISP de chaque document audio, l'étape finale du processus de comparaison est différente de celle de BLAST. Elle consiste à une simple comparaison entre la transcription ALISP de la requête audio et les références candidates avec la distance de Levenshtein. La référence candidate

ayant la distance de Levenshtein la plus faible et inférieure à un certain seuil est relative à l'appariement optimale de la requête audio.

1.3.3 Système Générique d'Indexation Audio à Base d'ALISP

L'objectif principal de nos travaux est d'indexer et identifier la majorité des éléments audio présents dans un flux radio. Ces éléments sont généralement: la musique, publicité, jingle, la parole et la vocalisation non linguistique (rire, toux, ...). À cette fin, un système d'indexation audio générique et unsupervisé basé sur la méthode ALISP est développé et appliqué pour l'identification audio, la découverte de motif audio, la segmentation et regroupement en locuteurs et la détection de rire. Bien que ces systèmes soient différents, ils utilisent une architecture commune basée sur la méthode ALISP. Comme le montre la figure 1.3, cette architecture est composée de trois modules: modélisation et acquisition des modèles ALISP, segmentation ALISP et appariement approximatif des séquences ALISP.

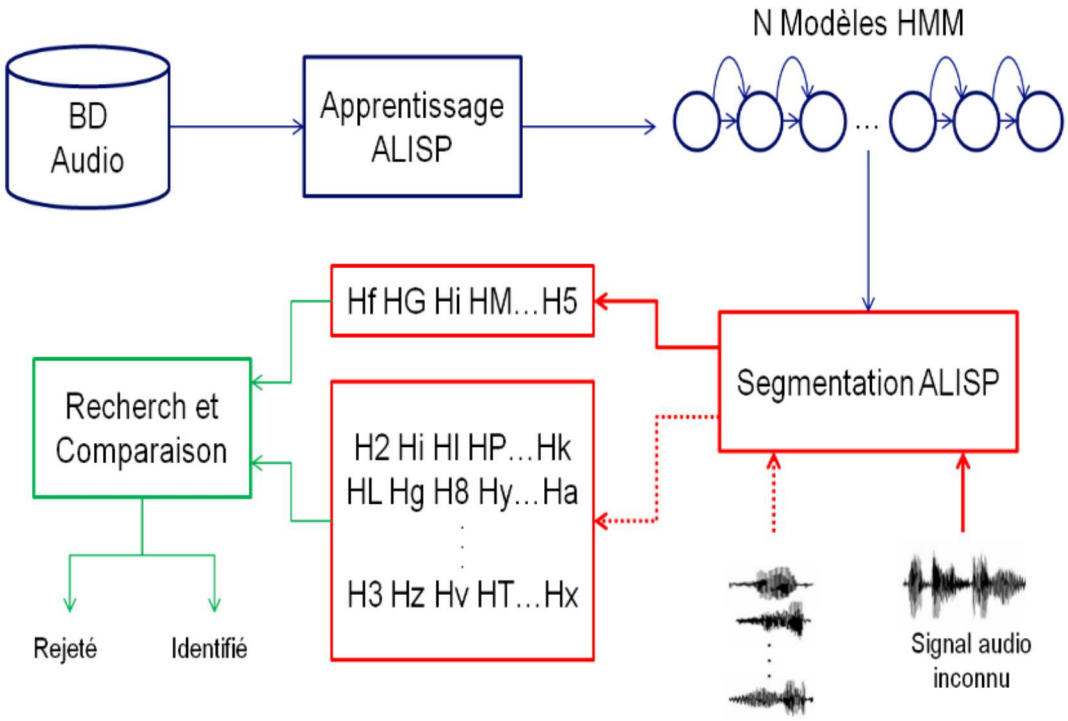


Figure 1.3: Architecture générale du système générique d'indexation audio à base d'ALISP.

Dans cette section, les principales contributions de cette thèse ont été présentées. D'abord, nous avons montré qu'en remplaçant la décomposition temporelle par la segmentation par stabilité spectrale le processus d'apprentissage des modèles ALISP pourrait être accéléré. Ensuite une méthode d'appariement approximatif des séquences ALISP inspirée de BLAST et la distance de Levenshtein est présentée. Enfin, un système d'indexation audio générique et unsupervisé basé sur ALISP est proposé. Dans la section suivante, le système d'indexation proposé est évalué sur les tâches d'identification audio, découverte de motifs audio, segmentation et regroupement en locuteurs et détection de rire.

1.4 Evaluations et Résultats

Dans cette section, nous présentons les protocoles expérimentaux et les résultats obtenus pour les différentes tâches auxquelles le système d'indexation audio est appliqué.

1.4.1 Identification Audio

Le système d'identification audio basée sur ALISP est utilisé pour identifier les publicités et les morceaux de musique dans les flux de radio. Pour évaluer ce système, deux protocoles expérimentaux sont proposés.

Le premier protocole, appelé protocole YACAST, correspond à 12 journées radios fournies dans le cadre du projet ANR-SurfOnHertz et divisées comme suit:

- Données développement : 5 jours radios sont utilisés pour étudier la stabilité des transcriptions ALISP et fixer le seuil de décision pour la distance de Levenshtein.
- Données de référence: elles contiennent 2,172 publicités et 7,000 morceaux de musique menant à 9,172 éléments de référence.
- Données d'évaluation: 7 jours de trois radios françaises. Ces jours sont différents de ceux utilisés dans les données de développement et dans le corpus d'apprentissage de modèles ALISP. Ces données contiennent 1,456 publicités et 4,880 chansons à identifier.

Système	R%	P%	Eléments non identifiés	Fausses alarmes
Décomposition temporelle	92	100	416	0
Stabilité spectrale	92	100	416	0
Segmentation uniforme	90	95	623	301
Segmentation phonétique	85	87	942	806

Table 1.3: Rappel (P%), Précision (R%), nombre d'éléments non identifiés et nombre de fausses alarmes pour les différentes techniques de segmentation avec le protocole YACAST.

Le deuxième protocole, appelé protocole QUAERO, a été utilisé lors de la campagne d'évaluation QUAERO 2010 (http ://www.quaero.org/)). Il est décrit comme suit :

- Données de développement: les mêmes que celles utilisées dans le protocole YACAST
- Données de référence: elles contiennent 7,309 extraits de morceaux de musique ayant une durée d'une minute chacune. La position de ces signatures dans les morceaux de musique est inconnue.
- Données d'évaluation: 7 jours de la radio française RTL (durée totale de 168 heures). Ces enregistrements contiennent 551 morceaux de musique..

Afin d'évaluer les performances de notre système d'identification, les mesures de rappel (R%) et précision (P%) sont utilisées. Pour le protocole YACAST l'influence des méthodes de segmentation initiale sur les performances du système proposé est étudiée. Le tableau 1.3 illustre les taux de rappel et précision pour chaque méthode de segmentation initiale.

Le tableau 1.3 montre que pour les modèles ALISP HMM utilisant la décomposition temporelle et la stabilité spectrale, le système n'était pas en mesure d'identifier 416 éléments. Ces éléments correspondent à 389 chansons et 27 publicités.

Pour la musique, 372 morceaux sont liés à des chansons qui ont une version différente de celle présente dans la base de référence. Par exemple, nous avons trouvé 302 morceaux de musique "live" dans le flux radio, tandis que les références associées sont interprétées en version studio. Pour les publicités, les 27 éléments non identifiés sont différents de leurs références. Ces résultats montrent que le système proposé permet de trouver les erreurs des annotations manuelles de la musique et des publicités.

Système	R%	P%	Eléments non identifiés	Fausses alarmes
Notre système	100	100	0	0
Fenet et al. [44]	97.4	100	12	0
Ramona et al. [109]	96.9	99	15	2
Yacast [108]	95.9	99	17	0

Table 1.4: Précision (P%), Rappel (R%), nombre d'éléments non identifiés et nombre de fausses alarmes pour le protocole QUAERO 2010.

D'autre part, le système utilisant la segmentation uniforme a obtenu plus d'éléments non identifiés et de fausses alarmes que ceux obtenus avec la décomposition temporelle et la stabilité spectrale. De plus, les modèles ALISP basés sur la segmentation phonétique ont obtenu les pires résultats. Cependant, ce système a correctement identifié tout les éléments audio où la parole est la partie dominante.

Le tableau 1.4 compare les performances de notre système par rapport à ceux participant à la campagne dévaluation QUAERO 2010. Notons que dans le protocole QUAERO la reconnaissance d'interprétations différentes du même titre est considérée comme hors du périmètre de l'identification audio.

Le tableau 1.4 montre que notre système se comporte aussi bien que les systèmes qui ont participé à la campagne d'évaluation. De plus notre système a montré sa robustesse à l'étirement temporel (plus connu sous le nom du "pitching"). En effet parmi les 459 morceaux de musique correctement identifiés, 209 morceaux ont été accélérés (ou ralentis) jusqu'à 7% par rapport à leurs versions de références.

Relativement au temps nécéssaire pour les différents modules, l'acquisition et la modélisation des unités ALISP se fait hors ligne. D'autre part, le temps nécessaire pour La transcription des flux audio avec les modèles ALISP est négligeable. Par conséquence, la complexité de calcul du système est actuellement limitée à la recherche de la plus proche séquence ALISP avec la distance de Levenshtein. Avec la méthode de recherche exhaustive, le temps nécessaire pour traiter une seconde du signal de test est de six secondes alors qu'en utilisant la nouvelle méthode de recherche basée sur BLAST le temps de traitement est réduit à 0.49 secondes avec 33 modèles ALISP et pour une base de références qui contient 9,000 éléments avec une machine 3.00GHz Intel Core 2 Duo 4 Go de RAM.

1.4.2 Découverte des Motifs Audio Récurrents

Pour l'identification audio, le système dispose d'une base de références qui contient les documents audio (musique, publicités) qu'il pourrait identifier. De ce fait un morceau de musique ou une publicité qui n'a pas une référence ne pourrait pas être identifié. C'est le cas des nouvelles chansons et publicités qui sont diffusées pour la première fois par les radios. Ce genre de document audio est généralement joué plusieurs fois par la radio. Par conséquent, la détection des répétitions d'éléments audio (appelé aussi découverte des motifs audio) dans les flux radio devrait conduire à la découverte automatique des publicités et des chansons sans avoir besoin d'une base de références.

La découverte des motifs audio est généralement basée sur l'extraction d'empreintes audio. Dans cette thèse, les outils ALISP sont utilisés pour convertir le flux audio hétérogène (contenant de la musique, jingles, publicités, parole, etc.) en une séquence de symboles. Ces symboles représentent l'empreinte nécessaire pour détecter les éléments répétitifs dans les flux audio. Par conséquent, le problème consistant à découvrir les motifs audio est transformé en un problème de recherche approximative des séquences ALIPS qui se répètent. Ce problème est traité à l'aide du système générique d'indexation audio où les outils ALISP sont utilisés pour calculer l'empreinte audio et la méthode de recherche inspiré de BLAST et la distance de Levenshtein est utilisée pour accélérer la recherche motifs audio dans le flux radio.

Afin d'évaluer notre système pour cette tâche, le protocole YACAST (utilisé aussi pour la tâche d'identification audio est utilisé). Dans ce protocole les données d'évaluation se constituent de 7 jours de trois radios françaises qui contiennent 1,456 publicités et 4,880 chansons. Dans ces données, il existe 1,315 répétitions pour les publicités et 3,081 pour la musique. La moyenne des répétitions est de 2 pour les publicités et 4 pour la musique.

L'évaluation du système proposé pour la tâche de la découverte de motifs audio a été réalisée avec les mesures de précision et rappel, exposées dans le tableau 1.5.

Pour la musique, le système n'était pas capable de détecter 21 répétitions. Ces répétitions sont liées à des morceaux de musique qui se chevauchent avec des tours de parole, ce qui perturbe le processus de détection. D'autre part, l'absence de fausses alarmes

	Répétitions	R%	P%	répétitions non détectées	Fausses alarmes
Songs	3081	99	100	21	0
Ads	1315	98	99	14	6

Table 1.5: Nombre de répétitions, précision (P%), rappel (R%), nombre des répétitions non détectées et nombre des fausses alarmes, obtenu pour le protocole d'évaluation YACAST.

confirme le résultat obtenu pour la tâche d'identification audio.

Pour les publicités, le système n'était pas capable de détecter 14 répétitions et a obtenu six fausses alarmes. En fait, ces erreurs sont liées à la détection de deux répétitions de deux publicités successives et une répétition de trois publicités successives. Alors que dans les transcriptions manuelles, ces publicités ont été annotées comme motif distinct ce qui a causé les erreurs de détection et les fausses alarmes.

D'autre part, en utilisant l'algorithme basé sur BLAST, le système a besoin de 10 heures pour traiter les 24 heures de flux radio avec une machine 3.00GHz Intel Core 2 Duo 4 Go de RAM, tandis que pour la recherche exhaustive le temps est estimé à 10 jours pour traiter un jour de flux radio.

1.4.3 Segmentation et Regroupement en Locuteurs

Dans les tâches d'identification audio et la découverte des motifs audio, le but était d'indexer et identifier les morceaux de musique et les publicités. Pour montrer la généricité de notre système d'indexation audio, nous nous intéressons à un autre type de document audio, la parole, à travers la tâche de segmentation et regroupement en locuteurs (appelé aussi "diarization").

La segmentation et regroupement en locuteurs a pour objectif de segmenter un signal audio en régions homogènes selon l'identité des locuteurs afin de répondre à la question "Qui parle quand?". Cette tache est composée généralement de deux étapes. Une étape de segmentation qui consiste à trouver les frontières des segments de parole homogènes en détectant les points de changement acoustique. Les segmentent trouvés devraient contenir la parole d'un seul locuteur ou un signal audio autre que la parole (silence, bruit, jingle, musique, etc.). Dans l'étape de regroupement, les segments de parole ayant prononcés par

le même locuteur sont étiquetés avec le même identifiant.

Généralement, un système de segmentation et regroupement en locuteurs est composé de quatre étapes :

- Paramétrisation: le signal audio est transformé en une séquence de vecteurs, généralement les MFCCs
- Détection d'activité vocale: la segmentation du signal audio en segments de parole et de non parole en utilisant les vecteurs calculés dans l'étape précédente.
- Segmentation: segmentation des régions de parole en segments homogènes (du même locuteur).
- Regroupement: classer les segments obtenus selon l'identité de locuteur.

Dans nos travaux, nous nous sommes intéressés par la segmentation et regroupement en locuteurs des émissions radio et TV. Généralement, ces émissions ont tendance à garder la même structure avec les mêmes présentateurs, journalistes, effets sonores, jingles, etc. Cette redondance est utilisée pour améliorer la performance du système de diarization.

L'idée principale de notre système est de comparer l'émission à segmentés avec la même émissions diffusée à une date ultérieure afin de trouver les éléments audio similaires, comme les tours de parole prononcé par le même locuteur, le silence, le bruit, les jingles, la musique et les publicités. Cette opération est effectuée par l'intermédiaire du système d'indexation audio basé sur ALISP. En effet, une séquence de symboles ALISP est extraite de chaque document audio stocké dans la base de références. Un extrait audio de test inconnu est déterminé en comparant son empreinte ALISP avec celles de la base de références à l'aide de notre algorithme de recherche approximative des symboles ALISP. Ensuite, les segments identifiés sont étiquetés selon leur nature (parole, jingle, silence, etc.), déterminée avec les éléments de la base de références. Tandis qu'une étiquette "inconnu" est attribuée aux segments non identifiés. Enfin le signal du test pré-étiqueté est traité avec un détecteur d'activité vocale, un module de segmentation et un module de regroupement.

Ce système a été évalué lors de la campagne d'évaluation ETAPE 2011 [55] (Gravier el al., 2012). Cette campagne d'évaluation vise à évaluer les différents systèmes de traite-

Genre	Train	Dev	Test	Sources
Journaux TV	7h30	1h35	1h35	BFM Story, Top Questions (LCP)
Débats TV	10h30	2h40	2h40	Pile et Face, Ca vous regarde Entre les lignes (LCP)
Variétés TV	-	1h05	1h05	La place du village (TV8)
Emissions Radio	7h50	3h00	3h00	Un temps de Pauchon, Service Public Le masque et la plume, Comme on nous parle Le fou du roi
Total	25h30	8h20	8h20	42h10

Table 1.6: Base de données ETAPE : apprentissage (train), développement (dev), évaluation (test) [55].

ment de la parole à travers la reconnaissance automatique de la parole, la segmentation et regroupement en locuteurs, la détection de la parole multiples et la détection des entités nommés.

Comme le montre le tableau 1.6, les données ETAPE sont divisé en trois sous-corpus. Notez que le nombre d'heures sont rapportés en termes d'enregistrements, et non de tours de parole. Plus précisément 77 % des enregistrements contiennent de la parole. La mesure d'évaluation utilisée est le Diarization Error Rate (DER).

Le tableau 1.7 donne les valeurs DER pour le système de base (sans l'étape d'indexation audio basée sur ALISP) et le système proposé. Ce tableau montre que l'introduction du système d'indexation audio basée sur ALISP a amélioré les performances du système de diarization pour toutes les émissions TV et radio. Cependant, ces améliorations ne sont pas significatives pour toutes les émissions. Pour l'émission "LCP-TopQuestions-213800" l'amélioration relative de la DER est 84,62%, tandis que pour l'émission "EST2BC-ENG-FR-0910" elle est de 5,38 %. D'une façon plus générale, l'amélioration globale relative est de 34.37% et l'amélioration absolue de 8.5%.

D'autre part, notre système a eu les meilleures performances lors de la campagne d'évaluation ETAPE 2011, sachant que 7 institutions ont participé à la tâche de diarization dans cette campagne et que le plus grand DER était de 29.32%.

Emission	Baseline	ALISP
BFMTV-BFMStory-175900	19.30	15.87 (-17.77%)
LCP-CaVousRegarde-235900	20.70	12.60 (-39.13%)
LCP-EntreLesLignes-192800-1	24.77	17.31 (-30.11%)
LCP-EntreLesLignes-192800-2	27.19	18.48 (-32.03%)
LCP-PilesEtFace-192800	28.42	19.76 (-30.04%)
LCP-TopQuestions-000400	35.46	29.55 (-16.66%)
LCP-TopQuestions-213800	15.87	2.44 (-84.62%)
TV8-LaPlaceDuVillage-201300	37.86	22.27 (-41.22%)
TV8-LaPlaceDuVillage-172800	35.82	20.40 (-43.04%)
EST2BC-FRE-FR-1000	14.55	13.75 (-5.49%)
EST2BC-FRE-FR-1750	39.41	22.93 (-41.81%)
EST2BC-FRE-FR-2152-1	41.83	27.34 (-34.64%)
EST2BC-FRE-FR-2152-2	29.91	23.93 (-19.99%)
EST2BC-FRE-FR-0910	8.73	8.26 (-5.38%)
EST2BC-FRE-FR-2004	21.13	15.48 (-26.73%)
ETAPE-2011	24.73	16.23 (-34.37%)

Table 1.7: DER du système de base (baseline) et le système proposé (ALISP) avec le protocole d'évaluation ETAPE 2011.

1.4.4 Détection du Rire

Dans les sections précédentes, le système d'indexation audio basée sur ALISP a été appliqué sur l'identification audio et la découverte de motifs audio pour la musique et les publicités, et la segmentation et regroupement en locuteurs pour la parole. Dans cette section, une catégorie différente de document audio, appelée vocalisation non linguistique, est étudiée.

Malgré tous les efforts déployés au cours des deux dernières décennies dans les systèmes de reconnaissance de la parole, la détection des vocalisations non linguistiques comme le rire, le soupire, la respiration, l'hésitation semble encore une tâche difficile cite Weninger-ICASSP 2011 (Weninger et al., 2011). Ces vocalisations sont plus fréquentes dans les émissions radio et TV ou dans les conversations quotidiennes.

Dans nos travaux, nous nous intéressons à un type de vocalisation non linguistique bien précis qui est le rire. Le rire est un type de vocalisation non linguistique complexe qui communique des messages avec des significations différentes. En outre, le rire est un signal très variable (variabilité intra-locuteur et iner-locuteur).

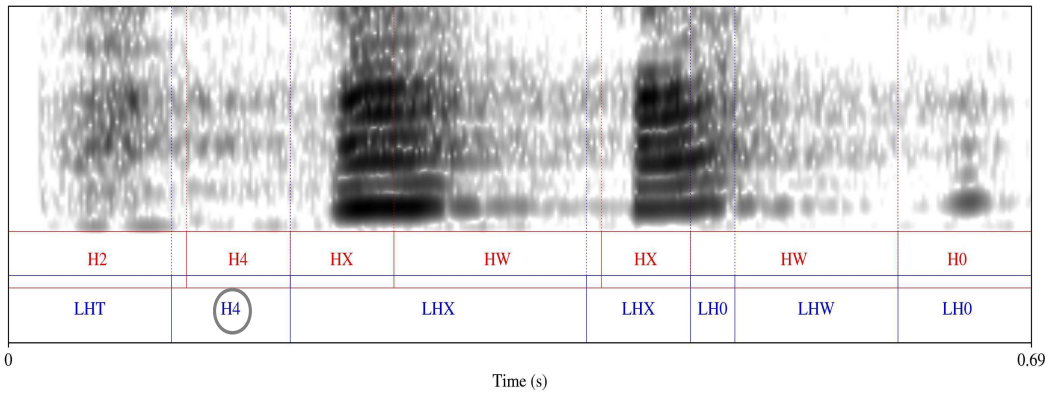


Figure 1.4: Segmentation ALISP d'un signal de rire obtenu par les modèles ALISP originaux (rouge) et par les ensembles de modèles spécifiques (bleu). Les symboles commençant par 'L' sont spécifique au rire et les autres symboles sont spécifiques aux éléments audio autre que le rire. Le symbole marqué par un cercle est une erreur de transcription qui pourrait être corrigée automatiquement avec un système de lissage.

Vu cette variabilité, une étape d'adaptation est ajoutée au système d'indexation audio basée sur ALISP pour améliorer les performances du système de détection du rire. Après avoir appris les modèles HMM ALISP originaux sur le corpus d'apprentissage YACAST (qui on été utilisés pour toutes les tâches précédentes), deux ensembles spécifiques de modèles HMM sont adaptés. Le premier est obtenu en adaptant les modèles ALISP originaux avec un corpus du rire et le deuxième avec un corpus qui ne contient pas de rire.

Les deux ensembles de modèles obtenus sont utilisés pour transformer le signal audio en une séquence de symboles en utilisant l'algorithme de Viterbi. La figure 1.4 illustre le spectrogramme d'un signal du rire el les transcriptions obtenus avec les deux ensembles de modèles spécifiques.

Après avoir transcrit le signal audio en symboles ALISP spécifique au rire et au non rire, une étape de lissage est réalisée pour corriger les éventuelles erreurs de transcriptions comme le montre la figure 1.4.

Afin d'évaluer le système proposé, les trois bases de données publiques, SEMAINE-DB [86] (McLeown et al., 2012), AVLaughterCycle [130] (Urbain et al., 2010) et Mahnob

laughter database [100] (Petridis et al., 2013) ont été utilisées. De plus, nous avons comparé notre système par rapport à des systèmes basés sur des modélisations GMM et HMM. Le tableau 1.8 montre la précision, le rappel et la F-mesure obtenus pour les différentes méthodes.

[%]	Précision	Rappel	F-measure
GMMs	70.8	78.6	74.5
HMMs en série	85.7	86.3	86.0
HMMs ergodique	92.8	84.5	88.5
ALISP-adapt	88.6	90.9	89.7
ALISP-adapt-sm3	92.4	92.7	92.6
ALISP-adapt-sm5	94.3	93.9	94.1

Table 1.8: Taux de précision, rappel et F-mesure pour les méthodes: GMM, HMM en série, HMM ergodique, le système proposé sans lissage (ALISP-adapt), le système proposé avec une fenêtre de lissage de taille 3 (ALIPS-sm3) et le système proposé avec une fenêtre de lissage de taille 5 (ALIPS-sm5).

Parmi les modèles acoustiques globaux, les HMM ergodiques performent mieux que les GMM et les HMM en série. Les HMM ergodiques montrent une grande précision (92,8%) à localiser les régions du rire, tandis que les HMM en série sont relativement mieux en rappel (86,3%). En se comparant avec ALISP-adapt, les HMM ergodiques sont toujours mieux de 4,2% de Précision. Cependant, ALISP-adapt obtient de meilleurs résultats en termes de F-mesure par rapport aux HMM globaux.

D'autre part, les modèles ALISP HMM avec une fenêtre de lissage offrent une flexibilité supplémentaire pour corriger les valeurs aberrantes à l'aide d'un système de vote majoritaire simple. Par conséquent, ALISP-adapter-SM3 et ALISP-adapter-sm5 montrent respectivement une amélioration en termes de F-mesure par rapport à ALISP-adapter de 2,9% et 4,4%. Dans l'ensemble, ALISP-adapter-sm5 a obtenu des performances relativement mieux que toutes les autres approches testé dans nos travaux.

1.5 Conclusions et Perspectives

Dans cette thèse, nous avons proposé un système générique d'indexation audio pour identifier la majorité des documents audio présents dans un flux radio. Ces documents sont: la musique, les publicités, la parole et les vocalisations non linguistiques (comme le rire, la toux, la vue, ...). De ce fait, le système d'indexation audio basé sur la méthode ALISP est appliqué pour différentes tâches qui sont : l'identification audio, découverte de motifs audio, segmentation et regroupement en locuteurs et la détection du rire. Le système proposé se compose de trois modules:

- Acquisition et modélisation des unités ALISP d'une manière unsupervisée
- Segmentation (aussi appelée transcription) ALISP, qui transforme les données audio en une séquence de symboles (en utilisant les modèles de Markov cachés ALISP précédemment acquis).
- Comparaison et décision qui comprend les algorithmes correspondants à la recherche approximative des séquences de symboles inspirées de la technique BLAST (Basic Local Alignment Search) et la distance de Levenshtein

Les principales contributions de cette thèse peuvent être divisées en trois parties:

1. Améliorer les outils ALISP en introduisant une méthode simple pour segmenter les données d'apprentissage en segments stables. Cette technique, appelée segmentation par stabilité spectrale, remplace la décomposition temporelle utilisée auparavant dans les outils ALISP. Le principal avantage de cette méthode est l'accélération du processus d'apprentissage non supervisé des modèles HMM ALISP.
2. Proposer une technique efficace pour la comparaison et la recherche des séquences ALISP utilisant l'algorithme BLAST et la distance de Levenshtein. Cette méthode accélère le processus de la recherche approximative des séquences de symboles sans affecter les performances du système d'indexation audio
3. Proposer un système générique pour l'indexation audio pour les flux radiophonique basé sur la segmentation ALISP. Ce système est appliqué dans différents domaines

d'indexation audio pour couvrir la majorité des documents audio qui pourraient être présents dans un flux radio.

L'évaluation du système pour la tâche d'identification audio en utilisant le protocole QUAERO 2010, montre la robustesse de l'empreinte ALISP par rapport aux autres systèmes. Pour la découverte de motif audio les résultats expérimentaux montrent que le système proposé est aussi performant que les systèmes utilisant les empreintes audio pour détecter des objets répétitifs dans les flux de radio. Pour la tâche de diarization le système a été évalué au cours de la campagne d'évaluation ETAPE 2011 et a obtenu les meilleurs résultats parmi les 7 participants. Enfin pour la détection de rire, les modèles HMM fournis par les outils ALISP ont obtenu de meilleurs résultats par rapport aux système utilisant des modélisations acoustiques globaux (GMM , HMM en série , HMM ergodique).

Les directions possibles de poursuite de ces travaux sont les suivantes. Tout d'abord, les informations sémantiques provenant des systèmes de reconnaissance de la parole pourront être exploitées pour améliorer les performances du système de segmentation et regroupement en locuteurs. De plus, un traitement parallèle pourrait être effectué afin d'accélérer le processus d'indexation et identification. En effet, le système proposé d'indexation audio pourrait être intégré dans un autoradio ce qui nécessite un traitement simultané de plusieurs stations radio. En outre, le calcul des MFCC, l'algorithme de Viterbi et la recherche approximative des séquences ALISP seront étudiés afin de détecter la partie qui pourrait être parallélisé et mis en oeuvre à l'aide des processeur graphique (GPU). Enfin, le système proposé pour la détection des vocalisations non linguistiques pourrait être aussi appliqués pour la détection des sons domestiques, telles que la fermeture des porte et le bruit des machines.

Chapter 2

General Introduction

2.1 Context and Motivation

For many decades, audio processing technologies have simplified the storage and accessibility to data. Actually, millions of audio documents are listened and hundreds of them are created every day. For example, more than 1 billion unique users visit YouTube each month and over 6 billion hours of video are watched each month on YouTube, that's almost an hour for every person on Earth[1]. However, there are not a lot of audio classification and retrieval tools to index, manage and characterize these data. Accordingly, few applications are developed to help users to search and browse the audio contents.

It was predictable that many researchers and industrials started focusing on audio indexing. There are some existing applications such as song classification, advertisement (commercial) detection, speaker diarization and identification, with various systems being developed to automatically analyze and summarize audio content for indexing and retrieval purposes. Within these systems audio data are treated differently depending on the applications. For example, song identification systems are generally based on audio fingerprinting using the energy and the spectrogram peaks such as SHAZAM and Philips systems. While speaker diarization and identification systems are using cepstral features and machine learning techniques such as Gaussian Mixture Models and/or Hidden Markov Models.

[1] http://www.youtube.com/yt/press/statistics.html

However, the diversity of the audio indexing techniques makes unsuitable the simultaneous treatment of audio streams where different types of audio content coexist. For example in radio streams, many types of audio data are found. These data are usually related to songs, commercials, jingles, speech and nonlinguistic vocalizations (such as laughter, sights and coughs). Therefore, a generic framework for audio indexing, retrieval and recognition is needed.

In this thesis we report our recent efforts in extending the ALISP (Automatic Language Independent Speech Processing) approach developed for speech as a generic method for audio indexing, retrieval and recognition. ALISP is a data-driven technique that was first developed for very low bit-rate speech coding, and then successfully adapted for other tasks such as speaker verification and forgery, and language identification. The particularity of ALISP tools is that no textual transcriptions are needed during the learning step, and only raw audio data is sufficient. In such a way any input speech data is transformed into a sequence of arbitrary symbols. These symbols can be used for indexing purposes.

2.2 Audio Indexing: Problematic

Audio indexing denotes the step in which relevant information is retrieved from unknown audio data. As shown in figure 2.1, such information, also referred as descriptive metadata, is usually linked to the type of audio content. Obtaining these metadata manually is tedious, time consuming, subjective and error-prone. Therefore, many systems are developed to automatically generate this information using minimal human intervention.

2.3 Contributions

As pointed out before, the general aim of this thesis is to use high-level information provided by ALISP tools for indexing purposes. In speech processing, high-level information represents the set of information that reflects the behavioral traits such as prosody, phonetic information, pronunciation, idiolectal word usage, conversational patterns, topics of conversations, etc. The main contribution of this thesis is the exploitation of the ALISP

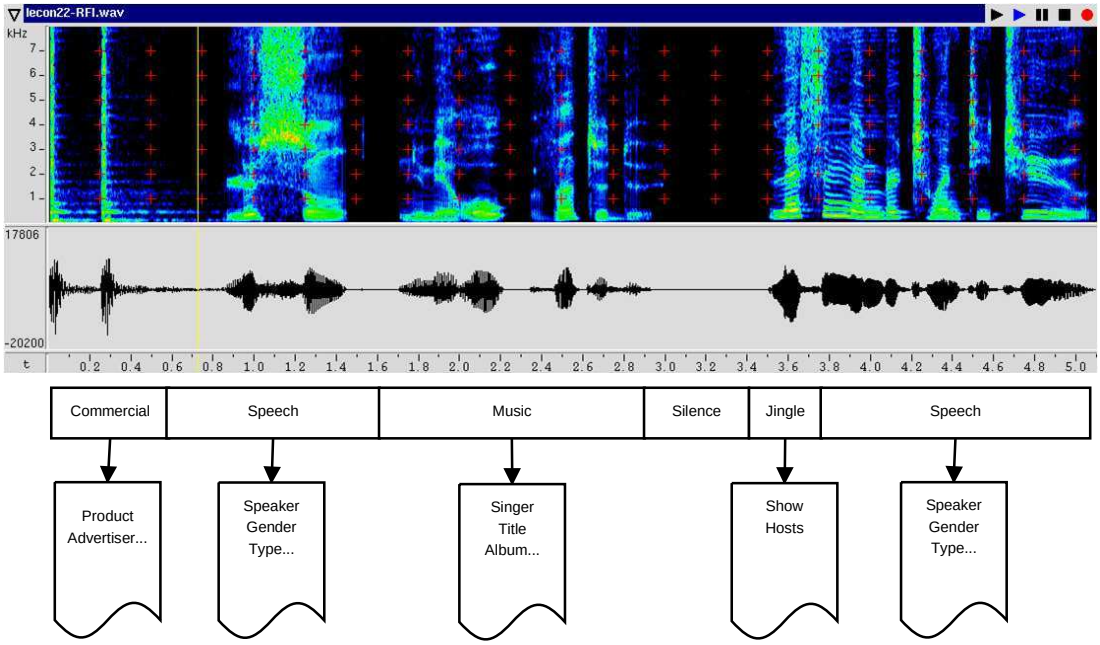


Figure 2.1: Audio indexing system.

approach as a generic method for audio (and not only speech) indexing and recognition. To this end, an audio indexing system based on the ALISP technique is introduced. The proposed architecture is composed of three modules:

- Automated acquisition (with unsupervised machine learning methods) and Hidden Markov Modeling (HMM) of ALISP audio models.
- Segmentation (also referred as sequencing and transcription) module that transforms the audio data into a sequence of symbols (using the previously acquired ALISP Hidden Markov Models).
- Comparison and decision module, including approximate matching algorithms inspired form the Basic Local Alignment Search (BLAST) tool widely used in bioinformatics and the Levenshtein distance, to search for a sequence of ALISP symbols of unknown audio data in the reference database (related to different audio items).

Our main contributions in this Ph.D can be divided into three parts:

1. Improving the ALISP tools by introducing a simple method to find stable segments within the audio data. This technique, referred as spectral stability segmentation, is replacing the temporal decomposition used before for speech processing. The main advantage of this method is its computation requirements which are very low comparing to temporal decomposition.

2. Proposing an efficient technique to retrieve relevant information from ALISP sequences using BLAST algorithm and Levenshtein distance. This method speeds up the retrieval process without affecting the accuracy of the audio indexing process.

3. Proposing a generic audio indexing system, based on data-driven ALISP sequencing, for radio streams indexing. This system is applied for different fields of audio indexing to cover the majority of audio items that could be present in a radio stream:

   - audio identification: detection of occurrences of a specific audio content (music, advertisement, jingle) in a radio stream;
   - audio motif discovery: detection of repeating objects in audio streams (music, advertisement, and jingle);
   - speaker diarization: segmentation of an input audio stream into homogenous regions according to speaker's identities in order to answer the question: "Who spoke when?";
   - nonlinguistic vocalization detection: detection of nonlinguistic sounds such as laughter, sighs, cough, or hesitation;

2.4 Thesis Structure

The thesis is organized as follows:

Chapter 2: State of the Art of Data-driven Speech Processing and Audio Indexing, focuses on the state of the art of data-driven speech processing and audio indexing. An overview of the techniques used to extract relevant information from unannotated speech data without using any linguistic information and rules is reported. Moreover the

ALISP data-driven segmentation method is presented. Finally a literature review of audio indexing systems based on fingerprinting is given.

Chapter 3: Databases, describes all the databases used in our work. They include the audio corpus provided by YACAST[2] that is used to train the ALISP models and to evaluate the ALISP-based audio indexing systems. Then, the ETAPE database used in the evaluation campaigns for automatic speech processing is described. Next, the MOBIO database exploited to evaluate our speaker verification system is presented. Finally, the databases needed for laughter detection are described.

Chapter 4: Contributions to Data-driven Audio Indexing, presents the main contributions of our work. The first contribution is related to the ALISP segmenter where the temporal decomposition is replaced by a simpler technique to find stable segments within the audio data. Second, an efficient technique to retrieve relevant information from ALISP sequences is proposed. Third, a generic audio indexing system, based on data-driven ALISP sequencing is developed, to cover the majority of audio items that could be present in a radio stream (song, advertisement, audio motif, speaker turn, laughter).

Chapter 5: Audio Identification, presents the ALISP-based audio indexing system applied to the audio identification task. Experimental studies about the number of Gaussian components, number of ALISP units and the method used for the initial segmentation are reported. Moreover, a comparison of the performances of our system with the systems participating in the 2010 QUAERO evaluation campaign is given.

Chapter 6: Audio Motif Discovery, describes the exploitation of the ALISP-based audio indexing system for audio motif discovery. Related works to audio motif discovery are presented. In addition, the evaluation of the proposed method is given. This evaluation involves repeating songs and advertisement detection in radio streams.

Chapter 7: Speaker Diarization, reports the use of the ALISP-based audio indexing systems to perform speaker diarization. First, an overview of methods used for speaker diarization is given. Moreover, the performances of the proposed system in the ETAPE 2011 evaluation campaign are given. Finally, the evaluation of the proposed speaker verification

[2] http://www.yacast.fr/fr/index.html

system in the MOBIO 2013 evaluation campaign is reported.

Chapter 8: Nonlinguistic Vocalizations Detection, deals with the detection of laughter using the high-level information provided by the ALISP segmenter. A generic framework to detect nonlinguistic vocalizations is proposed. The evaluation of the system is performed on three publicly available databases.

Finally, Chapter 9 closes this thesis with conclusions, discussions and perspectives.

Chapter 3

State of the Art of Data-driven Speech Processing and Audio Indexing

3.1 Introduction

The main purpose of our work is to exploit data-driven approaches, usually applied for speech processing, to develop a generic audio indexing system. In this chapter two states of the art are reviewed. The first one is related to data-driven approaches for speech processing while the second one deals with audio indexing.

Two categories of speech and language processing systems could be found in the literature:

- Supervised systems that use linguistic information and rules.
- Unsupervised systems that exploit machine learning techniques to extract relevant information from a set of representative examples.

The first category requires the availability of a number of linguistic information, such as phonetic inventories, lexicons and language models, and annotated training corpora consisting of manual transcriptions of speech data. While such systems have proven its robust-

ness and effectiveness for many problems where the human contributions are essential and labeled speech data are easy available, they have many disadvantages, such as:

- Language dependency: supervised systems are developed for a specific language, which make them not portable across languages.
- Human effort: significant human expertise is required to acquire the knowledge base such as pronunciation lexicon and labeled speech data.
- Diversity of linguistic models: representing all linguistic rules and phenomena using a common theory of linguistic seems to be impossible.

On the other hand, unsupervised systems (also referred as data-driven systems or zero resource systems) do not require transcriptions, annotations nor prior linguistic knowledge. The amount of available speech data, such as broadcast news archives, radio recordings, podcasts or various internet media is constantly increasing. Therefore, most of these systems exploit machine learning techniques to automatically determine the linguistic units and information required from representative examples of data.

The second part of this chapter deals with audio indexing. Audio indexing denotes the step in which relevant information are retrieved from unknown audio data. In our work we are interested in a particular field of audio indexing, which is the audio identification (known also as audio detection or audio information retrieval).

Audio identification involves detecting (and eventually locating) occurrences of a specific audio content (music, advertisement, jingle,...) in audio streams or audio database. In the literature the majority of proposed audio identification systems rely on the same underlying concept: audio fingerprinting. An audio fingerprint is a compact content-based signature that represents an audio recording. This technique consists of two parts: a fingerprint extraction module and a comparison module. First a fingerprint is extracted from each audio document stored in a reference database. An unlabeled audio excerpt is identified by comparing its fingerprint with those of the reference database.

In this chapter, the state of the art of unsupervised techniques for speech processing is reviewed. Then the adopted data-driven system based on the Automatic Language In-

dependent Speech Processing (ALISP) method is detailed. After that an overview of the application of the ALISP approach to speech processing, in particular, very low bite rate speech coding, speaker verification, voice forgery and language identification is presented. Finally a literature review of audio indexing based on fingerprinting techniques is presented.

3.2 Toward Unsupervised Techniques for Speech Processing

Automatic Speech Recognition is the most mature field of speech processing. Developing speech recognition systems requires the availability of large speech corpora with the corresponding world-level annotations. Huge linguistic resources associated with the constantly increasing computational and storage power have significantly reduced the word error rates on increasingly challenging tasks is speech processing [16] (Beyerlein et al., 2002) [84] (Martin and Garofolo, 2007) [38] (Deligne et al., 2002).

Supervised techniques provide good performances for scenarios where human expertise and annotated data are available. However, they remain ineffective when transcribed data are not available. Therefore, many frameworks are proposed to develop increasingly unsupervised data-driven systems which are less reliant on linguistic expertise and annotated corpora.

In [53] (Glass, 2012), speech processing techniques are divided into four groups depending on the scenario for which they are applied. Each scenario requires decreasing amount of human expertise and annotated resources, and increasing amount of unsupervised learning. These groups are illustrated in figure 3.1 [53] (Glass, 2012).

3.2.1 Expert-based Speech Processing

Expert-based speech processing denotes systems that use human expertise associated with annotated speech corpora. Human expertise is often provided in the form of a pronunciation lexicon that gives the relation between vocabulary words and their associated sub-word unit realizations. This scenario represents the most developed speech recognition system using the Hidden Markov Model (HMM) to represent the speech data [11] (Baker et al., 2009).

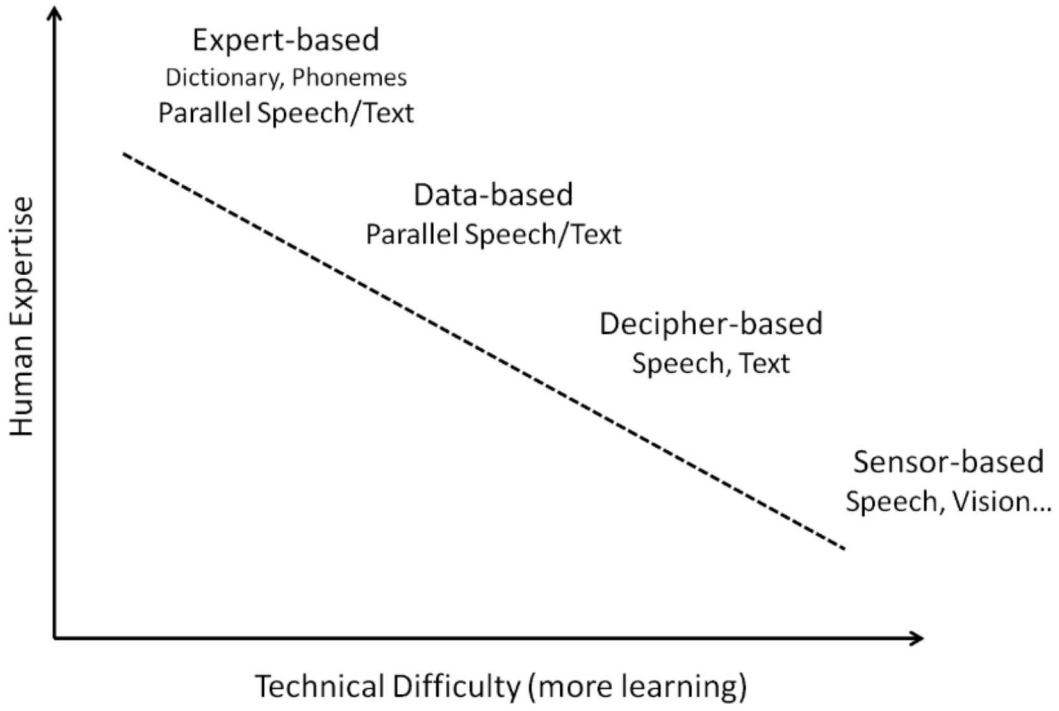


Figure 3.1: Potential scenarios for speech processing depending on human expertise and unsupervised training.

3.2.2 Data-based Speech Processing

Data-based speech processing systems aim to learn the pronunciation lexicon and provide automatically linguistic information, if no language expert knowledge (e.g. phonetic transcriptions) is available or affordable.

In [70] (Killer et al., 2003) the authors automate the pronunciation dictionary creation process for languages that have straightforward world-to-sound mappings (English, Spanish, German). A grapheme-based approach is proposed to develop a speech recognition system. Pronunciation dictionaries for grapheme based recognizers are built by simply splitting a word into its graphemes. Each grapheme is modeled by a 3-state Hidden Markov Model (HMM) consisting of a begin, a middle, and an end-state. The evaluation of this system shows that English has the worse correspondence between phonemes and graphemes while the best one is relative to Spanish language.

When a straightforward letter-to-sound mapping is not possible, a pronunciation mixture model could be used to perform a grapheme-to-phoneme conversion [10] (Badr et al., 2011). A joint-multigram approach [19] (Bisani and Ney, 2008) is employed to model the relationship between graphemes and phonetic units and to build a pronunciation mixture model. The evaluation of this approach shows that learned lexicons outperform expert, hand-crafted lexicons for a weather information retrieval spoken dialogue system and for the academic lectures domain.

Data-based scenarios could also involve the combination of annotated data from several languages to easily adapt the speech processing parameters to a new language [118] (Schultz and Kirchhoff, 2006). Moreover, in [54] (Gollan et al., 2007), the authors propose to combine untranscribed and transcribed data to improve the performances of a speaker adaptive acoustic model. Furthermore in [94] (Novotney et al., 2009), initial models are trained from a small amount of transcribed data. Then these models are used to decode a larger amount of speech data. Finally, new models are iteratively learned from these automatic transcripts.

3.2.3 Decipher-based Speech Processing

A more challenging scenario consists of building an automatic speech processing system from speech only corpus combined with non-parallel text data. The decipher-based speech processing systems represent a major breakthrough from conventional speech recognition systems.

In the last few years, many systems are developed to automatically extract useful information from not annotated speech data. This scenario is generally referred to a zero resource scenario. Most of these systems are used to identify word-like patterns in the speech signal by extracting recurrent speech sequences.

In [97] (Park and Glass, 2008), an unsupervised speech pattern discovery framework is proposed. It is based on a segmental variant of Dynamic Time Warping, which is used to search for matching acoustic patterns between spoken utterances. Similar acoustic sequences are grouped together to form clusters corresponding to lexical entities such as words and short multiword phrases. The evaluation of the proposed system on a corpus of academic lecture material shows that the obtained clusters are relevant to summarize the audio stream.

Muscariello et al. [92] (Muscariello et al., 2012) develop a similar system to extract speech motifs or patterns by unsupervised word discovery. The proposed system is based on a template matching technique to identify recurrent acoustic segments (using segmental Dynamic Time Warping metric combined with self-similarity matrix). A searching strategy based on the ARGOS framework [61] (Herley, 2006) to detect repetitions is designed. It consists of a sequential algorithm to find repetition in audio stream. First, the speech signal is divided into two parts: the query pattern and the past stream. Then the query pattern is searched in a library of motifs that are already extracted. After that, if a positive match is found a new occurrence of the corresponding motif is created, otherwise the pattern is searched in the past stream. Finally, if a positive match is found in the past stream, an extension of the query matching is performed to find the entire occurrence. The proposed system is evaluated on a French radio broadcast data and shows good results.

Another system using a Gaussian posteriorgram based representation for unsuper-

vised discovery of speech patterns is proposed by [136] (Yaodong and Glass, 2010). This framework is composed of three steps. First a Gaussian posteriorgram technique is performed to train an unsupervised Gaussian Mixture Model and associates each speech frame with a Gaussian posteriorgram representation. Then the segmental Dynamic Time Warping metric is used to detect similar sequences of Gaussian posteriorgram vectors. Finally a graph clustering procedure is carried out to group similar segments into clusters.

In addition to systems described above, many other frameworks are developed to detect automatically speech patterns in large audio corpora. In [122] (Siu et al., 2011) an unsupervised Hidden Markov Model-based recognizer system is built to convert speech data into self-organized units which are used to detect common audio patterns. Moreover an unsupervised speaker recognition system is proposed in [67](Kanthak and Ney, 2003), that combines grapheme-based units with multilingual acoustic modeling. Furthermore, a Polish speech recognition system is developed by combining the exploitation of the cross-language bootstrapping and confidence based unsupervised acoustic model training [79] (Loof et al., 2009). In addition, a multigram model is proposed in [37] (Deligne et Bimbot, 1997) to retrieve sequential variable-length regularities within streams of text data, which are exploited for automatic speech recognition. Speech motif discovery is useful for several applications, including spoken term detection [91] (Muscariello et al., 2011), nonnegative convolutive sparse coding [134] (Wang et al., 2011), topic segmentation [82] (Malioutov et al., 2007), topic classification [51] (Gish et al., 2009), spoken corpus summarization [60] (Harwath et al., 2013) and unit learning [64] (Jansen and Church, 2011).

In [28] (Chollet et al., 1999), a data-driven system, referred as Automatic Language Independent Speech Processing (ALISP) is proposed. ALISP method consists in segmenting the speech into data-driven speech units, denoted in this chapter and in the followings as ALISP units (or data-driven units or pseudo-phonemes). These units are automatically determined from the training corpus with no need of phonetic transcriptions and textual annotations of the corpus. As pointed out before, our objective through this thesis is to exploit high-level information for audio indexing by using data-driven units. To this end, we selected Automatic Language Independent Speech Processing (ALISP) tools as they

are versatile and have already been used in different applications for speech processing. A detailed description of these tools is given in section 3.3.

3.2.4 Sensor-based Speech Processing

In the sensor-based speech processing scenario, speech data are associated with other modalities such as vision. Most of the systems belonging to this category are relative to human-machine interaction. The aim is to jointly learn linguistic and perceptual models of semantic concepts [114] (Roy and Pentland, 2000). For example, it would be appropriate to teach robots new concepts through spoken interactions in a new environment.

3.3 Data-driven ALISP Segmentation

In the previous section, we described many techniques that are used to acquire relevant information from untranscribed speech data. For our work, we decided to use the ALISP method to exploit the resulting data-driven units for audio indexing purposes.

ALISP tools are selected as they are versatile and have already been used in different applications. First, they were used in Very Low Bit Rate coding based on recognition-synthesis [26] (Cernoky, 1998) [96] (Padellini et al., 2005). The second application was the use of those units for segmental speaker verification [40] (ElHannani et al., 2009) [39] (El-Hannani, 2007) [102] (Petrovska-Delacrétaz et al., 2000). Then, it was applied for voice forgery [99] (Perrot et al., 2005). They were also exploited for automatic language identification [29] (Chollet et al., 2005).

The set of ALISP units is automatically acquired through parameterization, temporal decomposition, vector quantization, and Hidden Markov Modeling as shown in figure 3.2. We detail hereafter each component of the figure.

3.3.1 Parameterization

The parameterization of audio data is done with Mel Frequency Cepstral Coefficients (MFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame, Hamming

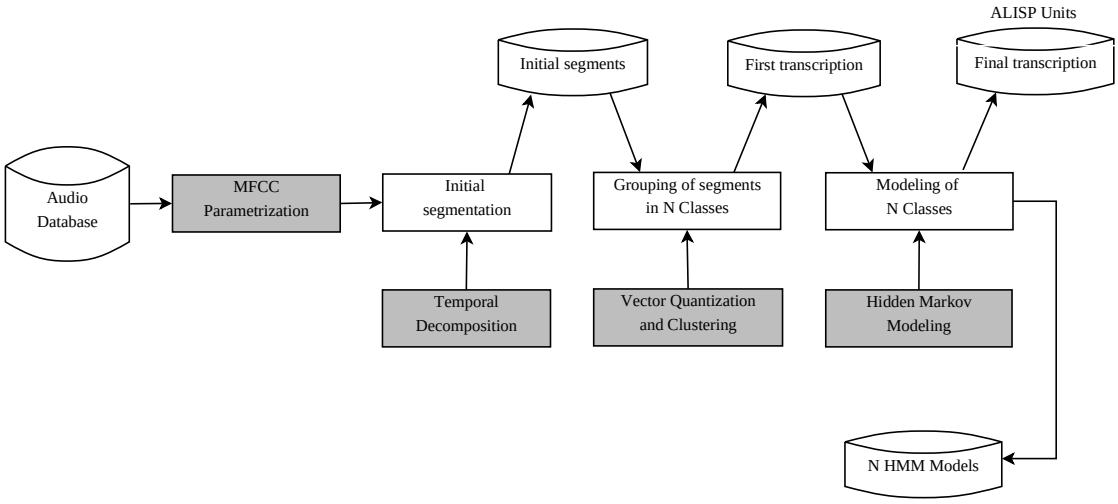


Figure 3.2: Automatic Language Independent Speech Processing (ALISP) units acquisition and their HMM modeling.

window is applied and a cepstral vector of dimension 15 is computed and appended with first and second order deltas.

3.3.2 Temporal Decomposition

After the parameterization step temporal decomposition is used to obtain an initial segmentation of the audio data into quasi-stationary segments. This method is introduced originally by Atal [8] (Atal, 1983) as nonuniform sampling and interpolation procedure for efficient parameter coding.

Temporal decomposition approximates a matrix X of N successive parameter vectors of dimension P by H target vectors a_{ph} with associated interpolation functions $\varphi_h(t)$). The trajectory of x_t^p, the p^{th} parameter of the t^{th} frame, is approximated as follows:

$$\hat{x}_t^p = \sum_{h=1}^{H} a_{ph}\varphi_h(t) \qquad p = \{1,\ldots,P\} \tag{3.1}$$

The previous equation can be written more compactly using matrix notation:

$$\hat{X} = A\Phi \tag{3.2}$$

where $\hat{X}$ is the approximated parameter matrix, A is the target-vectors matrix, and Φ is the interpolation functions matrix.

The procedure to find targets and interpolation functions consists of the following steps:

1. Initial search of interpolation functions using adaptive rectangular window and local singular value decomposition.
2. Post-processing of interpolation functions: smoothing, de-correlation and normalization.
3. Target vectors computation: $A = X\Phi^{\#}$, where $\Phi^{\#}$ is the pseudo-inverse of initial interpolation functions matrix.
4. Local adaptive refinement of interpolation functions and targets by iterations minimizing the distance of X and $\hat{X}$.

The detailed algorithm can be found in [18] (Bimbot and Atal, 1991).

Once interpolation functions are computed, their intersections are used to determine segment boundaries. The audio segments correspond at this point to spectrally stable portions of the signal. These segments will be further clustered using Vector Quantization. Then, boundaries together with labels will be used as initial transcription for Hidden Markov Modeling. This step corresponds, in traditional phonetic recognizer systems, to the use of phonetically transcribed data to initialize phone models.

3.3.3 Vector Quantization

The next step in the ALISP process is the unsupervised clustering procedure performed via vector quantization [81] (Makhoul et al., 1985). This method maps the P-dimensional vectors of each segment provided by the temporal decomposition into a finite

set of L vectors $Y = \{y_i; 1 \leq i \leq L\}$. Each vector y_i is called a code vector or a codeword and the set of all codewords is called a codebook. The codebook size L defines the number of ALISP units. Codebook training is performed using vectors located in gravity centers of segments computed with temporal decomposition (one vector per segment).

This training is done by a K-means algorithm with binary splitting. This method, called Linde-Buzo-Gray or LBG [77] (Linde et al., 1980), results in a codebook size which is power of 2. The LBG procedure is as follows:

1. Compute the initial centroid as the average of all vectors in the training set.
2. Split each centroid into two by moving it in opposite directions. This is done by adding small noise values $\pm \varepsilon$.
3. Redistribute vectors between the two centroids using the nearest neighbor rule.
4. Compute new positions for the two centroids by obtaining the average of their respective clusters. Then iterate in Step 3 until the change of the average distortion is relatively small.
5. Go to step 2 if the desired codebook size not yet reached, otherwise terminate.

The initial labeling of the entire audio segments is achieved by assigning segments to classes using minimization of the cumulated distances of all the vectors x_t from the audio segment to the nearest centroid of the codebook.

$$y_s^l = \min_i \left[\sum_{t=b_s}^{e_s} d(x_t, y_i) \right] \tag{3.3}$$

where s denotes a particular segment with the beginning b_s and the end e_s. All vectors in segment s are labeled with the label l of the winner centroid. The result of this step is an initial segmentation and labeling of the training corpus.

3.3.4 Hidden Markov Modeling

The final component in figure 3.2 represents the Hidden Markov Modeling procedure. The objective here is to train robust models of ALISP units on the basis of the

initial segments resulting from the temporal decomposition and vector quantization steps. HMMs training is performed using the HTK toolkit [1]. It is mainly based on Baum-Welch re-estimations and on an iterative procedure of refinement of the models that may be summarized as follows:

1. Initialization of parameters: This step provides initial estimates for the parameters of HMMs using a set of observation sequences. First, a prototype HMM definition must be specified in order to fix the model topology. In this system, each ALISP unit is modeled by a left-right HMM having three emitting states with no skips. Covariance matrices are diagonal, and computed for each mixture. The initialization of models is performed via HInit tool. Let each audio segment be represented by a sequence of feature vectors or observations defined as $O = \{o_1 \ldots o_T\}$, where o_t is the feature vector observed at time t. HInit first divides the training observation vectors equally amongst the model states and then initializes values for the mean and variance of each state j using the equations 3.4 and 3.5:

$$\hat{\mu}_j = \frac{1}{T} \sum_{t=1}^{T} o_t \tag{3.4}$$

$$\hat{\Sigma}_j = \frac{1}{T} \sum_{t=1}^{T} (o_t - \mu_j)(o_t - \mu_j)' \tag{3.5}$$

2. Context independent re-estimation: The initial parameter values computed by HInit are then further re-estimated by HRest tool using the Baum-Welch re-estimation procedure. In the contrary of HInit in which each observation vector o_t is assigned to a unique state, HRest assigns each observation to every state in proportion to the probability of the model being in that state when the vector was observed.

   Thus, if $P_j(t)$ denotes the probability of being in state j at time t then the equations 3.4 and 3.5 given above become the following weighted average:

$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} P_j(t) o_t}{\sum_{t=1}^{T} P_j(t)} \tag{3.6}$$

$$\hat{\Sigma}_j = \frac{\sum_{t=1}^{T} P_j(t)(o_t - \mu_j)(o_t - \mu_j)'}{\sum_{t=1}^{T} P_j(t)} \tag{3.7}$$

where the summations in the denominators are included to give the required normalization. Equations 3.6 and 3.7 are the Baum-Welch re-estimation formula. The probability of state occupation $P_j(t)$ is computed using the so-called Forward-Backward algorithm.

3. Context dependent re-estimation: This re-estimation step uses the same Baum-Welch procedure as for the context independent re-estimation but rather than training each model individually all models are trained in parallel. This re-estimation is done by HERest tool. For each training utterance, the corresponding segment models are concatenated to construct a composite HMM which spans the whole audio segment. This composite HMM is made by concatenating instances of the ALISP classes HMMs corresponding to each label in the transcription. The forward-backward algorithm is then used to accumulate, for each HMM in the sequence, the statistics of state occupation, means, variances, etc. When all of the training data has been processed, the accumulated statistics are used to compute re-estimates of the HMM parameters. It is important to emphasize that in this process, the transcriptions are only needed to identify the sequence of labels in each segment. No segment boundary information is needed.

4. Model refinement: This step consists in an iterative refinement of these HMMs by successive segmentation of the training data followed by re-estimations of parameters. The segmentation is performed using the HVite tool which is based on the Viterbi algorithm called the Token Passing Model [137] (Young et al., 1989). HVite matches an audio file against a network of HMMs and outputs its transcription. A simple grammar, in which each class can follow any other class, is used for decoding. The procedure of refinement can be summarized as follows:

    - use the previous models to segment the training data to produce new transcriptions. For the first iteration the models used are the one obtained in step 3;

- re-estimate new set of HMMs parameters using transcriptions obtained in the previous step. Old parameters are used as initial values and the re-estimation procedure used in this step is the one described in step 3;
- if the maximal number of iterations is reached (in this work 8) stop the refinement procedure, otherwise return to the first step;

5. Final refinement: This final step of the Hidden Markov Modeling aims at incrementing the number of mixture components in each model in a dynamic manner. The operation of increasing the number of components in a mixture is done by a process called mixture splitting using the HHEd tool. The procedure of final refinement may be summarized as follows:

   a. denote HMMs parameters λ^m and the list of models Γ^m. Set the iteration number m to 0. λ^0 corresponds to the parameters resulting from the first refinement and the list Γ^0 contains all ALISP models;

   b. increment the number of mixture components in HMMs for each ALISP classes in the list Γ^m. Denote new parameters Γ^{m+1};

   c. re-estimate the new set of HMMs parameters λ^{m+1} using transcriptions obtained at the end of the first refinement;

   d. perform a forced alignment of the all training data using λ^{m+1};

   e. for each ALISP class, compute the difference of recognition likelihoods using λ^{m+1} and λ^m. Update the ALISP list Γ^{m+1} by removing all ALISP classes for which the likelihood difference is relatively small;

   f. terminate the procedure if Γ^{m+1} is empty or otherwise return to b;

The resulting HMM models will then be used to transcribe any incoming audio data. This transcription will be referred in this chapter and the following as ALISP segmentation (or ALISP sequencing or ALISP transcriptions). Figure 3.3 shows the spectrogram of the sentence "Bonjour Christophe" and its ALISP transcription.

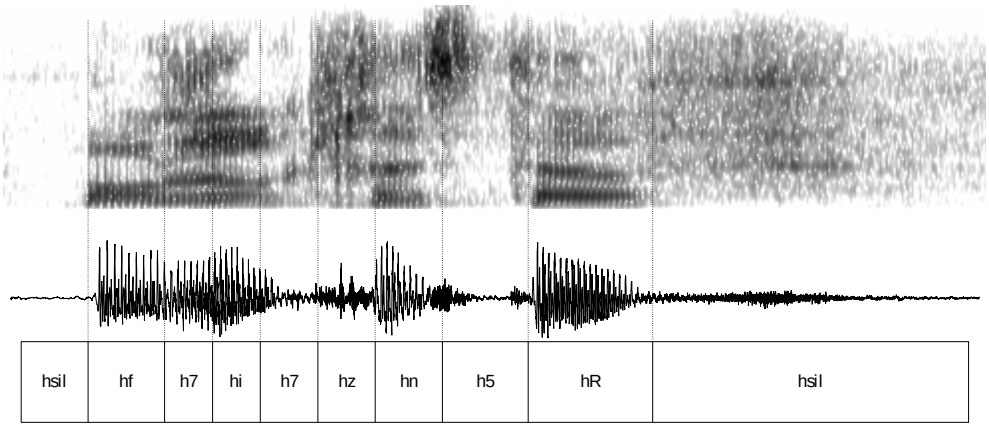


Figure 3.3: Spectrogram of a French speech sentence "Bonjour Christophe" and its ALISP transcription (hf, h7, hz,... are the name of ALISP units).

3.4 ALISP-based Speech Processing

As mentioned before, ALISP is a data-driven technique that was first developed for Very Low Bit Rate speech coding, and then successfully adapted for other tasks such as speaker verification and forgery and language recognition. In this section the ALISP-based speech processing systems are shortly described.

3.4.1 Very Low Bite Rate Speech Coding

Most of speech coding systems that achieve a bit rates lower than 600 bits/s are based on a recognition-synthesis approach. The ALISP-based Very Low Bit Rate speech coding system [26] (Cernoky, 1998) [96] (Padellini et al., 2005) is composed of two phases:

- Encoding phase: The Viterbi algorithm is used to segment the speech file to be coded using the ALISP HMM models. Then, the prosodic information is extracted from each resulting segment. This information is used to find the nearest synthesis unit in the reference codebook. The synthesis codebook is organized in such a way that each class (the ALISP unit) is associated with the previous identified class to take

into account the backward context information.

- Decoder phase: From the unit indexes sent by the encoder, synthesis units are found in the reference codebook and concatenated using the Harmonic plus Noise Model algorithm [123] (Stylianou, 1996) to recover the target speech file.

3.4.2 Speaker Verification

In [40] (ElHannani et al., 2009) [39] (ElHannani, 2007) [102] (Petrovska-Delacrétaz et al., 2000), high-level features derived from the speech data using the ALISP segmentation are used to develop three speaker verification systems:

1. Idiolectal system: In this approach, only the labels associated to the ALISP segments are used as source of information. The speaker models and the background model are computed using a simple n-gram frequency count. The background model is estimated using a large number of speakers while the speaker models are obtained by adapting the background model. In the evaluation phase, each ALISP-sequence is tested against the speaker specific model and the background model using a likelihood ratio.

2. ALISP language models system: The symbol sequences produced by data-driven ALISP tools are used to train ALISP n-grams models. These models are built as follows. Firstly, the training text is scanned and the n-grams are counted and stored in a database of gram files. Secondly, the resulting gram files are used to compute n-gram probabilities which are stored in the language models file. In the evaluation phase, the test file is transcribed using the ALISP tools. Then a log-likelihood ratio is computed to obtain the recognition score.

3. Duration models system: In this system, the duration of the ALISP units are used as features to model speakers. The duration of each ALISP unit is extracted and used to train background models. Each speaker is represented by 64 GMMs each of them models the duration of an ALISP class. The speaker specific 64 models are adapted from the 64 ALISP class dependent background models. During the evaluation phase, the duration vectors of the test file are extracted. Then, the test duration vectors are

compared to the hypothesized speaker model and to the background model of the corresponding ALISP class using the log-likelihood ratio.

The proposed systems were evaluated on English trials from NIST 2006 SRE and compared with phonetic approaches. It was shown that data-driven units provide better results than phonetic approaches.

3.4.3 Voice Forgery

Voice forgery aims to covert the voice of an arbitrary person (the impostor), in such a way that it seems to be the voice of another person (the client). In order to automatically transform the voice, the ALISP-based speech coder is used [99] (Perrot et al., 2005). First, speech corpus of the client is used to train the ALISP HMM models which provides a segmentation of this corpus. Then Harmonic plus Noise Model parameters are extracted from each segment of the client speech. After that, the impostor voice is encoded using the ALISP codebook of the client. Finally, in the decoding phase, synthesis units of the client voice are used to build the transformed speech signal.

3.4.4 Language Identification

In this part we report about the application of the ALISP data-driven segmentation method for Automatic Language Identification task [29] (Chollet et al., 2005). Two ALISP-based systems are developed to perform this task:

1. ALISP HMM based system: In this system, each test utterance is decoded by all language-dependent ALISP-recognizers, producing a transcription into ALISP-units along with their log-likelihood scores. For a given language, these segmental scores are summed up and normalized by the utterance length to produce a score for the test utterance. In this summation process, the segments previously identified as silence are simply skipped. Finally, each score produced for a language is divided by the mean of the other languages scores.

2. ALISP N-gram language models system: For each language, statistics of the 2-grams occurring in the transcription produced by that language recognizer are gathered. For each 2-gram ALISP sequence in the test utterance, its probability in the given language is divided by the mean of its probability in the other language models. The final score is the sum of the 2-gram probabilities normalized by the number of 2-grams in the test utterance.

At this point, ALISP method was only used for speech processing. It was exploited for very low bite rate speech coding, speaker verification and forgery and language identification. In this thesis we report our efforts in applying the ALISP approach as a generic method for audio (and not only speech) indexing and recognition. The next section presents the state of the art of audio indexing based on audio fingerprinting.

3.5 Audio Indexing Based on Fingerprinting: State of the Art

Audio indexing denotes the step in which relevant information are retrieved from unknown audio data. In our work we are interested in a particular field of audio indexing, which is the audio identification based on fingerprinting (known also as audio detection or audio information retrieval).

The general architecture of an audio identification system is described in figure 3.4. This figure shows that an audio identification system based on audio fingerprinting consists of 2 modules:

- A fingerprint extraction module
- A comparison module

The first step in an audio fingerprinting system is to create a fingerprint database from a reference database. The reference database contains audio files (Music pieces, jingles, advertisements,...) to be identified. In the second step an unlabeled audio excerpt is identified by comparing its fingerprint with those of the reference database.

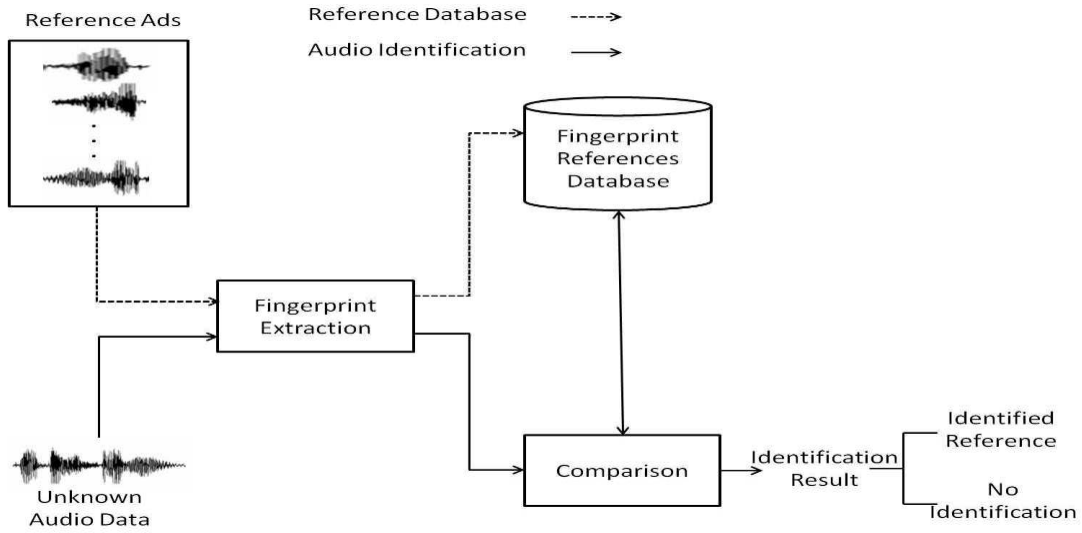


Figure 3.4: Audio identification system based on audio fingerprinting.

3.5.1 Properties of Audio Fingerprinting

An audio fingerprinting system should meet certain properties in order to take into account several requirements. The importance of each requirement varies depending on the application. Such systems have to be computationally efficient and robust. The requirements are the following [25] (Cano et al., 2005):

- Accuracy: The number of correct identifications, missed identifications and false alarms (wrong identification).

- Robustness: The ability of the system to work satisfactorily under the presence of different types of degradation.

- Granularity: The minimum duration of the query fingerprint needed to uniquely identify the audio file. For example, the average duration of advertisements varying from 5 to 30 seconds, it is necessary to have granularity less than 5 seconds.

- Complexity: The complexity of the system determines the computational costs.

They include costs needed to create the fingerprint, to search for it and to add a new entry in the existing reference database.

- Scalability: Performance in the presence of larger database. This has direct relationship with the complexity and the accuracy of the system.

Improving one parameter often puts additional constraints on other parameters. To improve the robustness of the system, complex features need to be computed which affects complexity and scalability. Using large databases often affects the accuracy of the system. Due to these reasons, trade-offs need to be made to find an optimal solution which has satisfactory performance with respect to all parameters.

3.5.2 Audio Degradations

Even though the audio files are made available to the audio identification system for building the reference database, the unknown audio signal (especially broadcasted audio) goes through several processes which degrade the quality to a certain extent. A very diverse panel of audio degradations are reported in the literature, designed to reproduce most of the audio effects that can be applied to an audio signal, affecting its quality, without changing its semantic content. Some of the degradations which occur quite often are listed in [24] (Cano et al., 2002) and reported below:

- Dynamic Amplitude compression: Affecting the dynamic range of the signal, these effects are used in order to ensure the headroom of digital systems while avoiding clippings.

- Channel filtering: When the audio is broadcasted through a channel, it passes through various filters which change the frequency spectrum of the audio.

- Real world noise addition: Due to poor transmission quality, but also sounds superposed on the original document such as speech utterance in the beginning of a music track.

- Pitching: Playing songs or advertisements in a radio broadcast faster or slower, where both the pitch and the tempo change. Sometimes stations use pitching up to 2,5% to achieve two goals: playing more songs per hour and getting more attraction for listeners.

- Equalization: Boosting or cutting the level of certain audio frequency bands compared to other bands.

- Perceptual Audio Coding: With the increasing amount of available audio, various compression techniques are used to store the audio clips such as *.mp3 and *.aac format. Compression techniques degrade the quality of audio, but perception of the clip remains the same.

3.5.3 Literature Review of Audio Fingerprinting Systems

There have been many studies in the field of fingerprinting of audio using different features. Several fingerprinting systems have been reviewed by Cano et al. in [25] (Cano et al., 2005). The main challenge of these systems is to create a robust fingerprint against different types of distortions and to propose a fast matching method that can satisfy real-time requirements regardless the size of the reference database.

By reviewing papers published on the subject, we group them in three main families based on the strategy of fingerprint extraction:

- Spectral representation techniques: These methods are generally based on the division of the spectrum into sub-bands.

- Computer vision techniques: These systems involve the processing of the audio signal as a 2-D image. They are usually using the wavelet transform to extract the audio fingerprint.

- Machine learning techniques: This family includes approaches based on vector quantization and data-driven techniques. These systems propose a fingerprint model that mimics the modeling and classification techniques used in speech processing.

We present audio fingerprinting systems according to the three families identified above. First, we describe few works that provide fingerprint models based on the spectral representation of the signal, then we present some approaches that rely on computer vision techniques and we end up by with works based on machine learning approaches.

3.5.3.1 Spectral Representations Techniques

Most of audio identification systems based on fingerprinting operate directly on the spectral representations of the signal to extract the fingerprint. This fingerprint is generally easy to extract and does not require significant computing resources.

Haistma et al. [57] (Haitsma and Kalker, 2002) developed an audio identification system for the company Philips. They use 33 non-overlapping logarithmic bands covering the range of 300Hz to 2kHz as their base feature. To improve the robustness of the system and to reduce the computational requirement, the change in the energy difference of adjacent bands on frame to frame basis is computed and stored as a single bit. This process of quantization gives a robust 32-bit feature vector per frame. Such a vector is calculated every 12 ms, giving about 86 frames per seconds. In this case the number of vectors to store for a music database of 10,000 songs of 5 minutes is about 260 million vectors. Different variations of this fingerprint have been developed, some by the authors themselves [58] (Haitsma and Kalker, 2003), in order to make the fingerprint more robust to deformations such as changing the speed (pitching). Improvements are however not very important.

For the comparison phase, the similarity measure used is the bit error rate, which is the number of erroneous bits divided by the total number of bits. The unknown fingerprint is considered identified if the bit error rate is less than a certain threshold. The authors show that for a value of 0.35, the probability of a false alarm is around 10^{-20}. The search method is based on indexing every reference frame in a look-up table. If the number of sub-bands used is N_b, then each frame will be a vector of $(N_b - 1)$ bits and the look-up table will have a 2^{N_b} entries. Each entry, called a key, points to all objects that have exactly this entry in the corresponding time. To reduce the identification time, candidate selection from the reference templates in the database is done with an assumption that at least one

feature vector among the block is the same as its original. This mechanism is faced with two difficulties. The first is the limitation of available memory, the look-up table is indeed too large to be loaded into memory. The second difficulty involves the distortions that the unknown signal will suffer. In fact, the assumption that at least one feature vector among the block is the same as its original, is not respected if same degradations affect the entire block. In order to test the robustness of this system, a hash block is extracted from four audio excerpts. All the excerpts are then subjected to different kinds of synthetic degradations. Experimental results show the robustness of the system against these degradations and the discrimination power of the hash blocs.

Wang et al. [133] (Wang et al., 2006) propose an audio search engine for the company SHAZAM. The algorithm uses a combinatorial hashed time-frequency constellation analysis of the signal. The fingerprint is based on the concept of landmarks. The landmark point is the spectrogram peak which has a higher energy content than all its neighbors in a region centered on itself. Candidate peaks are chosen according to a density criterion in order to assure that the time-frequency strip for the audio file has reasonably uniform coverage. Once the landmark points are identified, they are combined to increase the provided amount of information. Authors propose to create a key for each pair of landmarks, in fact, for two spectrogram peaks (f_1, t_1) and (f_2, t_2), the key will be the triplet $(f_1, f_2, t_2 - t_1)$. Each key presents an entry in the look-up table and each entry will contain the list of references (p, t) having this key, where p is the reference ID and t the time inside the reference.

In the searching process, triplet key of the unknown signal $(f_1, f_2, t_2 - t_1)$ is matched to select the possible candidates. For each candidate (p, t), the temporal offset histogram is computed where the offset is equal to $t_1 - t$. Once all candidate offset histograms are computed, the one with the maximum peak, which must be superior to a certain threshold, is considered as the best match to the unknown signal. The system is evaluated with 250 music samples of varying length and noise levels against a reference database of 10,000 tracks consisting of popular music. Audio excerpts of 15, 10, and 5 seconds in length are taken from the middle of each test track, each of which was taken from the test database. For each test excerpt, the relative power of the noise was normalized to the desired Signal-

to-Noise Ratio (SNR), then linearly added to the sample. Recognition rates are 50% for 15, 10, and 5 second samples at approximately -9, -6, and -3 dB SNR and more than 80% at 3, 6 and 9 db SNR. For the same analysis, except that the resulting music+ noise mixture is further subjected to GSM 6.10 compression, then reconverted to PCM audio, the 50% recognition rate level for 15, 10, and 5 second samples occurs at approximately -3, 0, and + 4 dB SNR. This method is intended to identify long audio objects. Indeed, short objects, like an advertisement or a radio jingle will not have enough landmark points, to ensure the reliability of the measure used in this case. In addition Fenet et al. [44] (Fenet et al., 2011) propose to use the Constant Q Transform (CQT) to improve the performances of the SHAZAM audio identification engine.

Pinquier et al. [104] (Pinquier and André-Obrecht) propose a method based on simple spectral coefficients. A total of 29 coefficients spanning the range of 100Hz to 8000Hz are used as feature vector. The final fingerprint consists of blocks of N such vectors where N depends on size of the training file. During the identification process, a block of the input feature vector is compared with the block stored in database using Euclidean distance. To achieve the goal of real-time processing, instead of performing comparison at each frame, fixed number of frames are skipped before next comparison. Test database is made up of six different corpora. The total duration is about 10 hours. The reference database is composed of 32 jingles with duration between 1 and 5 s. Among 132 jingles which had to be detected and identified, 130 are identified (98.5% of accuracy).

3.5.3.2 Computer Vision Techniques

There have been several experiments of using computer vision techniques for audio fingerprinting. The main idea is to treat the spectrogram of each audio clip as a 2-D image that transforms music identification into a corrupted sub-image retrieval problem.

In [12] (Baluja and Covell, 2008), the authors exploit the applicability of wavelet in image queries for large databases in fingerprinting applications by processing the audio spectrogram as a 2-D image. They generate a spectrogram of an audio with exactly the same parameter as described in [57] (Haitsma and Kalker, 2002). Then the audio spectrogram

is divided into smaller spectral images. Wavelet decomposition of each spectral image is carried out using Haar wavelet. Signs of top 200 wavelet magnitudes are retained in the final fingerprint. Then a hash table is used to find the best fingerprint segments and Hamming distance is computed between the candidate fingerprints and the query fingerprints. The system is evaluated with 1,000 independent probe snippets against a reference database of 10,000 songs, with an average song duration of 3.5 minutes. For this configuration, the recognition rate is 97.9%.

In [68] (Ke et al., 2005), the spectrogram of each music clip is treated as a 2-D image and transforms music identification into a corrupted sub-image retrieval problem. By employing pair-wise boosting on a large set of Viola-Jones features [132] (Viola and Jones, 2001), the system learns compact, discriminative, local descriptors. The system is evaluated on 220 songs captured with a very noisy recording setup against a reference database of 1,862 songs. The precision rate obtained with this configuration is 93% with the corresponding recall value of 80%.

Many other audio fingerprinting methods based on computer-vision techniques are proposed in the literature. For example in [138] (Zhu et al., 2010) an audio fingerprinting system is proposed to solve the problem of time scale modification and pitch shifting by extracting the Scale Invariant Feature Transform (SIFT) features from the spectrogram image.

3.5.3.3 Machine Learning Techniques

The last category of audio fingerprinting systems is based on machine learning techniques usually exploited for speech processing. These systems generally rely on vector quantization and HMM modeling.

In [31] (Cremer et al., 2001), the authors exploit low-level signal features standardized in MPEG-7 framework to develop a fingerprinting system. The system uses loudness, Spectral Flatness Measure and Spectral Crest Measure as the base feature. The features extracted from the training data are further processed with vector quantization method to obtain a set of code vectors by minimizing the Root Mean Square Error criterion. The

obtained set of code vectors is stored in a database forming a codebook which represents a particular class (audio item). The music identification task here is a N-class classification problem. For each of the music items in the database one class, i.e. the associated codebook, is generated.

To identify an unknown music item which is included in the reference database, a sequence of feature vectors is generated from the unknown item and these features are compared to the codebooks stored in the database. The class with minimal cumulative distance is assigned as the resultant class to the query input. A correct identification rate of 98% is achieved on a test set comprising 15,000 songs. The system runs about 80 times real-time on a Pentium III 500MHz class PC.

In [24] (Cano et al., 2002) a system based on HMM modeling is proposed. 32 models called AudioDNA are used to segment the audio signal into Audio Gens using the Viterbi algorithm. The final fingerprint consists of a sequence of letters (the Gens) and their temporal information (start time and duration). During the matching process, short subsequences of AudioDNA from an observed audio stream are continuously extracted and compared with the fingerprints in the references database. To reduce the computational processing time, string search algorithm called FASTA [98] (Pearson and Lipman, 1988) is proposed. The FASTA algorithm is initially deployed for bioinformatics. They report results with a reference database containing 50,000 music titles. In a preliminary experiment 12 hours of continuously broadcasted audio of different stations are captured to test the recognition performance of the system. All the 104 titles included in the reference database are detected.

3.5.3.4 Comparing System Performances

In the previous section, we presented some representative works of the state of the art in the field of audio identification based on fingerprinting. The main challenge of these systems is to create a robust fingerprint against different types of distortions and to propose a fast matching method that can satisfy real-time requirements regardless the size of the reference database. Table 4.1 illustrates the performance of the systems described above

Systems	Accuracy	Robustness	Granularity	Complexity	Scalability
Haitsma et al. [57]	+	NA	NA	+	-
Wang et al. [133]	-	-	-	+	+
Pinquier et al. [104]	+	NA	+	+	-
Baluja et al. [12]	+	+	NA	-	NA
KE et al. [68]	-	-	NA	-	NA
Cremer et al. [31]	+	NA	NA	-	-
Cano et al. [24]	+	+	NA	-	NA

Table 3.1: Performance of audio fingerprinting systems described in section refch02.sec.5.subsec.3, according to accuracy, robustness, granularity, complexity and scalability.

Systems	Reference database	Test database	Precision	Recall
Haitsma et al. [57]	4 excerpts	4 excerpts	100%	100%
Wang et al. [133]	10,000 songs	250	NA	80%
Pinquier et al. [104]	32 jingles	10h broadcast	100%	98,5%
Baluja et al. [12]	10,000 songs	1,000 songs	NA	97,9%
KE et al. [68]	1,862 songs	220 songs	93%	80%
Cremer et al. [31]	15,000 songs	15,000 excerpts	NA	98%
Cano et al. [24]	50,000 songs	12h broadcast	100%	100%

Table 3.2: Comparison of the performances of the systems described in 3.5.3, involving database and corpus sizes, precision and recall.

according to the criteria (accuracy, robustness, granularity, complexity and scalability) described in section 3.5.1.

Moreover, different experimental protocols are used in order to evaluate the audio identification systems based on fingerprinting. Table 4.2 summarizes the evaluation protocols used by the systems described in the previous section. The performance measure computed to evaluate audio identification systems based on fingerprinting are usually:

- Precision: The number of audio items correctly detected / Total number of detected audio items.
- Recall: The number of audio items correctly detected / The number audio items that should be detected.

Most of the audio fingerprinting systems described in tables 4.1 and 4.2 are only evaluated on a specific type of audio content (song or jingle) using private corpora which

makes the comparison between each other impossible. In chapter 6, we will present an audio fingerprinting system based on ALISP method to detect simultaneously, songs and advertisements in radio broadcast streams. The proposed system will be evaluated during the 2010 QUAERO evaluation campaign [108] (Ramona et al., 2012).

In this section, a system overview of audio identification based on fingerprinting was described. In fact audio fingerprinting system should meet certain criteria as granularity and accuracy. In addition the fingerprint must be robust to different degradations that the audio signal could suffer. We also described in this chapter the most representative techniques of the state of the art. These systems used different techniques to extract the fingerprint and proposed several matching process to search that fingerprint in the references database.

3.6 Conclusion

In this chapter, the state of the art of the unsupervised techniques for speech processing was reviewed. These techniques are generally requiring decreasing amount of human expertise and annotated resources, and increasing amount of unsupervised learning. Then, the adopted unsupervised technique used in our works was presented. This method is based on ALISP data-driven segmentation which consists of four steps: parameterization, temporal decomposition, vector quantization and HMM modeling. Finally, the ALISP-based speech processing systems are described. These systems are relative to very low bite rate speech coding, speaker verification and forgery and language identification.

At this point, ALISP method was only used for speech processing. In this thesis, a generic audio indexing system based on ALISP segmentation is proposed. The main purpose of this system is to retrieve and identify all the items present in a radio streams. These items are usually: music, commercial, jingle, speech and nonlinguistic vocalization (such as laughter, cough, applause,...). To this end, an audio indexing system based on data-driven ALISP technique is exploited for radio streams indexing and used for different audio indexing tasks, which are audio identification, audio motif discovery, speaker diarization and nonlinguistic vocalizations detection.

Chapter 4

Databases

4.1 Introduction

In this chapter, we present the audio databases exploited during this thesis. A radio broadcast corpus of radio data is provided by YACAST[1]. We had 26 days of annotated audio data that separated to train the ALISP models and to evaluate the ALISP-based audio indexing systems. In order to validate our proposal for speaker diarization we participated to the ETAPE'2011 evaluation campaign. Moreover, the proposed speaker verification system is evaluated during the MOBIO'2013 evaluation campaign. Finally, we use three publicly available corpora to evaluate our system for laughter detection.

4.2 Radio Broadcast Corpus

In the framework of the ANR-SurfOnHertz project we had at disposal the YACAST database. We had 26 days of annotated audio data from 13 French radio stations.

Three types of annotation are available, music, commercial and speaker turn, described below:

- Music annotation: it provides information about the songs present in the radio stream. This ground truth is given by an XML file for each radio station. The XML

[1] http://www.yacast.fr/fr/index.html

structure is given as follows:

```
<MusicTrack>
  <id>4134305</id>
  <idMedia>553</idMedia>
  <title>Belly dancer</title>
  <artist>Bob Sinclar, Kevin Lyttle</artist>
  <album>Born in 69</album>
  <genre>Dance</genre>
  <sousGenre>House</sousGenre>
  <startDate>2009-12-03 23:57:54</startDate>
  <endDate>2009-12-04 00:00:47</endDate>
</MusicTrack>
```

Music annotations are available for 10 radio stations which are: Virgin Radio, NRJ, RFM, RTL, RTL2, Cherie FM, FUN Radio, Europe1, RMC, and France Inter.

- Commercial's annotation: As for music, it provides information about the diffused commercial in the radio station. Its XML structure is given as follows:

```
<Advertisement>
  <id>5917143</id>
  <idMedia>1</idMedia>
  <name>MUSE CONCERT</name>
  <brand>MUSE CONCERT</brand>
  <advertiser>MUSE CONCERT</advertiser>
  <startDate>2009-12-01 00:45:08</startDate>
  <endDate>2009-12-01 00:45:51</endDate>
</Advertisement>
```

Commercial's annotations are available for the 10 radios given above plus the "France Info" radio.

- Speaker turn annotation: It gives the start and end time of each utterance in the radio streams with the identity of the corresponding speaker. Its XML structure is given as follows:

```
<TalkPassage>
  <idMedia>175712</idMedia>
  <mediaName>France Culture Temps de parole</mediaName>
  <idSpeaker>13355</idSpeaker>
  <speakerName>SARKOZY NICOLAS</speakerName>
  <startDate>2010-06-27 05:00:31:404</startDate>
  <endDate>2010-06-27 05:00:51:438</endDate>
</TalkPassage>
```

This database contains 2,172 different commercials that are broadcasted between 2 and 117 times. The mean duration of these commercials is 24 seconds and their total number in the 26 days of recordings is 14,953. Moreover, 8,694 different songs are present in this database. The mean duration of these songs is 229 seconds (3 minutes and 49 seconds) and their total number in these recoding is 56,902. Regarding the speaker turn annotations, this database contains 283 annotated speakers with a total duration of 42h46min.

The ALISP HMM models are trained on a part of this corpus. In fact, the ALISP models are trained on one day of audio stream from 12 radios (leading to 288 h). It is important to insist that the training database remains the same for all the proposed audio indexing systems. This training database is referred in this chapter and in the followings as the ALISP training database, for the HMM ALISP models.

Three audio indexing systems are evaluated using this corpus: audio identification, audio motif discovery and speaker diarization that is also evaluated during the ETAPE'2011 evaluation campaign.

genre	train	dev	test	sources
TV news	7h30	1h35	1h35	BFM Story, Top Questions (LCP)
TV debates	10h30	2h40	2h40	Pile et Face, Ca vous regarde Entre les lignes (LCP)
TV amusements	-	1h05	1h05	La place du village (TV8)
Radio shows	7h50	3h00	3h00	Un temps de Pauchon, Service Public Le masque et la plume, Comme on nous parle Le fou du roi
Total	25h30	8h20	8h20	42h10

Table 4.1: ETAPE dataset composition [55].

4.3 ETAPE Corpus

ETAPE is an evaluation campaign for automatic speech processing [55] (Gravier el al., 2012). As illustrated in table 4.1, the ETAPE data are divided into three subsets; train, development and evaluation data. Note that the number of hours are reported in terms of recordings, not speech. As reported, in the ETAPE TV data, about 77% of the recordings contain speech.

4.4 MOBIO Corpus

In this thesis we are interested by measuring the speech time of politicians in radio streams. This task involves two fields of speaker-based processing: speaker identification and speaker diarization. As pointed out before, the speaker diarization system is evaluated on the YACAST database and during the ETAPE'2011 evaluation campaign. Following the same spirit, the speaker identification system is evaluated during the MOBIO'2013 evaluation campaign.

The MOBIO database is a bimodal (face/ speaker) database recorded from 152 people. The database has a female-male ratio of nearly 1:2 (100 males and 52 females) and was collected from August 2008 to July 2010 in six different sites from five different countries.

In total 12 sessions were captured for each individual. The database was recorded using two types of mobile devices: mobile phones (NOKIA N93i) and laptop computers

(standard 2008 MacBook). In this evaluation we will only use the mobile phone data with a sampling rate of 16kHz.

The MOBIO database is a challenging database since the data is acquired on Mobile devices possibly with real noise, and the speech segments can be very short (less than 2sec). More technical details about the MOBIO database can be found in [85] (McCool et al., 2012). Based on the gender of the clients, two different evaluation protocols for male and female were generated. In order to have an unbiased evaluation, the clients are split up into three different sets: training, development and evaluation sets:

- Training set: The data of this set is used to learn the background parameters of the algorithm (UBM, subspaces, etc.). They can also be used for score normalization (cohort, etc.). It is worth noting that participants can use external data in their background training, however they should explicitly precise it in their system description.
- Development set: The data of this set is used to tune meta-parameters of the algorithm (e.g. number of Gaussians, dimension of the subspaces, etc.). For the enrollment of a client model, 5 audio files of the client are provided, and it is forbidden to use the information of other clients of the development set. The remaining audio files of the clients serve as probe files, and likelihood scores have to be computed between all probe files and all client models. In systems that require score calibration these scores can be used to train the calibration parameters.
- Evaluation set: The data of this set is used for computing the final evaluation performance. It has a structure similar to the development set. The only difference is that the file names are anonymized in order to prevent participants to optimize their system on the evaluation set.

Table 4.2 details each of the sets described above. It specifies the number of files and the number of targets.

	Background		Development				Evaluation			
			Enrollment		Test		Enrollment		Test	
	Spks	Files	Targets	Files	Spks	Files	Targets	Files	Spks	Files
MALE	37	7104	24	120	24	2520	38	190	38	3990
FEMALE	13	2496	18	90	18	1890	20	100	20	2100
TOTAL	50	9600	42	210	42	4410	58	290	58	6090

Table 4.2: Number of targets and audio files of the training set, the number of targets and enrollment audio files, and the number of test segments for the development and the evaluation set, in the MOBIO audio data.

4.5 Laughter Detection Corpus

In order to evaluate the proposed laughter detection system, three publicly available sources are used. These databases are:

- SEMAINE-DB [86] (McLeown et al., 2012): A large audiovisual database recorded from 150 participants. The youngest participant is 22, the oldest 60, and the average age is 32.8 years old. Thirty-eight percent are male. A manual transcriptions of laughter are available. The total duration of the database is 15h5min.
- AVLaughterCycle [130] (Urbain et al., 2010): An audiovisual laughter database recorded from 24 subjects. Annotations of the recordings, focusing on laughter description with more than 1,000 spontaneous laughs and 27 acted laughs. The laughter duration ranges from 250ms to 82s.
- Mahnob laughter databases [100] (Petridis et al., 2013): Au audiovisual laughter database recorded from 22 participants. The total duration of the database is 3h49min. It contains 563 laughter sequences, 849 speech utterances, 51 posed laughs, 67 speech laughs episodes and 167 other vocalizations.

4.6 Conclusion

In this chapter, all the corpora used to develop and evaluate our audio indexing system are described. The radio broadcast corpus is essentially used to train the ALISP

HMM models and evaluate the ALISP-based audio identification, audio motif discovery and speaker diarization systems. While ETAPE and MOBIO databases are related to evaluation campaigns that we have participated. On the other hand the laughter detection database are exploited to evaluate the proposed nonlinguistic vocalization detection system.

In the next chapter, the main contributions of our works are presented. These contribution are related to the ALISP segmenter, approximate matching process of ALISP units and the generic audio indexing system.

Chapter 5

Contributions to Data-driven Audio Indexing

5.1 Introduction

This chapter presents our main contributions in this Ph.D, which can be divided into three parts:

1. Improving the ALISP tools by introducing a simple method to find stable segments within the audio data. This technique, referred as spectral stability segmentation, is replacing the temporal decomposition used before for speech processing. The main advantage of this method is its computation requirements which are very low comparing to temporal decomposition.

2. Proposing an efficient technique to retrieve relevant information from ALISP sequences using BLAST algorithm [3] (Altschul et al., 1990) and Levenshtein distance [76] (Levenshtein, 1966), with the goal to speed up the retrieval process without affecting the accuracy of the audio indexing process.

3. Proposing a generic audio indexing system, based on data-driven ALISP sequencing, for radio streams indexing. This system is applied for different fields of audio indexing to cover the majority of audio items that could be present in a radio stream:

- audio identification: detection of occurrences of a specific audio content (music, advertisement, jingle) in a radio stream;
- audio motif discovery: detection of repeating objects in audio streams. (music, advertisement, and jingle);
- speaker diarization: segmentation of an input audio stream into homogenous regions according to speaker's identities in order to answer the question "Who spoke when?";
- nonlinguistic vocalization detection: detection of nonlinguistic sounds such as laughter, sighs, cough, or hesitation;

As pointed out before, ALISP tools have already been used for very low bit-rate speech coding, speaker and language recognition, and voice forgery.

The objective through this thesis is to exploit high-level information provided by data-driven units in order to build an unified data-driven platform for audio indexing, retrieval and recognition. To this end, the ALISP method is used as a data-driven segmentation tool. ALISP method consists in segmenting the audio data in data-driven segments. The particularity of ALISP tools is that no textual transcriptions are needed during the learning step, and only raw audio data is sufficient. In such a way any input audio data is transformed into a sequence of arbitrary symbols. These symbols can be used for indexing purposes.

This chapter is divided into three parts according to our contributions. The first section deals with the improvements made on the ALISP tools. In the second section a new technique to retrieve relevant information from ALISP sequences is described. Finally, a generic audio indexing system, based on data-driven ALISP sequencing, for radio streams indexing is introduced.

5.2 Improving the ALISP Segmenter

ALISP tools are the basis for the data-driven segmenter we are using in this thesis. One part of our work is related to adapt and improve these tools with regard to the task and the database we are using for audio indexing. The improvements that we have made concern

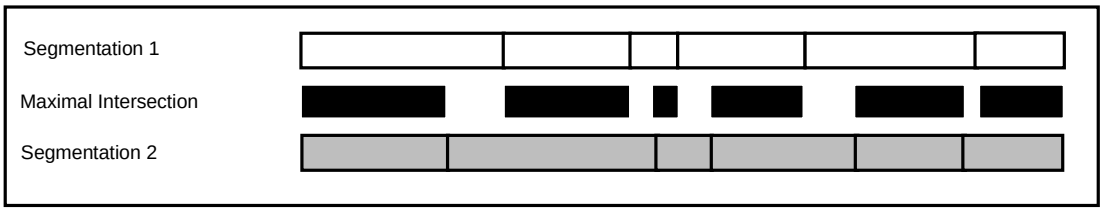


Figure 5.1: Maximal intersection between two segmentations.

the initial segmentation made by the temporal decomposition. As mentioned before, the temporal decomposition is used to obtain an initial segmentation of the audio data into quasi-stationary segments. These segments are clustered using vector quantization. Then, the boundaries together with labels are used as initial transcription for Hidden Markov Modeling.

In this section, other segmentation methods (for the HMM models) are explored in order to study the influence of the initial segmentation on the ALISP training process. These methods are:

- Uniform segmentation;
- Spectral stability segmentation;
- Phonetic segmentation.

A set of ALISP HMM models is trained for each initial segmentation technique using the training database (288 hours) described in the previous chapter, with 65 units (except for the phonetic segmentation).

In order to compare the proposed segmentation techniques with the temporal decomposition, the maximal intersection segmentation measure is computed. This measure is introduced in [66] (Joley et al., 2007) in order to extract the maximal intersection between two segmentation as shown in figure 5.1.

This comparison is performed at two levels:

1. The initial segmentation provided by each of the proposed segmentation technique is compared to the one provided by the temporal decomposition.

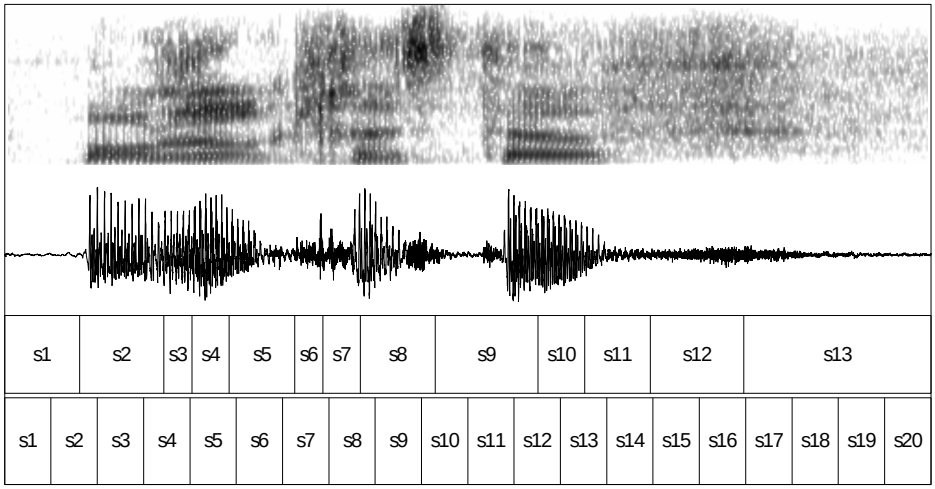


Figure 5.2: Spectrogram of an audio excerpt with two segmentations obtained by temporal decomposition (below) and the uniform segmentation (above).

2. The segmentation given by the ALISP HMM model using the proposed segmentation technique is compared to the one given by the ALISP HMM model using the temporal decomposition. The set of ALISP HMM models is acquired as shown in figure 3.2.

5.2.1 Uniform Segmentation

Generally, the uniform segmentation is the most direct approach to segment audio data. It consists on segmenting the audio data into an equal size frames, for example MFCCs are calculated for each 20ms frame. This process is similar to performing the vector quantization directly on audio frames. More precisely, after the parameterization step, the audio signal is divided into an equal size segments. Then a centroid frame (central frame) is taken as the representative frame of each segment. After that, the vector quantization and Hidden Markov Modeling are performed, as described in section 3.3, only the centroid frames are taken to build the dictionary. In this work the size of segments is equal to 50ms (5 frames).

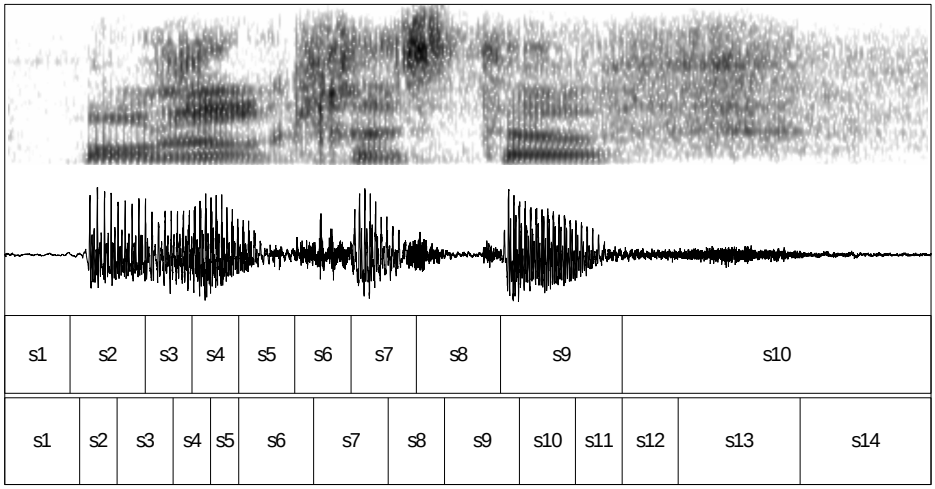


Figure 5.3: Spectrogram of an audio excerpt with two segmentations obtained by the ALISP HMM models after re-estimation using the temporal decomposition (below) and the uniform segmentation (above).

Figure 5.2 shows the spectrogram of an audio excerpt with the initial segmentations obtained by the temporal decomposition and the uniform segmentation. The maximal intersection between the initial segmentations of both methods is equal to 24%. This result was predictable, since the temporal decomposition aims to find quasi-stationary segments while in the uniform segmentation, the characteristics of the audio signal are not considered and the obtained segmentation relies only on the size of the segment.

In addition, ALISP segmentation obtained by the HMM models using the uniform segmentation is computed and compared with the one obtained by the HMM models using the temporal decomposition. Figure 5.3 shows an example of these segmentations. The maximal intersection between the segmentations provided by both HMM models is equal to 52%. This result shows that the initial segmentation has a significant effect on the final ALISP HMM models. On the other hand, the HMM modeling process has increased the maximal intersection from 24% to 52%, leading to an absolute improvement of 28%.

5.2.2 Spectral Stability Segmentation

The goal of this method is to find the stable regions of the audio signal. These regions represent the spectrally stable segments of the audio data. This process is performed using the spectral stability curve obtained by computing the Euclidian distance between two successive feature vectors as follows:

$$d = \sqrt{\sum_{i=1}^{n} (C_{i,j} - C_{i,j+1})^2} \tag{5.1}$$

where $C_{i,j}$ and $C_{i,j+1}$ are two successive feature vectors and n is their size. The local maxima of this curve represent the segment boundaries while the minima represent the "stable" frames of the audio signal.

Figure 5.4 shows the spectrogram of an audio excerpt with the initial segmentations obtained by the temporal decomposition and the spectral stability segmentation. The maximal intersection between segmentations provided by the temporal decomposition and the one obtained by the spectral stability segmentation of the ALISP training database is 78%. This result shows that the temporal decomposition method could be replaced by the spectral stability segmentation which is much easier to compute.

Figure 5.5 shows the spectrogram of an audio excerpt with the segmentations provided by the ALISP HMM models using the temporal decomposition and the spectral stability segmentation. The second comparison between HMM models using both techniques gives a 89% of maximal intersection, which leads to an absolute improvement of 11%. This result confirms our previous assumption that the spectral stability segmentation is an appropriate method to provide the initial segmentation.

5.2.3 Phonetic Segmentation

A phonetic segmentation method consists of using a HMM phonetic model to initially segment the audio data. This method is used to find out wether a phonetic model could be used for audio indexing purposes such as audio identification or audio motif discovery.

The HMM phonetic models are trained using ESTER database (French radio broad-

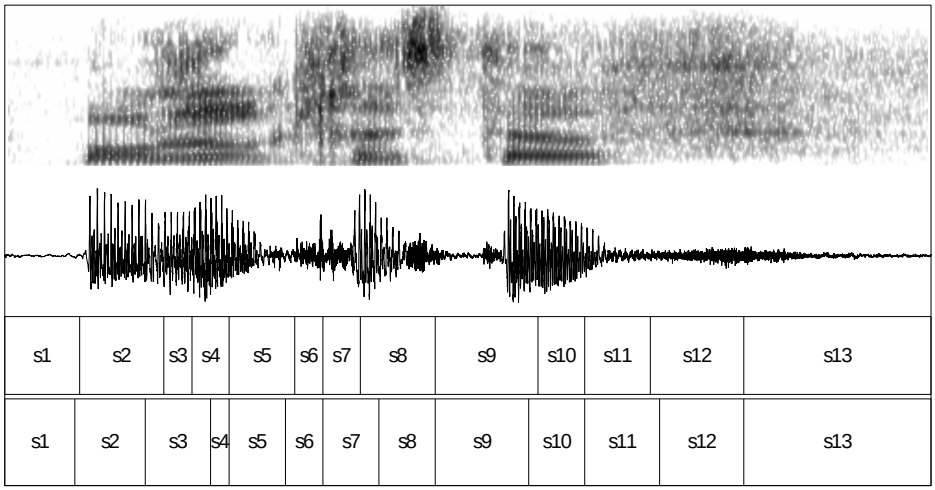


Figure 5.4: Spectrogram of an audio excerpt with two initial segmentations obtained by temporal decomposition (below) and the spectral stability segmentation (above).

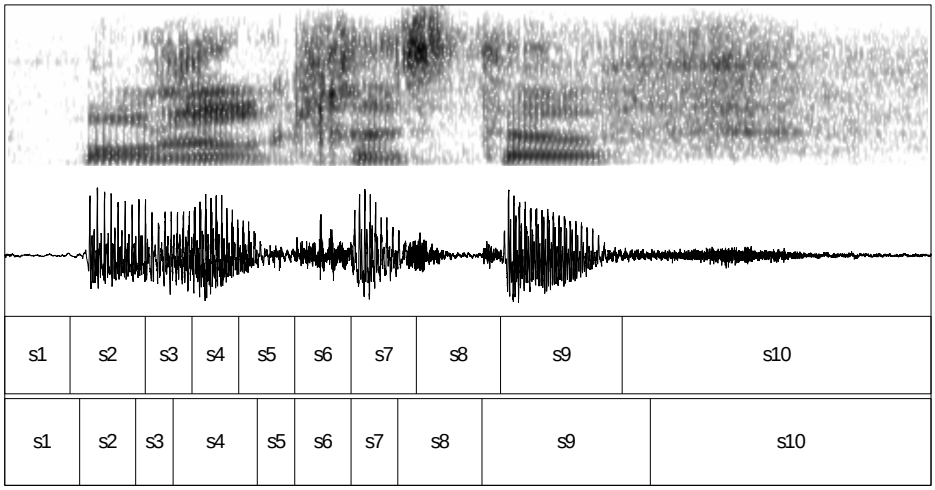


Figure 5.5: Spectrogram of an audio excerpt with two segmentations obtained by the ALISP HMM models using the temporal decomposition (below) and the spectral stability segmentation (above).

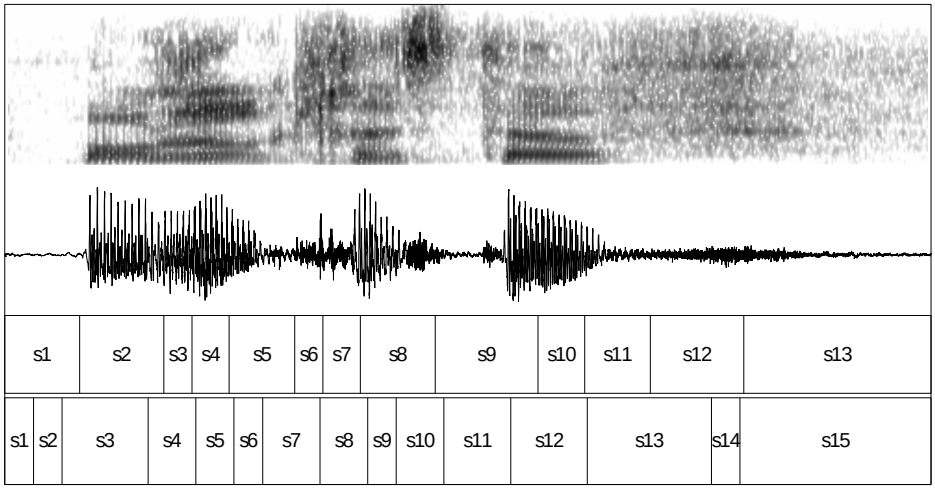


Figure 5.6: Spectrogram of an audio excerpt with two initial segmentations provided by temporal decomposition (below) and the phonetic segmentation (above).

cast database) [49] (Galliano et al., 2009). As for ALISP units, each phone (41 phones) is modeled by a left-right HMM having three emitting states with no skips. The phonetic segmentation is replacing the temporal decomposition and the vector quantization. In fact, the phonetic segmentation is used as initial transcription for Hidden Markov Modeling step of the ALISP modeling.

Figure 5.6 shows the spectrogram of an audio excerpt with the initial segmentations obtained by the temporal decomposition and the phonetic segmentation. The maximal intersection between the initial segmentations provided by both techniques is equal to 21%. This result is predictable given that the phonetic models are trained only on speech. In fact, using phone models on audio items other than speech (such as music and advertisement) could lead to a random segmentation.

In the next experience, the segmentations provided by the ALISP HMM models using both techniques are compared. Figure 5.7 shows the spectrogram of an audio excerpt with the segmentations provided by the ALISP HMM models using the temporal decomposition

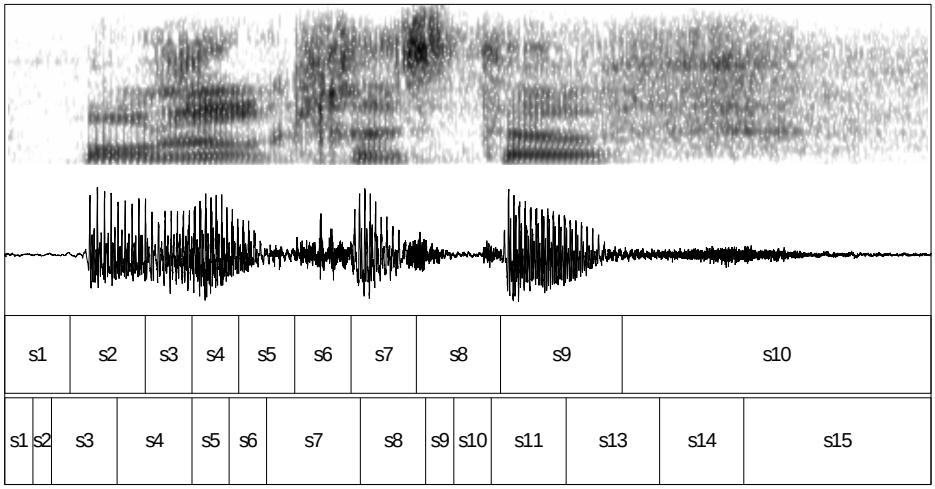


Figure 5.7: Spectrogram of an audio excerpt with two segmentations provided by the ALISP HMM models using the temporal decomposition (below) and the phonetic segmentation (above).

and the phonetic segmentation. The maximal intersection between both segmentations is equal to 32%, which gives an absolute improvement of 9%. This result shows that the use of a phonetic segmentation as an initial segmentation leads to a different ALISP HMM models.

5.2.4 Comparing Segmentation Techniques

In this part, a comparison between the different segmentation techniques is presented. Table 5.1 shows the maximal intersection between each of the proposed segmentation method and the temporal decomposition for the initial segmentation and HMM segmentation, using the 288-hours radio broadcast database to train the final ALISP HMM models.

This table shows that the spectral stability technique provides the nearest segmentation to the one provided by the temporal decomposition. On the other hand, phonetic and uniform segmentations are not appropriate to obtain a segmentation of the audio data into

Method	Initial segmentation	HMM segmentation
Uniform segmentation	24	52
Spectral stability segmentation	78	89
Phonetic segmentation	21	32

Table 5.1: Maximal intersection between segmentations provided from each of the proposed methods and the temporal decomposition for the initial segmentation and the HMM segmentation.

quasi-stationary segments.

In this section, the temporal decomposition was compared to three different segmentation techniques in terms of initial segmentation and HMM modeling. In the next chapter, the influence of the four segmentation methods on the performances of the proposed audio indexing system will be carried out.

5.3 Approximate Matching Process of ALISP Sequences

As mentioned before, the proposed audio indexing system is composed of three modules: automatic acquisition and modeling of ALISP units, segmentation module and comparison module. In the previous section, we presented our contributions related to the first and second modules. In this section, a new technique for approximate matching of ALISP sequences is proposed. This technique is used to compare relevant information from ALISP transcriptions using BLAST algorithm [3], (Altschul et al., 1990) and Levenshtein distance [76] (Levenshtein, 1966).

5.3.1 ALISP Sequencing

ALISP unit recognition involves the transformation of audio data into a sequence of ALISP units. The most likely ALISP sequence given a sequence of feature $Y = y_1 \ldots y_T$ is found by searching all possible state sequences arising from all possible ALISP units sequences for the sequence that was most likely to have generated the observed data Y. An efficient way to solve this problem is to use Viterbi algorithm [137] (Young et al., 1989).

In the previous section, we show that the temporal decomposition could be replaced

by the spectral stability segmentation. Therefore, a new scheme to acquire and model ALISP units is presented in figure 5.8.

5.3.2 Similarity Measure and Searching Method

An important part of the proposed audio indexing system is the matching process. As the main requirement of the proposed audio indexing system is robustness against several types of signal distortions, the actual ALISP unit sequences extracted from an observed signal will not be fully identical to the reference database. Two techniques are developed to perform the approximate matching process of ALISP sequences. The first one is related to the baseline method where a full search (or brute search) is applied, while the second technique is inspired form the Basic Local Alignment Search Tool (BLAST) [3] (Altschul et al., 1990), widely used in bioinformatics.

5.3.2.1 Full Search

The full search module compares the ALISP sequences extracted from observed audio signal against reference ALISP transcriptions stored in the reference database. First, the transcriptions of each reference advertisement (the ones that we are going to look for in the newly incoming audio stream) into a sequence of reference ALISP symbols has to be done. Then the test audio stream is transformed into a sequence of ALISP symbols. Once the ALISP transcriptions of reference and test data are done, we can proceed to the matching step.

The similarity measure used to compare ALISP transcriptions is the Levenshtein distance [76] (Levenshtein, 1966). The Levenshtein distance is a special case of an edit distance. The edit distance between two strings of characters is the number of operations required to transform one of them into the other. When edit operations are limited to insertion, deletion and substitution this distance is called Levenshtein distance. At this stage the matching component used in our system is very elementary. In each step we move on by one ALISP unit in the test stream and Levenshtein distance is computed between reference advertisement transcription and the transcription of the selected excerpt from the

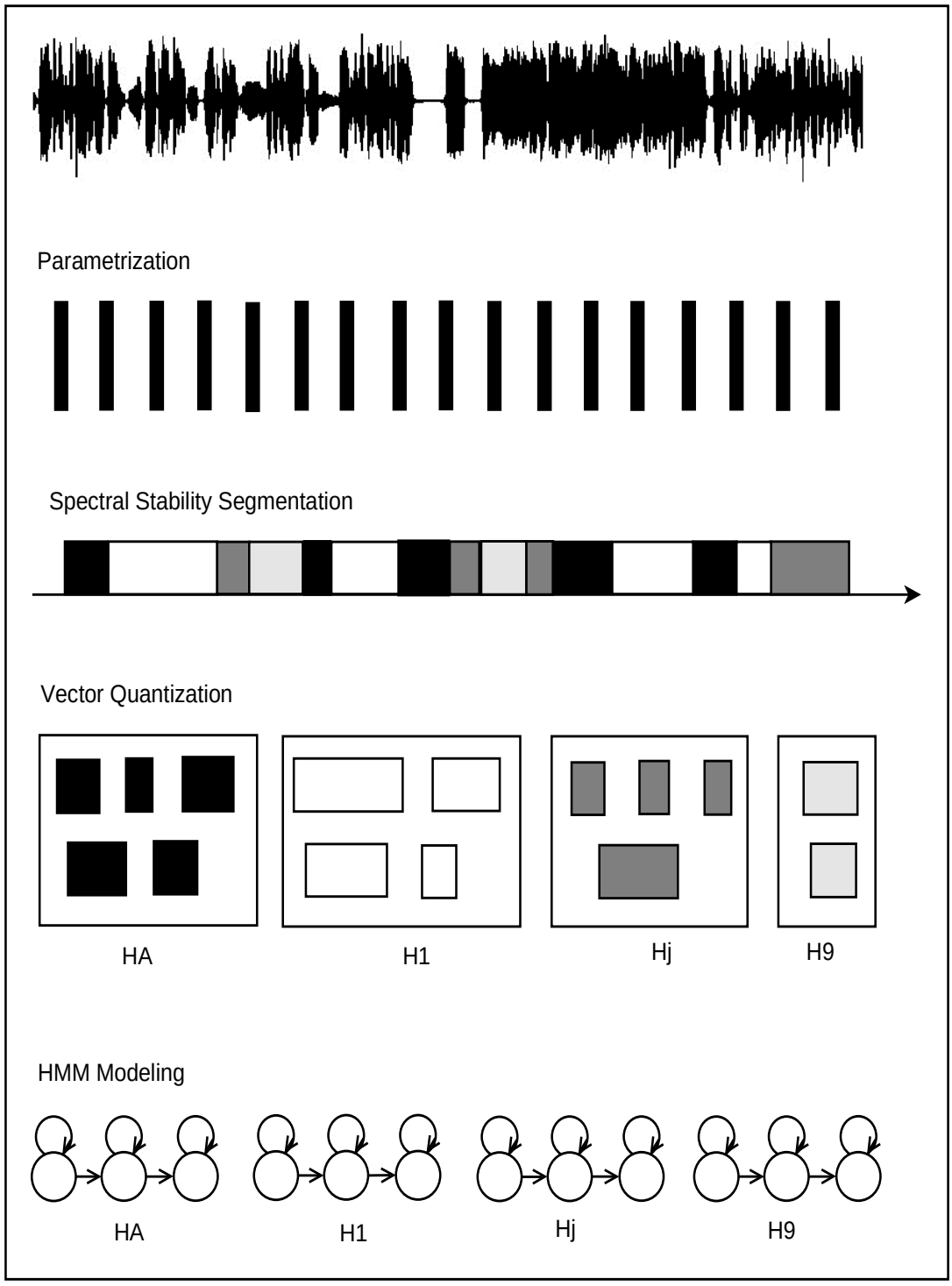


Figure 5.8: Illustration of the different steps of the ALISP units acquisition and their HMM modeling.

audio stream. At the point when the Levenshtein distance is below a predefined threshold it means that we have an overlap with the reference. Then we continue the Levenshtein distance comparison by stepping on by one ALISP symbol until the Levenshtein distance increases relatively to its value in the previous step. This point indicates the optimal match, where the entire reference has been detected.

In order to speed up the search stage, an alternative approximate matching method is developed. Approximate string matching algorithms are a traditional area of study in computer science. With the huge increase of nucleotide and protein sequence data produced by various genome projects, fast string matching algorithms are developed. Our approximate string matching algorithm is based on the BLAST technique [3] (Altschul et al., 1990), widely used in bioinformatics.

5.3.2.2 BLAST Algorithm

The BLAST [3] (Altschul et al., 1990) algorithm can be summarized as follows. It is an algorithm for comparing primary biological sequence information, such as amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. Note that BLAST considers that the library is formed by one long string sequence.

Let q be the querying string sequence and D the database. From the string q a substring w is considered. The first step in the algorithm is to build a lookup table (LUT) for all w-length words in D and to let the entries in that LUT point to the position where w-length word occurs. In the second step, for each w-length substring in q, a list of seeds is generated using the LUT. This list contains all w-length seeds with a similarity score with the relative substring greater than a certain threshold T. The final step of the algorithm consists of extending each candidate seed on either side to find the optimal alignment with the querying string sequence q. A candidate is considered as the optimal alignment if its similarity score with the query q is greater than a certain threshold S.

In our case the query sequence is a long sequence of ALISP symbols where we are

looking for occurrences of reference advertisements and songs. In order to deal with this, the BLAST algorithm was adapted as follows.

5.3.2.3 Approximate Matching Process of ALISP Sequences

The approximate matching process depicted in figure 5.9 is proposed. First, a LUT is created by all possible ALISP sequences of w units but with an offset of k units that occur in the ALISP transcriptions of the reference database. This database contains all the audio item references that we can identify, such as songs, advertisements, speech segments and audio motifs.

Each entry in the LUT points to the audio item reference and the position in that item where the respective ALISP unit sequence occurs. Since an ALISP unit sequence can occur at multiple positions in multiple audio items the pointers are stored in a linked list. Therefore one ALISP sequence can generate multiple pointers and positions.

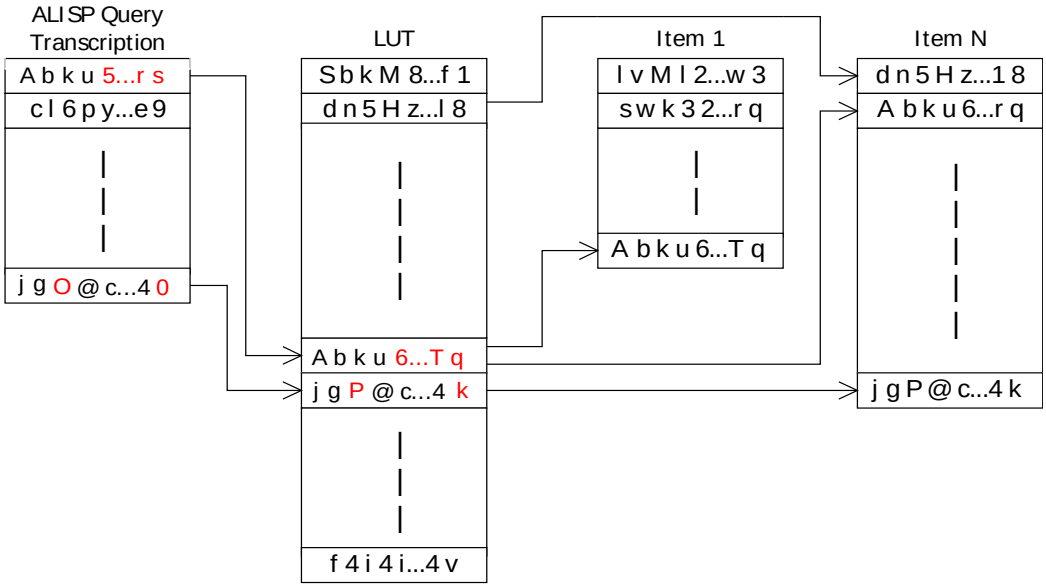


Figure 5.9: Approximate matching process of an ALISP query transcription using a lookup table (LUT) and a reference database containing N items.

Then, the ALISP transcription is computed from the query audio stream, and for each subsequence of w units with an offset of k units of that query a set of candidate subsequences is found using the LUT. In contrast to the original BLAST algorithm, we are not looking to the exact occurrence of the subsequence of w units in the LUT. We are rather searching the subsequences that have some differences with w units subsequence. This operation is motivated by the fact that the assumption that at least one subsequence among the query audio stream is the same as its original, is not respected if the same degradations affect the entire block.

From this set of subsequences, a list of candidate references and the position where the candidate subsequences occur in that reference is generated for each subsequence of the query data.

Since our reference database is formed by each ALISP transcription of the reference audio item (not one string sequence as in BLAST), the final step of the matching process is different from the BLAST one. It consists of a simple comparison between the ALISP transcription of the query audio stream and the corresponding candidates references with the Levenshtein distance [76] (Levenshtein, 1966). The candidate audio item selected as the best match of the unknown audio stream is the reference having the lowest Levenshtein distance among all candidates and providing a Levenshtein distance below a certain threshold.

The approximate matching of ALISP sequences is used to identify an audio items such as commercial, music, audio motif or speech segment. This module is very crucial since it defines whether the proposed audio indexing system follows the real-time requirements or not.

5.4 Generic ALISP-based Audio Indexing System

The next main contribution of this thesis is the exploitation of the ALISP approach and the proposed approximate matching process as an unified method for audio indexing and recognition. There are many existing applications for audio processing, such as song classification, advertisement (commercial) detection, speaker diarization and identification, with various systems being developed to automatically analyze and summarize audio content

for indexing and retrieval purposes. With these systems audio data are treated differently depending on the applications. For example, song identification systems are generally based on audio fingerprinting such as SHAZAM and Philips systems. While speaker diarization and identification systems are using cepstral features and machine learning techniques such as Gaussian Mixture Models and Hidden Markov Models. The diversity of audio indexing techniques makes unsuitable the simultaneous treatment of audio streams where different types of audio (music, commercials, jingles, speech, laughter, etc.) coexist. Hence the need for a generic system portable across domains.

One of the contributions of this thesis is the exploitation of ALISP approach as a generic method for audio indexing and recognition. As pointed before, ALISP is a data-driven technique that transforms any input audio data into a sequence of arbitrary symbols. These symbols can be used for indexing purposes.

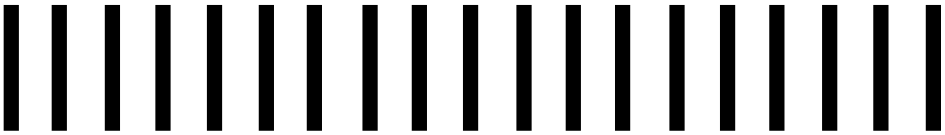
5.4.1 System Overview

The main purpose of our works is to retrieve and identify the majority of audio items present in a radio streams. These items are usually: music, commercial, jingle, speech and nonlinguistic vocalization (laughter, cough, sight,...). To this end, a generic audio indexing system based on data-driven ALISP technique is developed and exploited for radio streams indexing and applied for different fields to cover the different items that could be present in a radio stream.

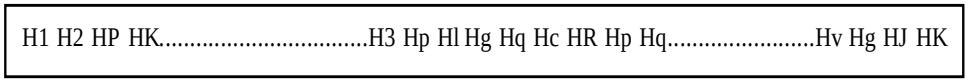
Figure 5.10 shows the proposed ALISP-based system overview. As shown in this figure, the proposed audio indexing system is composed of four sub-systems based on the same ALISP sequencing method:

- Audio identification: detection of occurrences of a specific audio content (music, advertisement, jingle) in a radio stream.
- Audio motif discovery: detection of repeating objects in audio streams.
- Speaker diarization: segmentation of an input audio stream into homogenous regions according to speaker's identities in order to answer the question "Who spoke

Parametrization



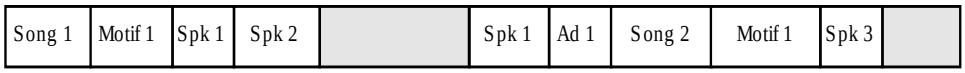
ALISP Sequencing



Audio Identification

Audio Motif Discovery

Speaker Diarization



Nonlinguistic Vocalizations Detection

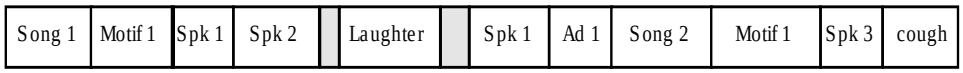


Figure 5.10: ALISP-based audio indexing system.

when?"

- Nonlinguistic vocalization detection: detection of nonlinguistic sounds such as laughter, sighs, cough, or hesitation.

Speaker diarization should only be applied to speech data. Therefore, performing the audio identification and audio motif discovery at the beginning is important to remove music and advertisement data. The ALISP-based audio indexing system is composed of 4 sub-systems. Although these systems are different, they are using a common architecture based on ALISP method. This architecture is composed of two modules: ALISP sequencing and approximate matching of ALISP sequences.

5.4.2 Audio Indexing: Fields of Interest

As pointed out before, the audio indexing system, based on ALISP sequencing and approximate matching technique, is composed of four sub-systems. These systems are chosen to cover the majority of audio items that could be present is a radio broadcast stream.

5.4.2.1 Audio Identification

As shown in figure 5.11, the proposed system uses automatically acquired units provided by ALISP tools to search for advertisements and music pieces in radio broadcast streams. In this sense ALISP transcriptions of advertisements and songs, present in the reference database, are computed using HMM models provided by ALISP tools and Viterbi algorithm and compared to transcriptions of the radio stream using the BLAST algorithm [3] (Altschul et al., 1990).

5.4.2.2 Audio Motif Discovery

Radio streams often contain redundant parts. Commercials on radio or television stations, songs on music channels and jingles broadcasted before a specific radio or TV show, are some of the repeating objects in multimedia streams. The ALISP-based audio

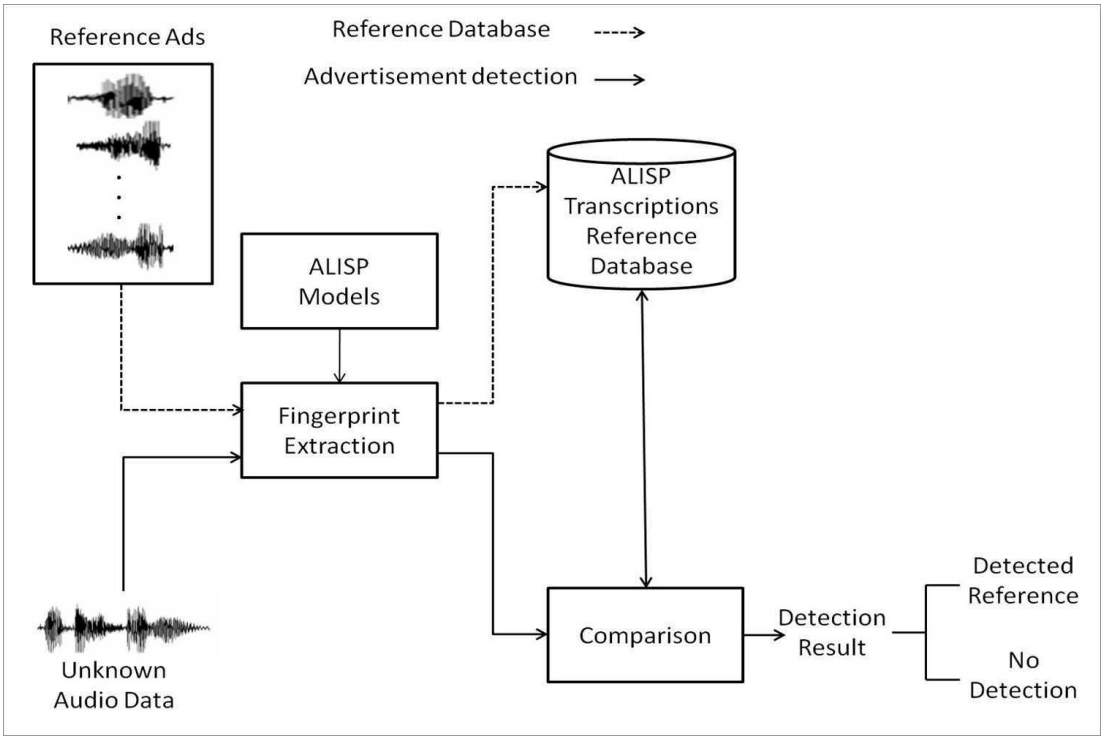


Figure 5.11: Audio identification system based on ALISP fingerprinting.

indexing system is used to detect repeating objects in audio streams. In order to resolve this problem, the ARGOS segmentation framework proposed in [61] (Herley, 2006) is used. This framework is combined with ALISP sequencing technique to build an audio motif detection system. The BLAST algorithm is applied to speed up the approximate string matching to find the repeating items in the audio streams.

As was previously mentioned, the data-driven ALISP technique converts the raw audio data into a sequence of symbols. These symbols represent the fingerprint used to detect the repeating items in audio streams. Thus, the problem of audio motif detection is transformed into a string matching problem.

5.4.2.3 Speaker Diarization

In our work, we are interested in speaker diarization for TV and radio shows which include various acoustic sources such as studio/telephone speech, music, or speech over music. Usually these shows keep the same structure with same presenters and jingles. This redundancy is used in order to improve the performance of the speaker diarization system.

The main idea of our system is to compare the show to be segmented with the same show broadcasted before in order to find the common audio segments. This operation is performed via audio fingerprinting which involves the extraction of a fingerprint for each audio document stored in a reference database. An unlabeled audio excerpt is identified by comparing its fingerprint with those of the reference database.

The reference database is built from audio segments provided by annotated databases. These segments represent speech sentences, silence, noise, jingles, music and advertisements. Then ALISP transcriptions of reference segments are computed using HMMs provided by the ALISP tools and compared to the transcriptions of the TV and radio shows stream using the BLAST algorithm.

5.4.2.4 Nonlinguistic Vocalizations Detection

Despite the best efforts made over past two decades in speech recognition systems, detection of nonlinguistic vocalizations such as laughter, sighs, breathing, or hesitation

sounds is still a challenging task. Such vocalizations are most frequent vocalizations in our daily conversational speech.

Laughter is one of the complex nonlinguistic vocalizations that conveys a wide range of messages with different meanings. Most of previous studies on automatic laughter detection from audio are based on frame level acoustic features as parameters to train machine learning techniques, such as Gaussian Mixture Models and Support Vector Machines.

A generic methodology to detect nonlinguistic vocalizations using ALISP method is proposed. Using Maximum Likelihood Linear Regression and Maximum A Posterior techniques, the proposed method adapts ALISP models, which then facilitate detection of local regions of nonlinguistic vocalizations with the standard Viterbi decoding algorithm. Moreover, a simple majority voting scheme, using a sliding window on ALISP sequences, can be helpful in eliminating outliers from the Viterbi-predicted sequence automatically. The evaluation of the proposed system is performed on laughter detection.

5.5 Conclusion

In this chapter, the main contributions of our works are presented. First we described how the ALISP segmenter is improved using other techniques to provide initial segmentation to initialize the ALISP HMM models. These techniques were described and compared with the temporal decomposition. The comparison of the segmentation attained by the spectral stability method and the temporal decomposition showed that there is a great correlation between both segmentations.

In the second part of this chapter, a new technique to extract relevant information from ALISP sequences is presented. The proposed approximate matching process is inspired form BLAST technique, widely exploited in bioinformatics, and the Levenshtein distance, used to compare ALISP transcriptions.

The third contribution is related to the generic ALISP-based audio indexing system. This system is composed of four sub-systems which are: audio identification, audio motif discovery, speaker diarization and nonlinguistic vocalizations detection. All these systems are using a common architecture based on ALISP method.

In the next chapter we treat the first task of audio indexing which is audio identification. Audio identification consists of detecting and locating occurrences of a specific audio content (music, advertisement, jingle,..) in audio streams or audio databases. The proposed ALISP-based system will be evaluated on the radio broadcast corpus and during the QUAERO project evaluation campaign.

Chapter 6

Audio Identification

6.1 Introduction

In this chapter we present the ALISP-based audio indexing system applied to the audio identification task. Audio identification consists of detecting and locating occurrences of a specific audio content (music, advertisement, jingle,..) in audio streams or audio databases. There are many potential applications of audio identification, most have recently emerged. We can distinguish three categories according the application to which they are intended [15] (Betser, 2008):

- Seeking information in an audio document: identify the tracks of a CD audio from a reference database of CDs, or more generally retrieve metadata of an unknown audio file, delete audio duplicate in a database, identify a "live" song (broadcasted on the radio) via a mobile phone or any other recording device, etc.

- Audio structuring: search for occurrences like jingles as a first step for the analysis of radio or television contents (information retrieval, summarization).

- Media monitoring: confirm for advertisers if the planned advertisements were really broadcasted, detection of illegal use of multimedia content, etc.

Performing the audio identification task manually is quite tedious. Moreover, manual methods are slow and prone to errors. In automated systems, audio identification

is typically accomplished by audio watermarking [30] (Cox et al., 1996) or audio fingerprinting [25] (Cano et al., 2005). They are based on two different principles. The first one is intended to hide the essential information of identification in the audio document. In the second technique a signature (or fingerprint) is extracted from the audio content and compared to the reference fingerprints stored in a database. An audio fingerprint is a compact content-based signature that represents an audio recording. We are interested in methods based on audio fingerprinting, which are more appropriate for radio broadcast monitoring [25] (Cano et al., 2005).

This chapter is organized as follows. The ALISP-based audio fingerprinting system is presented in section 6.2. Then the experimental setup to evaluate the proposed systems is described in section 6.3. Studies about the number of Gaussian components, number of ALISP units and the method used for the initial segmentation are, respectively, reported in section 6.4, section 6.5 and section 6.6. Finally a comparison of the performances of our system with the systems participating in the 2010 QUAERO evaluation campaign is given.

6.2 ALISP-based Audio Fingerprinting

Radio broadcast monitoring consist on keeping a record of the timing and the occurrence of an audio content. It has an important role in the media industry. Generally, radio stations must pay royalties for the music they play. Even for radio stations which can play music for free, many companies are interested in detecting these music tracks for statistics purposes. Moreover advertisers are willing to monitor radio streams to verify the fulfillment of contracts by the broadcast channel for broadcasting the specific commercial between the stipulated times. Many commercial systems are providing these services, such as Broadcast Data System(www.bdsonline.com), Music Reporter(www.musicreporter.net), Audible Magic (www.audiblemagic.com), and YACAST(www.yacast.fr).

Our proposed system is used to search for advertisements and songs in radio broadcast streams. In this sense ALISP transcriptions of advertisements and songs are computed using HMM models provided by ALISP tools and Viterbi algorithm and compared to transcriptions of the radio stream using the BLAST algorithm [3] (Altschul et al., 1990).

The ALISP HMM models are first trained on the ALISP training database (288h), described in section 4.2, the number of ALISP units is 65 (64+ Silence model) and the average length per model is around 100ms. Compared to [57] (Haitsma and Kalker, 2002) which extracts 32-bit vector per frame, leading to 5,160 vectors per minute, ALISP methodology provides a very compact way to represent the audio data with 600 ALISP units per minute. Moreover our fingerprinting method is as compact as the audioDNA described in [24] (Cano et al., 2002) which extracts 800 gens per minute.

Figure 6.1 shows a spectrogram of excerpts from a reference advertisement and two spectrograms of the same advertisement streamed on two different radios with their ALISP transcriptions. Note the presence of some differences between ALISP transcriptions of the three advertisements. These differences could be explained by the similarity between some ALISP classes which leads to confusion during the recognition of these classes.

6.3 Experimental Setup

The ALISP-based audio identification system is used to search for advertisements and songs in radio broadcasted streams. To evaluate this system, two experimental protocols are proposed. The first protocol, denoted the SurfOnHertz protocol, is used for the identification of advertisements and songs in radio streams. It corresponds to 12 annotated days of radio broadcast provided by the framework of the ANR-SurfOnHertz project and divided in different parts as follows:

- Development database: five days of audio stream are used to study the stability of ALISP transcriptions of advertisements and to set the decision threshold for the Levenshtein distance.
- Reference database: it contains 2,172 advertisements and 7,000 songs leading to 9,172 reference items. The advertisement references correspond to the whole commercial item while only a one-minute-long excerpts of each reference song is kept. The position of these signatures within the tracks is unknown. The radio stream from whom a given reference was extracted is not part of the evaluation set.

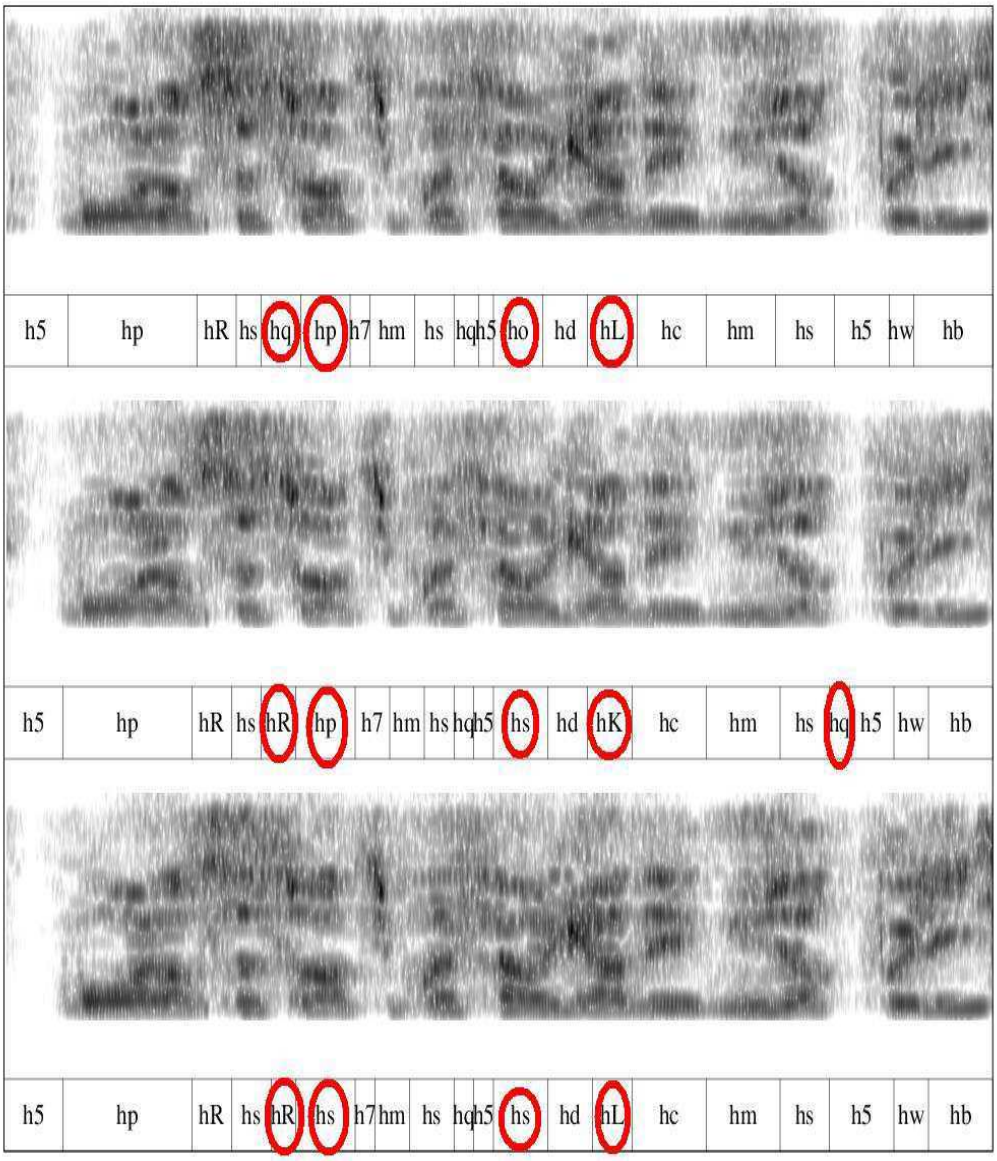


Figure 6.1: Advertisement spectrograms, taken from the radio broadcast corpus, with their ALISP transcriptions: first spectrogram is an excerpt from the reference advertisement, second one represents the same excerpt from French virgin radio and the last one represent the same excerpt from French NRJ radio.

- Evaluation database: seven days of audio stream from three French radios, these days are different from the ones used in the development database and the ALISP training database. This database contains 1,456 advertisements and 4880 songs.

The experimental protocol described above is also used to fix some parameters of the ALISP-based audio identification system:

- Number of Gaussian components: two ALISP HMM models are trained. The first is a mono-Gaussian model where each state is modeled with one Gaussian component. While the second is a multi-Gaussian model trained using the techniques described in section 3.3.

- Number of ALISP units: the initial number of ALISP units was 65 (64+ silence model). In order to find out if the set of possible ALISP units can be reduced, three new sets of ALISP models are trained and evaluated (using the same data) with 9, 17 and 33 units.

- Method of initial segmentation: as described in section 5.2, different methods for the initial segmentation of the ALISP training database are compared: temporal decomposition, uniform segmentation, spectral stability segmentation and phonetic segmentation.

Like the majority of the evaluations of audio fingerprinting systems, the evaluation protocol described above is applied on private corpora. Therefore a second experiment is done using a public evaluation framework for audio fingerprinting technologies. This framework is proposed by Ramona et al. [108] (Ramona et al., 2012) during the 2010 evaluation campaign of the QUAERO project (http://www.quaero.org). It is based on a scenario involving the detection of songs excerpts in broadcast radio streams. The evaluation protocol for this experiment on the QUAERO database is the following:

- Reference database : it contains only the 7,309 one-minute-long excerpts of songs.

- Evaluation database : It consists of the recording of 7 days of the French radio stream RTL captured and saved on disk in 5 minutes chunks. Therefore, the total duration

Day	Number of tracks
1	67
2	63
3	71
4	66
5	63
6	120
7	101

Table 6.1: Number of music track present in each day in the QUAERO evaluation set.

reaches 7 days x 24 hours per day = 168 hours. All of it was annotated by manually checking the output of an audio identification engine (with precision around 1 second). This database contains 459 music tracks distributed as it shown in table 6.1.

It is important to remind that all the sets of ALISP HMM models are trained on the ALISP training database which contains 288 hours from 12 French radios.

6.4 Number of Gaussian Components

In order to evaluate the contribution of the dynamic mixture splitting described in section 3.3.4, two sets of ALISP HMM models are trained with mono and multi-Gaussian models. The number of ALISP units for both models is 65 and temporal decomposition is used for initial segmentation. For the multi-Gaussian HMM model the number of Gaussian components per mixture is shown in figure 6.2. The mean value of Gaussian components used per mixture is 6.

The evaluation of the proposed audio identification system is performed using both ALISP HMM models to find out whether the use of multi-Gaussian models could improve the accuracy of the system. But before that, the stability of ALISP transcriptions of advertisement is studied. As shown in figure 6.1, the ALISP transcriptions of the same advertisements broadcasted on different radios is different. Therefore a comparison between advertisements present in the development data set is performed to fix the decision threshold of the Levenshtein distance.

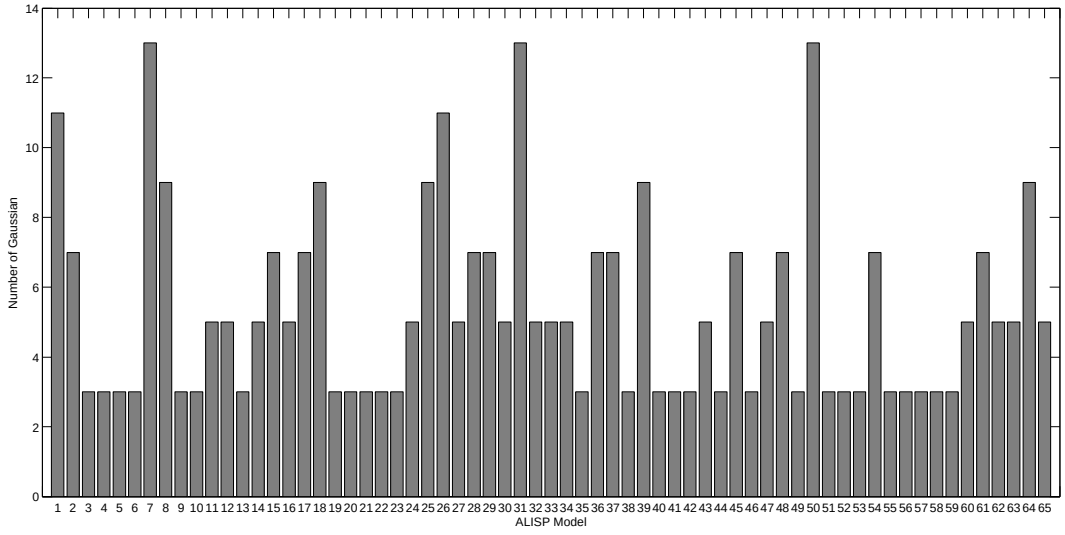


Figure 6.2: Number of Gaussian components used per mixture for the multi-Gaussian HMM model trained on the ALISP training database (288h).

6.4.1 Threshold Setting

To study the stability of ALISP transcriptions and determine the decision threshold, two experiments are realized on the mono and multi-Gaussian models obtained after the dynamic split of states mixtures on the development database:

- Compare ALISP transcriptions of the reference advertisements to the commercials in the radio recording (intra-pub experience).
- Compare ALISP transcriptions of reference advertisements to data that does not contain advertisements (extra-pub experience).

Figure 6.3 shows the distribution of the Levenshtein distances between ALISP transcriptions of references and advertisements in the radio recordings (denoted as mono-intra-pub and multi-intra-pub) and the distribution of the Levenshtein distances between ALISP transcriptions of references and data that do not contain advertisements (denoted as mono-extra-pub and multi-extra-pub).

Note that for both sets of HMM models, the two distributions (intra-pub and extra-pub) for the Levenshtein distance are disjoint. This result means that by choosing an appropriate decision threshold for the Levenshtein distance, there is a big chance that all advertisements in the radio streams can be detected.

As commonly observed for speech recognition systems, at a phone-like level with current ALISP models the transcriptions of audio data are not perfect. Therefore, when two different repetitions of the same advertisement are analyzed there are differences (that is the reason why we need to apply the Levenshtein distance). The number of transcription errors is proportional to the length of the advertisement. For long advertisement, there is a larger risk to find more transcription errors that lead to a bigger Levenshtein distance. On the other side, this study shows that ALISP transcriptions made with multi-Gaussian models are more precise than those made with mono-Gaussian models.

Once we have tuned the threshold of the Levenshtein distance on the development set, we can proceed to evaluate the proposed audio identification system with mono and multi-Gaussian models.

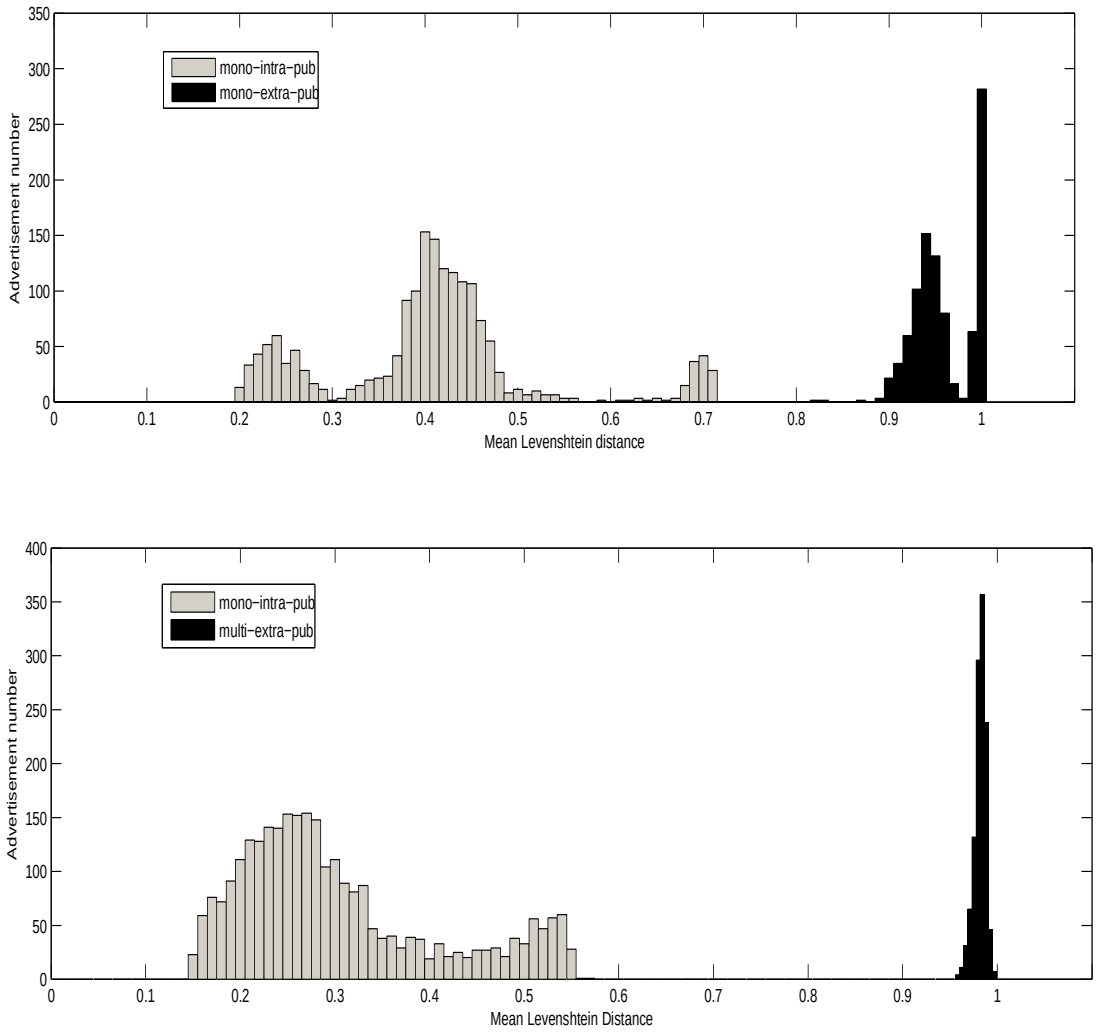


Figure 6.3: Distribution of the Levenshtein distance between ALISP transcriptions of references and advertisements in the development radio recordings for the mono-Gaussian model (denoted as mono-intra-pub) and the multi-Gaussian model (denoted as multi-extra-pub) and distribution of the Levenshtein distance between ALISP transcriptions of references and data that do not contain advertisements for mono-Gaussian model (denoted as mono-extra-pub) and multi-Gaussian model (denoted as multi-extra-pub).

Item	R%		P%		Missed item		False alarms	
	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2	Exp1	Exp2
Advertisement	98	98	93	100	27	27	102	0
Songs	92	92	96	100	389	389	176	0
All	93	93	95	100	416	416	278	0

Table 6.2: Precision (P%), recall value (R%), number of missed ads and number of false alarms found for each audio item. Results for the SurfOnHertz protocol (Seven days of audio stream for 3 French radios, containing 1,456 advertisements and 4,880 songs from YACAST database) with a threshold of 0.75 for mono-Gaussian model (Exp1) and 0.65 for multi-Gaussian model (Exp2).

6.4.2 Experimental Results

To detect commercials and songs in the test database we proceed as follows:

- Transcription of reference items by 65 ALISP HMM models (acquired from the ALISP development data set).
- Transcription of the test data to obtain its ALISP sequences.
- Setting the decision threshold on the development set of the Levenshtein distance to 0.75 for mono-Gaussian model and 0.65 for multi-Gaussian model to be sure to detect all items.
- Searching for each ALISP transcription of audio items in the ALISP transcriptions of each test audio stream using the proposed approximate matching process describes in section 5.3.

In order to evaluate the detection performance precision (P%) and recall (R%) rates are given in table 6.2:

- Recall : The number of items correctly detected / The number items that should be detected.
- Precision : The number of items correctly detected / Total number of detected items.

Table 6.2 shows that for both sets of HMM models, the system was not able to detect 416 audio items. These missed items belong to 389 songs and 27 commercials.

For music identification, 372 tracks are related to songs that have a different version from the one present in the reference database. For example, 302 live version songs from the test radio stream correspond to the studio version in the references. For commercial identification, the 27 missed advertisements are different from the reference ones. For example, there are 9 commercials spoken by different speakers who say the same things. These results show that the proposed system allows us to find errors in the manual annotations of songs and advertisements.

Moreover, we note the presence of 278 false alarms for the mono-Gaussian HMM models, while with the multi-Gaussian HMM models we observe no false alarms. This result proves that using one Gaussian per state to model 288 hours of audio data is not sufficient and could lead to many errors of identification. Hence, the dynamic split of the states mixtures during the HMM modeling step of the ALISP units is a good solution to overcome this problem.

Related to the processing time, the computational complexity of the system is mainly limited to the search for the closest ALISP sequence through the Levenshtein distance. With the 7,000 songs and 2,172 commercials database, the system runs at a speed of 0.57 per second of signal using the 65 ALISP models on a 3.00GHz Intel Core 2 Duo 4GB RAM, while for the brute search described in section 5.3.2.1 the systems runs at a speed of 6 seconds per second of signal. It's important to note that the approximate matching technique algorithm speeds up the ALISP transcriptions search without affecting the identification scores.

6.5 Number of ALISP Units

In the previous section, the number of ALISP units was 65 (64+ silence model). In order to find out if the set of possible ALISP units can be reduced to speed up the matching process, three new sets of ALISP models are trained and evaluated (using the same data) with 9, 17 and 33 units. All these models are trained using the multi-Gaussian configuration.

The first part of this section deals with the stability of the ALISP transcriptions for each set of ALISP models. Then the results obtained in terms of precision and recall are presented.

6.5.1 Threshold Setting

As for the setting of the number of Gaussian components, two experiences are realized in order to study the stability of the ALISP transcriptions for each set of ALISP models.

Figure 6.4 shows the distribution of the Levenshtein distance between ALISP transcriptions of references and advertisements in the radio recordings (denoted as intra-pub) and the distribution of the Levenshtein distance between ALISP transcriptions of references and data that do not contain advertisements (denoted as extra-pub), for the four sets of ALISP models.

Note that for 17, 33 and 65 ALISP models, the two distributions (intra-pub and extra-pub) for the Levenshtein distance are disjoint. This result means that by choosing an appropriate decision threshold for the Levenshtein distance, there is a big chance that all items in radio streams can be detected. Whereas, for the 9 ALISP models the two distributions overlap. From theses distributions, the Levenshtein distance thresholds were set to 0.35, 0.45, 0.55 and 0.65 respectively for the sets constituted from of 9, 17, 33 and 65 ALISP models.

6.5.2 Experimental Results

In order to evaluate the ALISP-based audio identification system performance recall (R%) and precision (P%) rates are used. Table 6.3 shows that same results are obtained in terms of missed items as for the mono/multi-Gaussian system described in the previous section. In fact, the ALISP-based systems were not able to detect 416 audio items using the four ALISP sets. These missed items belong to 389 songs and 27 commercials.

ALISP Set	Threshold	R%	P%	Missed Items	False Alarms
AL-65	0.65	93	100	416	0
AL-33	0.55	93	100	416	0
AL-17	0.45	93	96	416	129
AL- 9	0.35	93	92	416	334

Table 6.3: Recall (P%), Precision (R%) values, number of missed item and number of false alarms found for the SurfOnHertz protocol with a threshold of 0.65, 0.55, 0.45 and 0.35 respectively for 65, 33, 17 and 9 ALISP models.

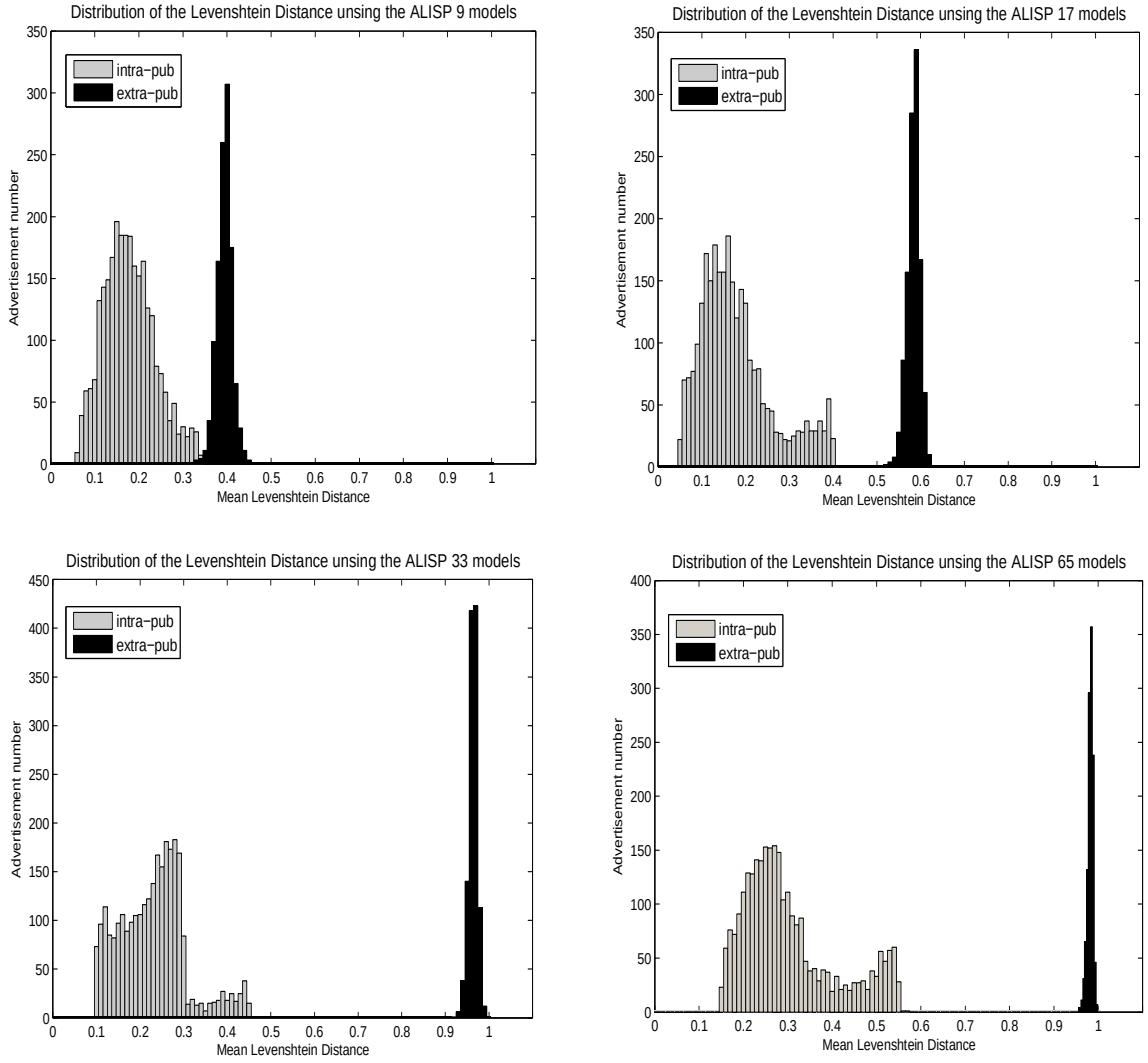


Figure 6.4: Distribution of the Levenshtein distance for the intra-pub and extra-pub experiences using the four sets of ALISP models, corresponding to 9, 17, 33 and 65 units.

Furthermore, the presence of 129 false alarms for 17 models and 334 false alarms for the 9 models were predictable given the decision threshold for the Levenshtein distance. This value was chosen to ensure the detection of all items even if false alarms have occurred.

Related to the processing time, the system runs at a speed of 0.49 per second of signal using the 33 ALISP models on a 3.00GHz Intel Core 2 Duo 4GB RAM, while for the 65 ALISP models the systems runs at a speed of 0.57 per second of signal.

6.6 Method of the Initial Segmentation

At this point, the temporal decomposition is used to obtain an initial segmentation of the audio data into quasi-stationary segments. Then, these segments are clustered using vector quantization. After that, boundaries together with labels will be used as initial transcription for Hidden Markov Modeling. However, other methods to provide an initial segmentation of the audio data are used. These methods are described in section 5.2.

In this section, the influence of the initial segmentation on the performances of the audio identification process is studied. Four different techniques to obtain an initial segmentation of the ALISP training database are used: temporal decomposition, uniform segmentation, spectral stability segmentation and phonetic segmentation. These techniques combined with the vector quantization are used to initialize the HMM models. All the acquired models are multi-Gaussian with 33 ALISP units.

The first part of this section involves the influence of the initial segmentation on the stability of the ALISP transcriptions of advertisements. Then the results obtained in terms of precision and recall are reported for each method of segmentation.

6.6.1 Threshold Setting

Same experiences used to study the stability of ALISP transcriptions and determine the decision threshold are realized. Figure 6.5 shows the distribution of the Levenshtein distances between ALISP transcriptions of references and advertisements in the radio recordings (denoted as intra-pub) and the distribution of the Levenshtein distances between ALISP transcriptions of references and data that do not contain advertisements (denoted as

System	R%	P%	Missed Items	False Alarms
Temporal decomposition	92	100	416	0
Spectral stability	92	100	416	0
Uniform segmentation	90	95	623	301
Phonetic segmentation	85	87	942	806

Table 6.4: Recall (P%), Precision (R%) values, number of missed item and number of false alarms found for the SurfOnHertz protocol for the different techniques of segmentation.

extra-pub), for the four techniques of segmentation.

Note, that the use of uniform segmentation, spectral stability segmentation and temporal decomposition leads to disjoint distributions (intra-pub and extra-pub) for the Levenshtein distance. Furthermore, this study shows that ALISP transcriptions obtained using the spectral stability segmentation and temporal decomposition are more precise than those with uniform segmentation.

On the other side, for the phonetic segmentation the two distributions overlap. This result is predictable given that the phonetic models are trained only on speech. In fact, the phonetic segmentation of audio data are used to initialize the ALISP HMM models, however the final ALISP models still vulnerable to distortions occurred in audio data which leads to many errors. From theses distributions, the Levenshtein distance thresholds were set to 0.55, 0.55, 0.65 and 0.65 respectively for temporal decomposition, spectral stability segmentation, uniform segmentation and phonetic segmentation.

6.6.2 Results

Table 6.4 shows the recall (R%) and precision (P%) rates when the different techniques of initial segmentation are used.

ALISP models using spectral stability segmentation perform as well as the models using temporal decomposition. This result confirms what is obtained in section 5.2.2, when maximal intersection between both models is computed.

For the uniform segmentation, the obtained results are slightly worse than those obtained with temporal decomposition and spectral stability segmentation. This could be explained by the fact that uniform segmentation don't take into account the audio signal

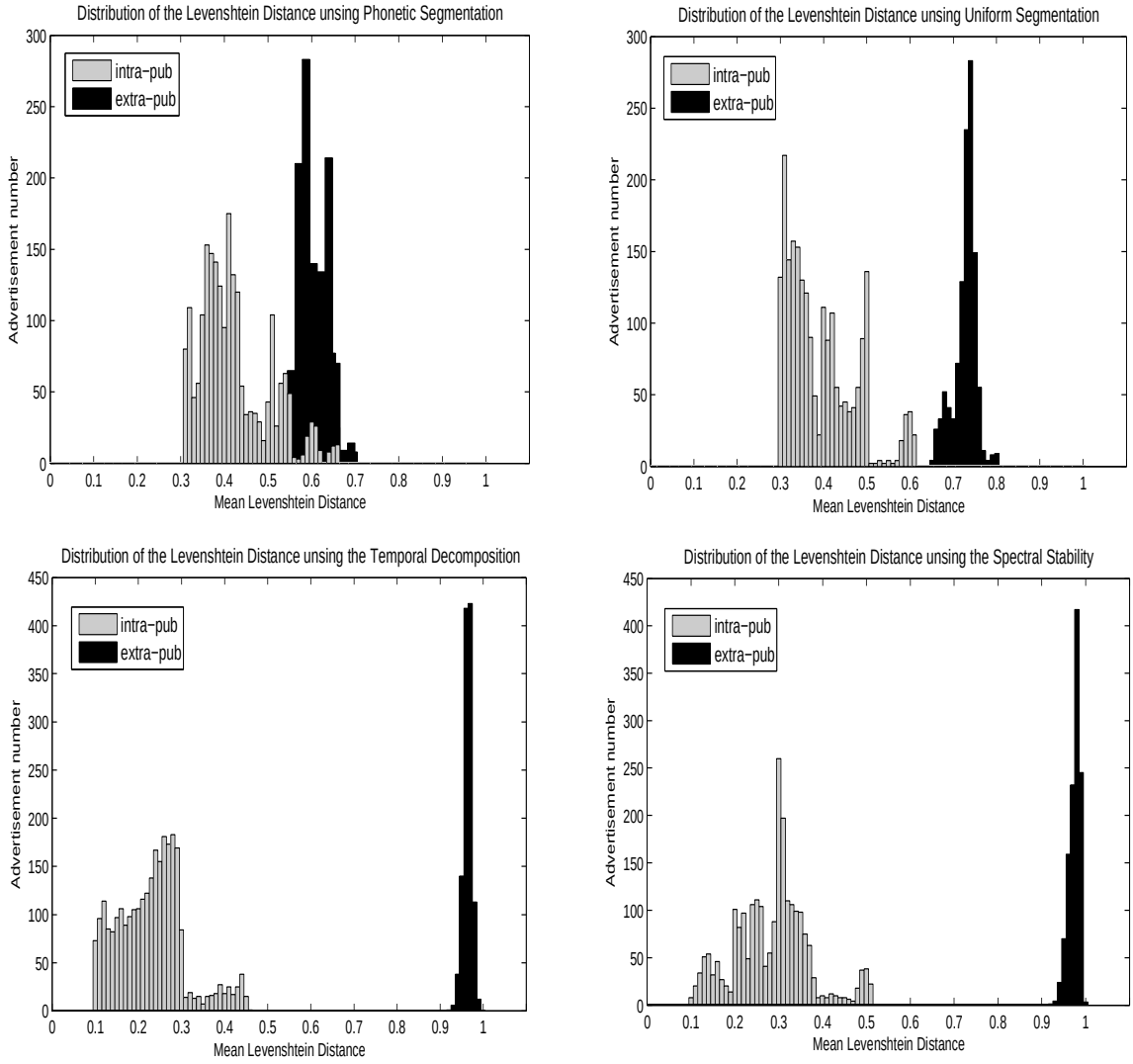


Figure 6.5: Distribution of the Levenshtein distance for the intra-pub and extra-pub experiences using the phonetic segmentation, uniform segmentation, spectral stability segmentation and temporal decomposition.

characteristics which affects the HMM modeling of the data.

As shown in the study of ALISP transcriptions stability of advertisements, the phonetic segmentation leads to many errors of identification and especially to a higher false alarms than the other techniques. This result confirms that phonetic segmentation are not appropriate to treat songs and advertisements.

6.7 Comparative Study

In this section, the ALISP-based audio identification system is evaluated using the QUAERO evaluation protocol in order to compare the performance of the proposed system with the systems provided by the participants of the 2010 QUAERO evaluation campaign [108] (Ramona et al., 2012), which are:

- Fenet et al. [44] (Fenet et al., 2011) developed a fingerprinting system based on the SHAZAM approach [133] (Wang, 2006).
- Ramona et al. [109] (Ramona et al., 2011) provided a system based on spectral modeling of bark-bands energy and synchronization through onset detection.
- YACAST implemented an audio fingerprinting system based on the Philips system [58] (Haitsma and Kalker, 2003)

Note that in this protocol the recognition of different versions of the same song title is considered outside the perimeter of audio identification. Therefore the number of music tracks to be detected is reduced to 459 tracks.

The set of ALISP HMM models used in this evaluation campaign is trained using the multi-Gaussian configuration with 33 ALISP units and the spectral stability method for the initial segmentation. The choice of spectral stability segmentation is motivated by its simplicity compared to the temporal decomposition.

Since the position of music signatures within the tracks is unknown, our task is limited only in detecting the music track in the test audio stream. For each detection of a reference, the system provides the name and the date of the song in the radio stream. If the date of

System	R%	P%	Missed Items	False Alarms
Our system	100	100	0	0
Fenet et al.	97.4	100	12	0
Ramona et al.	96.9	99	15	2
YACAST	95.9	99	17	0

Table 6.5: Precision (P%), recall rate (R%), number of missed tracks and number of false alarms found for the Quaero protocol (7 days of radio streams containing 459 songs to be identified).

detection is between the annotated start and end date of the same song, this song will be considered as detected.

Results are shown in table 6.5.

Table 6.5 clearly shows the relevance of our ALISP-based fingerprint, when compared to the other systems. However this protocol considers only one radio for evaluation which is not a real-world use-case. However, it gives an idea about the positioning of our system in the state of the art of audio fingerprinting.

6.8 Conclusion

In this chapter, the ALISP-based audio fingerprinting system was described. This system uses automatically acquired units provided by ALISP models to search for advertisements and music pieces in radio broadcast streams. In this sense ALISP transcriptions of advertisements and songs are computed using HMM models provided by ALISP tools and Viterbi algorithm and compared to transcriptions of the radio stream using the an approximate matching algorithm based on BLAST technique.

After that, many experiences were realized to determine the best configuration of ALISP HMM models. It was found that using the multi-Gaussian configuration with 33 ALISP units and the spectral stability method for the initial segmentation ensure the best performances of the proposed audio fingerprinting system. These models will be used in the following chapters to extract the ALISP units from audio data.

Regarding the results, on a set of 6,336 audio items (4,880 songs and 1,456 advertisements) 5,920 were detected without false alarms. 399 missed items are relative to songs

and commercials that are different from their references. Moreover, a comparative study showed that our system performs better than other audio fingerprinting systems on the task of music identification.

After all these studies and results, a legitimate question could be raised: "What if a song or a commercial is streamed for the first time in radio broadcast and do not have its signature in the reference database?". The next chapter will be dedicated on identifying audio items without references by extracting salient parts or by finding all repetitions of audio sequences in the entire database which should lead to automatic discovery of advertisements and songs.

Chapter 7

Audio Motif Discovery

7.1 Introduction

This chapter is a continuation of our work on indexing radio streams. In the previous chapter an audio identification system based on ALISP segmentation was described. The first step in the proposed system is the automatic acquisition and Hidden Markov Modeling (HMM) of ALISP audio models. Then a fingerprint database is created from a reference database using the automatically acquired units provided by ALISP tools. A reference database contains audio files (songs, jingles, advertisements,...) which the system can identify. In the last step an unlabeled audio excerpt is identified by comparing its fingerprint with those of the reference database using the BLAST algorithm.

In the case where an audio item is not in the reference database, the audio identification system could not detect it. An example of such a case is when a new song or advertisement is broadcasted by radio stations. These new items are usually played many times. Therefore the detection of repetitions of audio items in radio streams should lead to the automatic discovery of advertisements and songs without the need of a reference database.

The task of detecting repeating audio objects is also referred as audio motif discovery or near-duplicate discovery. As explained in [115] (Sandve and Drablos, 2006), the term "motif" is borrowed from comparative genomic, where it designates a family of symbol

sequences (each symbol representing a nucleotide or amino-acid). In our work, the term "motif" denotes the repeating objects in audio streams which are songs, advertisements and jingles. We are not dealing with motif discovery in speech data.

Performing this task is usually based on audio fingerprinting which involves a compact content-based signature that represents an audio recording. In this thesis, ALISP tools are used to convert the heterogeneous audio stream (containing music, jingles, commercials, speech, etc.) into a sequence of symbols. These symbols represent the fingerprint needed to detect the repeating items in audio streams. Therefore, the problem of repeating objects detection is transformed into a string matching problem.

This chapter is organized and structured in the following manner. The first section deals with related work to audio motif discovery. Then the ALISP-based near-duplicate discovery system is presented. In section 7.4, the evaluation of the proposed system is given.

7.2 Related Work

The amount of audio data available, such as broadcast news archives, radio recordings, music and songs collections, podcasts, etc, has increased exponentially in the past decades. However, most of these data have limited label information, or worse yet, have no label information.

Therefore, it is not easy for users to locate a desired song or speaker in such databases, or to skip unwanted contents. To overcome these limitations, various systems were developed to rapidly analyze and summarize audio content for indexing and retrieval purposes.

One of the purposes of these systems is to detect repeating items in audio streams, also refereed as audio motif discovery. Locating repetitions of unknown audio objects is useful for many reasons. Herley mentions in [61] (Herley, 2006) many applications for repeating objects detection such as:

- Commercial skipping: the detection of repeating objects in radio (or TV) streams allows the deletion of all unwanted contents (such as advertisements).

- Compression and archiving: the repeating objects might be used to compress the audio streams efficiently [7] (Apostolico et al., 2006).
- Broadcast monitoring: confirm for advertisers if the advertisement was broadcasted, detection of illegal use of multimedia content.
- Audio structuring: search for repeated occurrences, like jingles, as a first step for the analysis of radio or television contents.

7.2.1 Problem Formulation

Detecting repeating audio objects in audio streams consists of finding all pairs of disjoint audio segments $[x, y]$ and $[u, v]$ which verify these three conditions:

$$D([x, y], [u, v]) < thr \tag{7.1}$$

$$|x, y| < L_{min} \tag{7.2}$$

$$x < y < u < v \tag{7.3}$$

Condition 9.1 is relative to the similarity constraint. It considers that two audio segments are similar if their distance D is below a certain threshold thr. The second condition is used to define the minimum length (L_{min}) of the repeating item. The last condition is there to avoid the detection of two overlapping segments.

7.2.2 Literature Review of Audio Motif Discovery

Most of audio motif discovery systems rely on the same principle: audio fingerprinting [95] (Ogle and Ellis, 2007) [21] (Burges et al., 2005) [43] (Fenet et al., 2012) [121] (Sinitsyn, 2006) [90] (Muscariello et al., 2011). We review in this section the most representative works of audio motif discovery.

In [95] (Ogle and Ellis, 2007), a framework to identify repeating sound events in long-duration personal audio recording is proposed. This system adapts the sparse landmark

fingerprint and hashing technique proposed in [133] (Wang, 2006) to search and retrieve the repeating sound events. The system is evaluated on 40 hours of personal recording containing 30 songs and 45 telephone rings that are repeated 10 times. The system achieves a recall rate of 97% and a precision rate of 85 for songs while 69% and 100% are obtained, respectively, for recall and precision in the case of telephone rings. However, this method do not perform well with organic sounds, such as impact transients, machinery, door closure and speech.

Another framework for the detection of repeating objects in multimedia streams is described in [43] (Fenet et al., 2012). First, a fingerprint is extracted from each frame present in the audio stream. Then, each fingerprint is compared to the database containing the past frames fingerprints. Based on this comparison a repetition detection decision is taken. In case of positive match, the storage database, which contains all the processed fingerprints, is updated so that it will not store repeated frames in the database. Two audio motif discovery systems are used for this framework. The first system is based on the Constant-Q-Transform (CQT) [44] (Fenet et al., 2011) where a 2-dimensional peak-peaking in CQT spectrogram is extracted for each frame. Then, the extracted peaks are clustered in pairs. The time occurrences of these pairs are given by the temporal localizations of the first peaks in the pairs. The second system exploits a fingerprint based on a sparse decomposition of the signal in a redundant dictionary using the Matching Pursuits (MP) algorithm proposed in mallat-sp-1993 (Mallat and Zhang, 1993). The two systems are evaluated on a 24-hours radio stream that contains 191 repetitions of songs. For the CQT-based system, all the repetitions are detected without false alarms, while the MP-based system misses 13 repetitions and records one false alarm.

A different algorithm to detect duplicate songs is proposed in [21] (Burges et al., 2005). The system is based on the Robust Audio Recognition Engine (RARE) audio fingerprinting system [22] (Burges et al., 2003). It consists on transforming an audio segment into 64 floating-point numbers and using a weighted Euclidean distance to search for repeating songs. The system is evaluated on 21,322 songs for which one or more duplicates should be detected. 259 mismatches are found leading to a detection rate of 98.8%. This framework is

also exploited for audio thumbnail generation, where the task is to find a short representative summary of the music track. The proposed system is used to find repeating parts within the audio clip. In fact, if a song has a similar chorus segments, the system will be able to exploit that redundancy to generate a good thumbnail which outperforms the use of a random thumbnail.

In [90] (Muscariello et al., 2011), the authors adapt an existing system to identify short and highly variable patterns in speech to detect repeating songs in radio stream. The proposed system is based on the Automatic RepeatinG Object Segmentation (ARGOS) framework and Segmental Locally Normalized Dynamic Time Warping (SLNDTW)-based pattern matching of audio sequences. The fingerprint used in this system is a simple conversion of raw data to MFCC features. In order to speed up the search of repetitions a regular downsampling of MFCC sequences is integrated in the framework. The system is evaluated on 6 days of radio streams that contain 1,742 songs. For a set of 208 repeating songs a precision rate of 100% with the corresponding recall value of 70% are achieved.

Audio motif discovery has also been applied to analyze the musical structure and perform audio thumbnail generation. As explained in [33] (Dannenberg and Hu, 2002), the audio data are first transcribed into a sequences of representations such as monophonic pitch estimation [32] (Dannenberg, 2002), chroma representation [14] (Bartsch and Wakefield, 2001), and polyphonic transcription [83] (Marolt, 2001). Then similar segments are searched within the piece of music. After that, the obtained segments are grouped into classes in order to analyze the structure of the song.

Most of the system described above are evaluated on repeating songs with long duration (about 5min). In the next section, the ALISP-based audio motif discovery system is described and evaluated on advertisements and songs where the duration could vary from few seconds (3s) to some minutes (7min). This system uses the ALISP symbols to represent the audio data and the BLAST algorithm to search for repetitions.

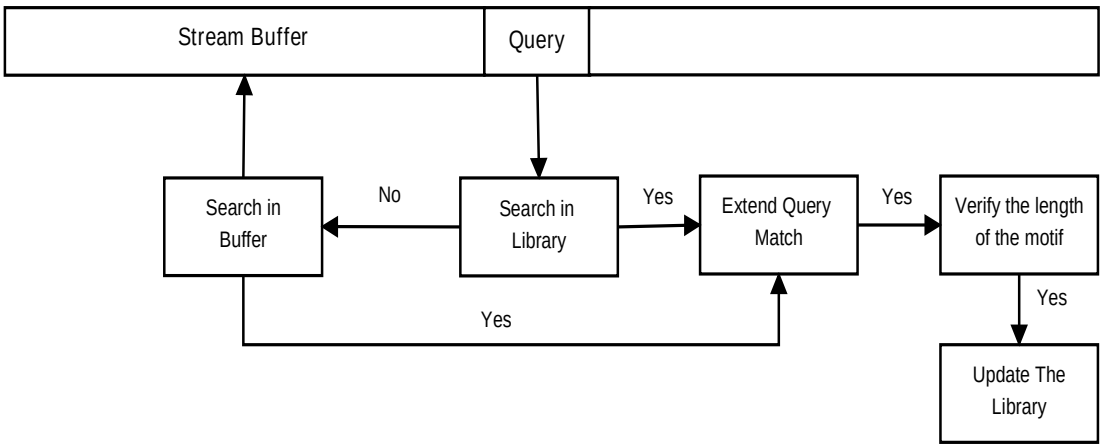


Figure 7.1: Main architecture of the ARGOS segmentation framework.

7.3 ALISP-based Audio Motif Discovery System

In order to detect repeating objects, the ARGOS segmentation framework proposed in [61] (Herley, 2006) is combined with ALISP tools. The main advantages of the ARGOS framework is its ability to work in a streaming mode (where the future data are not available). The architecture of the ARGOS framework is illustrated in figure 7.1. It consists of a sequential algorithm to find repetitions in audio stream. The query is an audio segment to be searched in the motif library and the received stream (buffer stream). If a positive match is found, an extension of the query matching is performed to find the entire audio item. This item is considered as a repetition if it meets the conditions described in section 7.2.1.

This framework is combined with the ALISP-based audio identification system to build a new audio motif detection system. As previously mentioned, the data-driven ALISP technique converts the raw audio data into a sequence of symbols. These symbols represent the fingerprint used to detect the repeating items in audio streams. Therefore, the problem of audio motif detection is transformed into a string matching problem. As in the ALISP-based audio indexing system described in section 5.3, the BLAST algorithm is implemented to speed up the approximate string matching. This algorithm is used in the ALISP-based audio motif detection system to accelerate the query search in the motif library and the

buffer stream.

7.4 Experimental Setup and Results

In our work, we are interested by detecting repeating songs and advertisements in radio streams. Comparing different motif detection systems remains impossible, since no public evaluation framework or corpus has been proposed. Therefore, the ALISP-based motif detection system is evaluated on the SurfOnHertz protocol previously used for audio identification.

The first part of this section deals with the experimental setups. Then, studies related to the stability of the ALISP transcriptions to set the decision threshold for the Levenshtein distance are presented. Finally results obtained in terms of precision and recall are exposed.

7.4.1 Experimental Protocol

In the previous chapter, we showed that the optimal configuration of ALISP HMM models is the one using the multi-Gaussian modeling with 33 ALISP units and the spectral stability method for the initial segmentation. These models were trained on 288 hours of audio data and already exploited for audio identification as described in the previous chapter. They will be also exploited in this chapter and the following to compute the ALISP transcription of audio data.

The ALISP-based motif detection system is evaluated the SurfOnHertz protocol where seven days of audio stream from three French radios (21 days) are considered.

7.4.2 Threshold Setting

To study the stability of ALISP transcriptions and determine the decision threshold, two experiments are realized:

- Compute the Levenshtein distance between the ALISP transcriptions of repeating songs (rep-song experience).

- Compute the Levenshtein distance between the ALISP transcriptions of different songs (diff-song experience).

These experiences realized on one day of audio from "Radio Nostalgie" with 347 broadcasted songs. Among these songs, 47 are repeated.

Figure 7.2 shows the distribution of the Levenshtein distances between ALISP transcriptions of repeating songs (denoted as rep-song) and the distribution of the Levenshtein distances between ALISP transcriptions of different songs (denoted as diff-song). Note that the two distributions (rep-song and diff-song) for the Levenshtein distance are disjoint. The mean Levenshtein distance for rep-song experience is 0.32 while for diff-song experience this value is equal to 0.85. This result means that by choosing an appropriate decision threshold for the Levenshtein distance, there is a big chance that all repeating items in radio streams can be detected.

7.4.3 Results

7 days of audio stream from 3 French radios (leading to 21 days) are used to evaluate the proposed system. These data contain 4,880 songs and 1456 advertisements, yielding, respectively, to an average duration of 210 seconds and 29 seconds. The shortest song and advertisement have, respectively, a duration of 59 seconds and 5 seconds, while the longest ones has, respectively, a duration of 411 seconds and 43 seconds.

Among all songs in the evaluation database, 348 are repeated with a total number of 3081 repetitions. The most repeated motif occurs 24 times while the average number of repetitions is 4. For advertisements, the most repeated motif occurs 16 times while the average number of repetitions is 2. The total number of repeated advertisements is 1315.

In order to evaluate the ALISP-based motif detection system performance precision (P%) and recall (R%) rates are used:

- Precision: the number of motifs correctly detected / Total number of detected motifs
- Recall: the number of motifs correctly detected / The number of motifs present in the audio stream

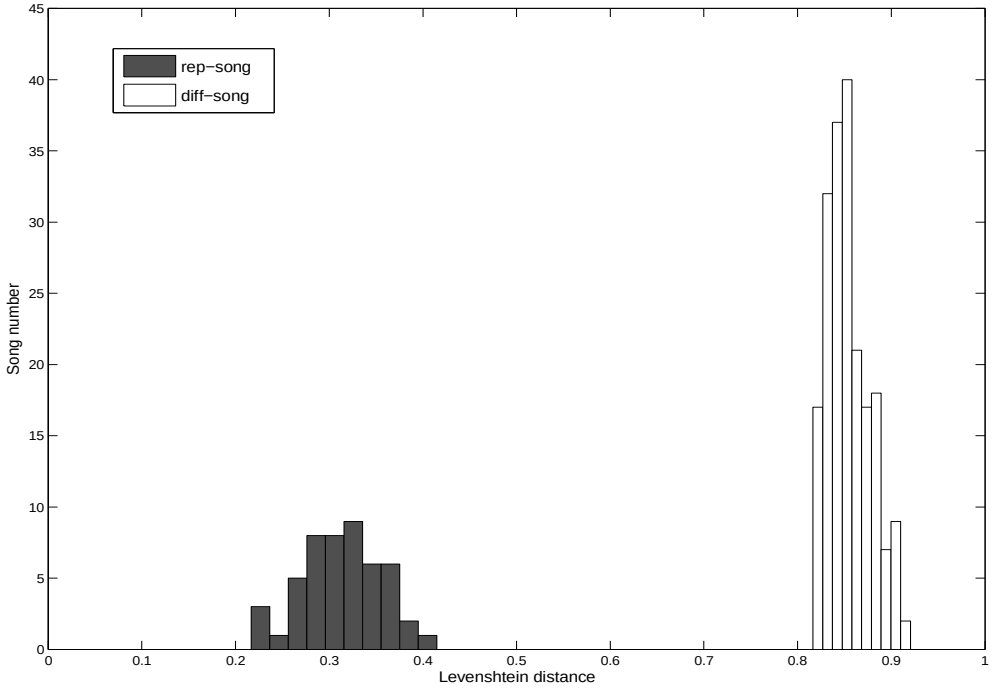


Figure 7.2: Distributions of the Levenshtein distance between ALISP transcriptions of repeating songs (denoted as rep-song) and different songs (denoted as diff-song).

The evaluation database contains various repeating objects other than songs and advertisements, such as jingles and speech segments. Since the manual annotations relative to these objects are not available, the detection of these items is considered to be out of the scope of this work.

The evaluation process is defined as follows. When the system detects a repetition, it gives its detection time to be checked if it does actually correspond to a repeating song. If a detected motif does not match any annotated repeating song, it will be considered as a false alarm. The results of the experiments are summarized in table 7.1.

	Rep	R%	P%	MD	FA
Songs	3081	99	100	21	0
Ads	1315	98	99	14	6

Table 7.1: Number of repetitions (Rep), precision (P%), recall value (R%), number of missed detection (MD) and number of false alarms (FA), found in the evaluation database for songs and advertisements

For songs, the system is not able to detect 21 repetitions. These repetitions are related to songs overlapped with speech which disturbs the detection process. On the other hand, the absence of false alarms confirms the results obtained on the development database where the Levenshtein distance distributions of rep-song and diff-song experiences are disjoint.

For the advertisements, the system is not able to detect 14 repetitions and leads to 6 false alarms. In fact, these errors are related to the detection of two repetitions of two successive advertisements and one repetition of three successive advertisements. In the manual annotation these repeated advertisements are annotated as separate motif. This is the origin of this errors.

Moreover, this evaluation database was also used for the audio identification task and a mean precision rate of 100% with the corresponding recall value of 95% were achieved. These results show that the ALISP-based audio indexing system is generic and could be applied on different tasks for the same radio streams.

It's important to note that the ALISP-based motif detection system performs as well as the two systems described in [43] where the evaluation database is a 24 hours of a French radio. The first system used a fingerprint based on the Shazam system while the second

one used a sparse decomposition-based fingerprint.

7.4.4 Runtime

The computation time required to detect the repeating objects is an important parameter to be considered. Therefore the BLAST algorithm is used to speed up the approximate ALISP symbols matching.

For the ALISP fingerprinting computing the processing time is 0.04s per second of signal. Accordingly, the computational complexity of the system is mainly limited to the search for the closest ALISP sequence through the Levenshtein distance. Using the BLAST algorithm, the system needs 15 hours to process 24 hours of radio streams using 33 ALISP models on a 3.00GHz Intel Core 2 Duo 4GB RAM, while for the brute search the runtime is estimated at 10 days to process 1 day of radio stream.

It's important to note that the BLAST algorithm speeds up the ALISP transcriptions comparison without affecting the detection scores.

7.5 Conclusion

In this chapter a motif discovery system to detect repeating songs and advertisements in radio streams was described. As for the audio identification, the proposed system is based on ALISP sequencing to represent the audio data and BLAST algorithm to accelerate the approximate ALISP units matching. As pointed before, this architecture is common for all the audio indexing systems presented in this thesis.

The ALISP-based audio motif discovery system was evaluated on 6 days broadcast corpus of 4 radios. On a set of 975 motifs a mean precision rate of 100% with the corresponding recall value of 97% were achieved. Moreover, BLAST algorithm was able to speed up the searching step without affecting the detection scores.

In the next chapter, we will focus on the speaker diarization task, that aims to segment an input audio stream into homogenous regions according to speaker's identities in order to answer the question "Who spoke when?". In our work, we are interested in speaker diarization for TV and radio shows. Usually these shows keep the same structure with

same presenters and jingles. This redundancy is used in order to improve the performance of the speaker diarization system. The main idea of our system is to compare the show to be segmented with the same show broadcasted before in order to find the common audio segments. This operation is performed using the ALISP sequencing and BLAST algorithm.

Chapter 8

Speaker Diarization

8.1 Introduction

In this chapter, the proposed audio indexing system is applied to speaker diarization on TV and radio shows, where the goal is to segment an input audio stream into homogenous regions according to speaker's identities in order to answer the question "Who spoke when?".

Speaker diarization is also known as a speaker segmentation and clustering. The speaker segmentation step aims to detect the boundaries of speech segments by finding the speaker change points or more generally the acoustic change points. Then, speaker clustering is applied to group together the speech segments that seem to be pronounced by the same speaker. The general architecture of a speaker diarization system is illustrated in figure 8.1 [41] (ElKhoury, 2010). It is generally composed of four steps:

- Parameterization.
- Voice Activity Detection (VAD).
- Speaker segmentation.
- Speaker clustering.

As mentioned before, speaker diarization provides useful information related to the speaker identities. Coupled with automatic speech recognition this knowledge is useful in many applications, such as:

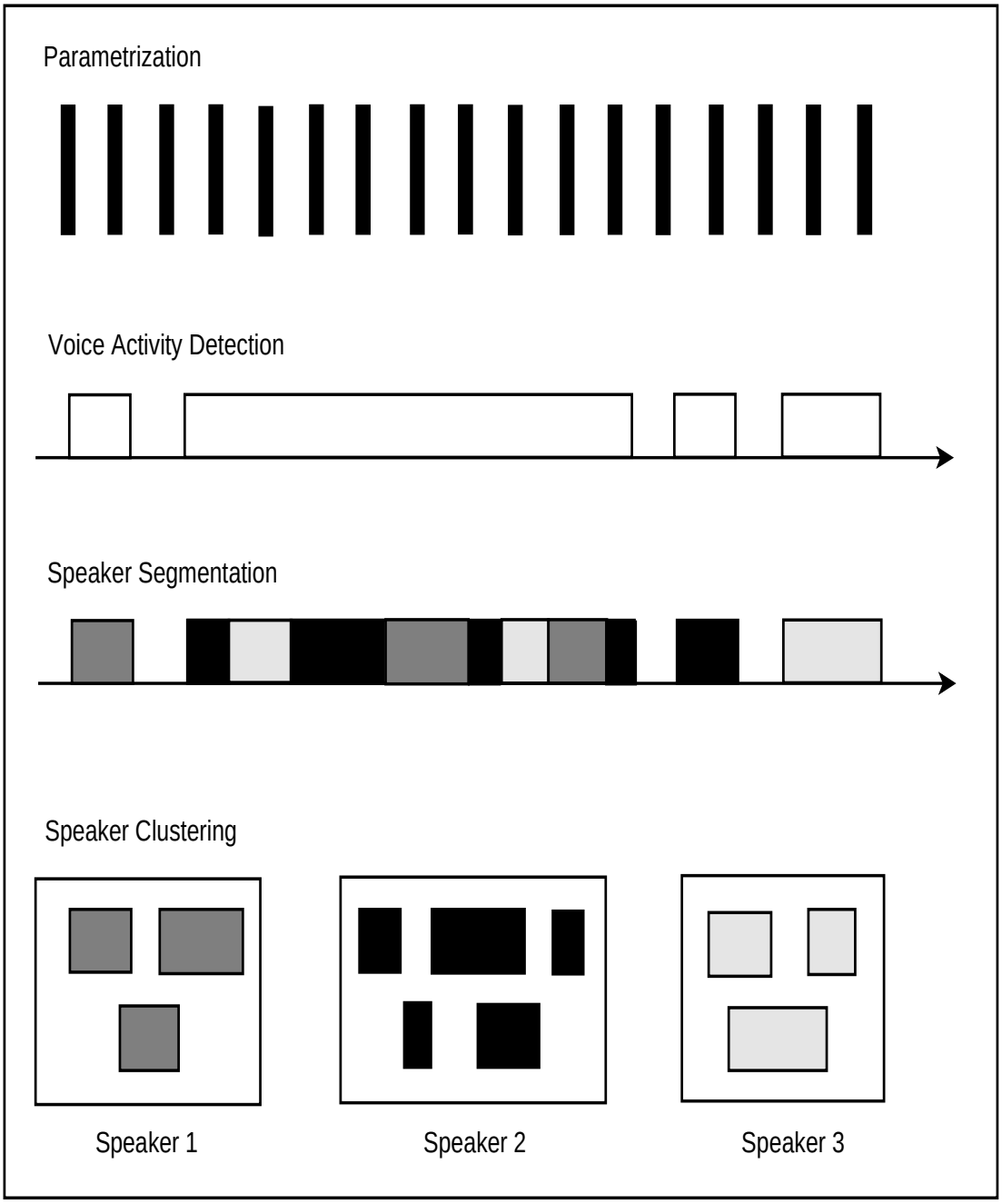


Figure 8.1: General architecture of a speaker diarization system.

- Rich transcription: speaker diarization is performed as a preliminary step in every task of information retrieval such as the speech duration of politicians during an election campaign or the tracking of a particular person for summarization and indexing purposes.
- Automatic speech recognition: the main goal of speaker diarization is to identify the speech segments pronounced by the same speaker. These segments are exploited for speaker adaptation to enhance the automatic speech recognition performance.
- Speaker-based algorithms: speaker diarization is the basis of several applications such as, speaker tracking, speaker verification, speaker identification and other speaker-based algorithms.

In this work, we are interested in speaker diarization for TV and radio shows which include various acoustic sources such as studio/telephone speech, music, or speech over music. Usually these programs tend to keep the same structure with same presenters, reporters, sound effects, jingles, etc. This redundancy is used in order to improve the performance of the speaker diarization system.

This chapter is organized as follows. The first section presents the state of the art of speaker diarization. In section 8.3, the ALISP-based speaker diarization system is detailed. Section 8.4 presents the experiments and the results.

8.2 State of the Art of Speaker Diarization

Many speaker diarization systems are described in [124] (Tranter and Reynolds, 2006) and [5] (Anguera et al., 2012). As previously mentioned, these systems are composed of four modules: acoustic features extraction, speech detection, speaker segmentation and speaker clustering. This section is organized as follows. In section 8.3.1, some of the acoustic features that are suitable for speaker diarization are listed. In the next section, a review of speech detection algorithms is presented. Then, the different techniques used for speaker segmentation and speaker clustering are respectively introduced in sections 8.3.3 and 8.3.4. Finally the main recent search directions for speaker diarization are described.

8.2.1 Acoustic Features

As for speaker and speech recognition systems, the parameterization in speaker diarization system is based on frame-level features such as the Mel Frequency Cepstrum Coefficients, the Linear Frequency Cepstrum Coefficients, the Linear Predictive Coding and Perceptual Linear Predictive.

Moreover, in the field of VAD, other features are proposed such as energy, spectrum divergence between speech and background noise, 4 Hertz modulation energy, pitch and zero crossing rate. In addition, for music detection other features like the number and the duration of the stationary segments obtained from a forward/backward segmentation are used. In the following part, some of these acoustic features are described:

- Mel Frequency Cepstrum Coefficients (MFCC): The MFCCs [88] (Mermelstein, 1976) are by far the most frequent features for speech processing. The MFCCs are commonly extracted as show in figure 8.2. First, the audio signal is windowed using a Hamming function. Then the Fourier transform is applied on each window. After that, the powers of the spectrum obtained above are transformed to the MEL scale using triangular overlapping windows. Finally, the Discrete Cosine Transform of the obtained powers are calculated. The MFCCs are the amplitudes of the resulting spectrum.

- 4 Hertz modulation energy: The speech signal has a characteristic energy modulation peak at 4Hz syllabic rate. The 4Hz modulation energy is exploited in [105] (Pinquier et al., 2003) in order to segment speech and music. To compute this feature, the audio signal is transformed into 40 perceptual channels according to the same process used to compute the MFCCs features. Then the energy of each band is filtered by a bandpass filter with a center frequency of 4Hz. After that, the filtered energies are summed and normalized. Finally, the desired feature is obtained by computing the variance of the filtered energy.

- Zero Crossing Rate (ZCR): The ZCR is the rate at which the sign of the audio signal changes. This feature is often extracted for speech/music segmentation. It is

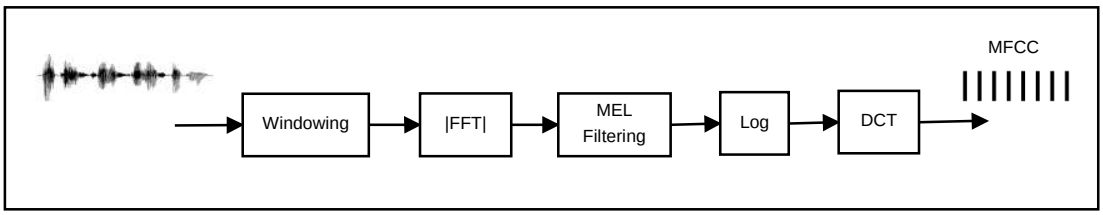


Figure 8.2: Extraction method of MFCC features.

computed for each frame as follows:

$$ZCR(i) = \frac{1}{2N} \sum_{n=1}^{N} |sign(x_n(i)) - sign(x_{n-1}(i))| \quad (8.1)$$

Where $x_n(i)$ is the n^th sample in the frame i and N is the size of frame i.

- Energy: The energy is a temporal feature commonly used in signal processing. It is computed as follows:

$$E(i) = \sum_{n=1}^{N} x_n^2(i) \quad (8.2)$$

After the parameterization, the next module in a speaker diarization system is the voice activity detection which allows the detection of the speech parts in the audio signal.

8.2.2 Voice Activity Detection

Voice activity detection involves the labeling of speech and non-speech segments in the audio signal. This module have a significant impact on speaker diarization performance. If a speaker segment is not labeled as speech, it is counted as a missed detection. On the other hand, a non-speech segment, which is labeled as speech, can affect negatively the speaker diarization process. Many different approaches and studies of such a system are proposed in the literature [106] (Ramirez et al., 2007). These systems are generally divided into two categories.

The first category concerns the model-based approaches. These approaches rely on a two-class detector, with models trained on external speech and non-speech data [6] (Anguera

et al., 2005) [42] (ElKhoury et al., 2009). The models are usually based on Gaussian Mixtures or Hidden Markov Models. Speech and non-speech models are usually adapted to specific conditions such as noise and channel. The main drawback of model-based approaches is their dependency on the training data.

The second category is related to unsupervised methods. These methods use acoustic features such as 4Hz modulation energy, energy or the number and the duration of segments, described in the previous section, to discard non speech regions [107] (Ramirez et al., 2003) [105] (Pinquier et al., 2003). The main drawback of these methods is the use of a threshold decision to detect speech. Generally, this threshold is determined empirically on a developmental corpus.

Hybrid approaches have been developed to overcome these limitations. Generally, a threshold-based method is first applied to detect speech and non speech segments with high confidence of classification. Then, the labeled data is employed to train speech and non-speech models. These models are then exploited to obtain the final segmentation of the audio signal [4] (Anguera et al., 2006) [42] (ElKhoury et al., 2009).

Once the voice activity detection is performed, the next step in the diarization process is the speaker segmentation.

8.2.3 Speaker Segmentation

During speaker segmentation (also referred as speaker change detection) the audio stream is split into homogeneous segments by detecting changes in speakers. Each segment should contain the speech of one speaker and two consecutive segments should contain the speech of two different speakers. Two main types of speaker segmentation systems can be found in the literature. The first category concerns the metric-based segmentation methods. While the second category involves other methods which rely on non metric-based segmentation approaches. In this works, we are only interested in metric-based speaker segmentation methods that are exposed in this section

Metric based speaker segmentation is the most common technique used to detect speaker changes in audio streams [5] (Anguera et al., 2012). It relies on the computation of

a distance between two adjacent segments, usually in overlap, to figure out if they belong to the same speaker. Most of the distances computed for speaker segmentation are also applied to speaker clustering.

Metric-based speaker segmentation systems are divided into two category. The first type of systems performs a single processing pass to detect the speaker turns boundaries. In the second class, a two-pass method is carried out. The first pass yields many change points with a high false alarm rate, in the second pass the detected change points are reconsidered to enhance the speaker segmentation output.

The most used distances are described in the following sections. These distance are calculated in the speaker segmentation step for either a single processing pass or two-pass processing method.

8.2.3.1 Generalized Likelihood Ratio

The Generalized Likelihood Ratio (GLR) is introduced by [52] (Gish et al., 1991). It considers that for each audio segment there are two possible hypotheses:

- H_0: This hypothesis supposes that the segment $X = x_1 \cdots x_N$ is produced by a single speaker. Therefore, the segment X is modeled by a multi-Gaussian distribution.
- H_1: This hypothesis supposes that the segment $X = x_1 \cdots x_N$ is produced by two different speakers representing two different segments: $X_1 = x_1 \cdots x_i$ and $X_2 = x_{i+1} \cdots x_N$. In this case, the segment X is modeled by two multi-Gaussian distribution.

As described in [65] (Jin et al., 2004) and [59] (Han and Narayanan, 2008), the GLR is determined to estimate the ratio between the probabilities of the hypothesis H_0 and the hypothesis H_1 as follows:

$$GLR = \frac{P(H_0)}{P(H_1)} \tag{8.3}$$

In terms of likelihood, the previous equation is given by:

$$GLR = \frac{L(X|M)}{L(X_1|M_1)L(X_2|M_2)} \tag{8.4}$$

where M, M_1 and M_2 are, respectively, the estimated model of X, X_1 and X_2 and $L(|)$ is the likelihood function. By considering a Gaussian distribution of the models, the previous expression becomes:

$$R(i) = -\log(GLR) = \frac{N}{2}\log|\Sigma_X| - \frac{N_1}{2}\log|\Sigma_{X_1}| - \frac{N_2}{2}\log|\Sigma_{X_2}| \tag{8.5}$$

where Σ_X, Σ_{X_1} and Σ_{X_2} are, respectively, the covariance matrices of X, X_1 and X_2 and N, N_1 and N_2, are respectively the size of X, , X_1 and X_2. The estimated value of the point of change is given by:

$$\hat{i} = \arg\max_i R(i) \tag{8.6}$$

Finally, a threshold T is defined in order to detect the point of speaker change. In fact, if $\hat{(}i)$ is greater that the threshold T, the segment X belongs to two different speakers and $\hat{i}$ is designed as the change point.

The main drawback of the GLR measure is the existence of the threshold T. This threshold is determined empirically using some external data.

8.2.3.2 Bayesian Information Criterion

The BIC distance is the most common approach for speaker segmentation [5] (Anguera et al., 2012), it is given by:

$$BIC(M) = \log L(X|M) - \frac{\lambda}{2} n \log N \tag{8.7}$$

where n represents the number of feature vectors used to build the model M. The BIC value is composed of two terms. The first one gives the log-likelihood of the data given the model. While the second term represents the complexity of the data. λ is a penalty coefficient, theoretically set to 1 [112] (Rissanen, 1989).

By considering the two hypothesis H_0 and H_1 defined in the previous section, the difference between BIC measures related to the two hypothesis is:

$$\Delta BIC(i) = R(i) - \lambda P \tag{8.8}$$

where R(i) is the likelihood ratio defined in equation 8.5 and P is the data complexity term:

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N \tag{8.9}$$

where d is the size of the feature vector. The larger the value of $\Delta BIC(i)$, the less similar the two segments is. Therefore, if:

$$\max_i \Delta BIC(i) > 0 \tag{8.10}$$

The time index corresponding to this maximum value is considered as a speaker change point. Unlike the GLR criterion, the BIC segmentation method do not require a threshold and the penalty coefficient λ is fixed theoretically to 1. However, other studies suggest that this value is not necessarily equal to 1 [125] (Tritschler and Gopinath, 1999) [35] (Delacourt et al., 1999) [80] (Lopez and Ellis, 2000). Moreover, BIC-based speaker segmentation are computationally more expensive than other metrics. Therefore, some systems propose to consider a two-pass processing method, where the BIC metric in employed in the second pass, while a faster metric is performed in the first pass [36] (Delacourt and Wellekens, 2000).

8.2.3.3 Kullback-Leibler Divergence

Like the GLR and the BIC metrics, The Kullback-Leibler (KL) divergence [72] (Kullback and Leibler, 1951) is a distance between two random probability distributions. Given two probability distribution P and Q, the KL divergence is given as follows:

$$KL(P||Q) = \int_{-\infty}^{+\infty} \ln \left[\frac{p(x)}{q(x)} \right] p(x)dx \tag{8.11}$$

As shown in the previous equation, the KL divergence is an asymmetric distance which make its use unsuitable for speaker segmentation. Therefore a symmetric version of the KL divergence, denoted as KL2, is proposed [139] (Zhu et al., 2006):

$$KL2 = \frac{KL(P||Q) + KL(Q||P)}{2} \tag{8.12}$$

By considering two adjacent windows of the audio signal with the Gaussian distributions $N_1(\mu_1, \sigma_1)$ and $N_2(\mu_2, \sigma_2)$ the previous equation becomes:

$$KL2 = \frac{1}{2}\left[\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + (\mu_1 - \mu_2)^2\left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right]\right] \tag{8.13}$$

where μ_i and σ_i are, respectively, the mean and the covariance of the Gaussian distribution N_i. The local maxima of the KL2 metric correspond to the speaker change points [119] (Siegler et al., 1997).

8.2.4 Speaker Clustering

As pointed out before, a speaker segmentation system aims to determine if two adjacent segments belong to the same speakers. While, for a speaker clustering system, the goal is to group the segments that seem to be pronounced by the same speaker. Unlike the speaker segmentation, these segments could be localized anywhere in the audio signal. The problem of measuring a distance between segments for the speaker clustering remains the same. Therefore, all the distance presented in section 8.3.3 are also used for speaker clustering.

As shown in figure 8.1, the optimal output of the speaker clustering system is a single cluster for each speaker. Since the number of clusters and the speakers identities are unknown, the speaker clustering process is considered as an unsupervised classification problem that is commonly solved with hierarchical clustering.

The hierarchical clustering is applied to iteratively agglutinate together a set of elements that belong to the same class. Figure 8.3 shows the two mostly used methods to perform hierarchical clustering: top-down clustering and bottom-up clustering.

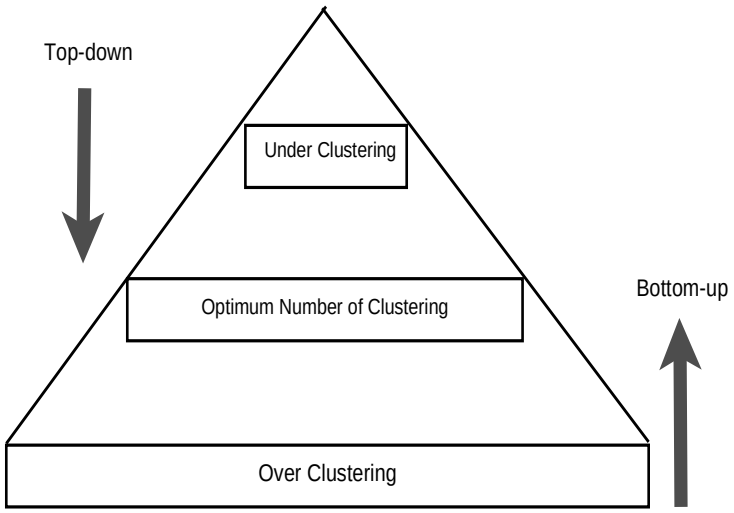


Figure 8.3: Hierarchical bottom-up or top-down clustering.

The bottom-up clustering method, also known as agglomerative hierarchical clustering, is by far the most common in the literature. In this method, each segment provided by the segmentation step is considered as a single cluster. Then a merging procedure is applied iteratively to reach the optimal number of clusters. On the other hand, the top-down clustering technique assigns the whole data to a single cluster. Then a splitting procedure is performed in order to obtain the optimum number of clusters.

The next subsections deal with the main systems for speaker clustering. These systems are developed for an offline configuration where the whole audio document is available. However, some of these systems are adapted for an online scenario where only the streamed data is available.

8.2.4.1 BIC-based Clustering Approach

As previously mentioned, all the distances applied to detect speaker changes remain useful for speaker clustering. The most commonly found distance is Bayesian Information Criterion [89] (Moraru et al., 2005). As shown in [124] (Tranter and Reynolds, 2006), the Agglomerative Hierarchical Clustering using a BIC-based distance is composed of the

following steps:

1. Each segment provided by the speaker change detection step is assigned to an independent cluster.
2. Compute the pair-wise distance matrix between each cluster using the BIC metric.
3. Merge the pair with the lowest ΔBIC distance.
4. Update the pair-wise distance between the remaining clusters and the merged one.
5. Iterate from step 2 to 4 until all pairs have a $\Delta BIC > 0$.

8.2.4.2 Hidden Markov Model Approach

A HMM-based approach for speaker clustering is introduced in [2] (Ajmera et al., 2002). An ergodic HMM model is proposed, where each state represents a cluster with a Gaussian distribution and the transitions matrix reflect the changes between speakers. The clustering process is performed by merging the closest clusters in terms of the log-likelihood ratio distance. Then the HMM are re-trained according to the new topology (one less state) and the overall likelihood is computed. When this likelihood decreases, the merging process is stopped.

An alternative HMM approach, called Evolutive-HMM training, is proposed in [87] (Meignier el al., 2001) and [46] (Fredouille et al., 2009). This method belongs to the top-down clustering category, where the segmentation and the clustering module are unified in a common step. First the whole speech file is assigned to a single speaker and modeled by a 1-state HMM model. Then a new state, representing a new speaker, is estimated from few feature vectors that maximize the likelihood ratios of the initial model. After that, each cluster (represented by a HMM state) is iteratively adapted and a Viterbi algorithm is performed to obtain the new segmentation. The stop criterion is reached when the recognition likelihood of the iteration $m-1$ is greater than the one of the iteration m.

8.2.4.3 Cross Likelihood Ratio Approach

The Cross Likelihood Ratio (CLR) method is generally performed after a first clustering based on the BIC metric [140] (Zhu et al., 2005). In the BIC clustering step, the acoustic features are not normalized in order to keep the background information which are useful to differentiate between speakers. However these information could lead to several clusters belonging to the same speaker. Therefore, a second step of Cross Likelihood Radio clustering is performed to overcome this problem.

In the CLR clustering step, the background environment effects are removed by normalizing the acoustic features. Then an universal background model is trained and adapted to each cluster provided by the BIC clustering step. After that, a pair-wise distance matrix is computed using the CLR metric and the closest clusters are merged.

The CLR metric is expressed as follows:

$$CLR(C_1, C_2) = \frac{1}{N_1} \log\left(\frac{L(C_1, M_2)}{L(C_1, UBM)}\right) \times \frac{1}{N_2} \log\left(\frac{L(C_2, M_1)}{L(C_2, UBM)}\right) \quad (8.14)$$

where N_1 and N_2 are the sizes of the clusters C_1 and C_2. M_1 and M 2 are respectively the adapted models of the clusters C_1 and C_2 and L(.) is the likelihood function. The clustering stops when $CLR(C_1, C_2)$ gets higher than a predefined threshold.

8.2.5 Recent Research Directions

In this section, a recent research direction for speaker diarization are described. These works did not confirm their robustness yet, but they show a considered potential to improve the diarization performances. Two directions are explored in this section: the use of prosodic information and the overlapping speech detection.

8.2.5.1 Prosodic Information Exploitation

For many years, speaker diarization systems were based on cepstral features, such as MFCC, LFCC, etc., to represent the audio signal. However some works propose to exploit the prosodic information to improve the performance of the diarization process.

In [47] (Friedland et al., 2009), a framework to study the effects of 70 different long-term features is described. These features belong to five different categories: pitch, energy, formants, harmonics-to-noise ratio, and long-term average spectrum. The authors demonstrate that by combining the 10 top ranked features , in terms of speaker discriminability, with the MFCC features the diarization performance increases dramatically.

8.2.5.2 Overlapping Speech Detection

Speech overlaps involve audio segments where simultaneous speakers are active. For many years, speaker diarization systems were designed for audio contents where speech overlaps are rare. However, it was shown in [63] ((Huijbregts et al., 2012)) that overlapping speech is one of the main source that decreases the performance of speaker diarization system. In fact, assigning an overlapping segment to a particular speaker could perturb the modeling of its cluster. Moreover, when an overlapping segment is missed by the diarization system, the error is accounted twice. Many works propose to detect overlapping speech in order to improve the performances of the speaker diarization. The main approaches are based on HMM/GMM modeling of overlapping and non-overlapping speech.

In [27] (Charlet et al., 2013), two systems are introduced. The first one combines a Gaussian Mixture Model classification system and a multi-pitch features detection approach. Three classes are considered: non-speech, non-overlapping speech and overlapping speech using the Perceptual Linear Predictive features. The multi-speech detection method is based on the pitch estimation algorithm proposed in [34] (De Cheveigné, 2006). In the second system three Gaussian models are trained, representing male non-overlapping speech, female non-overlapping speech and overlapping speech. Then, the obtained models are used to build a 2-class HMMs for overlapping and no-overlapping speech.

Two strategies are proposed to efficiently handle the overlapping speech information in a speaker diarization system:

- Discard the detected overlapping speech from the diarization process.
- Assign the overlapping speech segments to the two temporal closest speakers.

An oracle studies in [63] (Huijbregts et al., 2012) [120] (Sinclair and King, 2013) shows that by assigning an overlapping segment to at least one right speaker the error is halved and if the labeled second speaker is correct the error is halved again.

8.3 The ALISP-based Speaker Diarization System

The main goal of this thesis is to identify the majority of audio items that could be found in a radio broadcast streams using data-driven ALISP segmentation. In the previous chapters, the case of music and advertisement was treated and good performances of the ALISP-based system were showed. In this section, the generic audio indexing system is applied to speaker diarization.

As showed in the previous section, the most systems of speaker diarization involve the acoustic features extraction, the speech activity detection and the speaker segmentation and clustering. A new module based on ALISP tools is proposed at the top of the chain.

In this work, we are interested in speaker diarization for TV and radio shows which include various acoustic sources such as studio/telephone speech, music, or speech over music. Usually these programs tend to keep the same structure with same presenters, reporters, sound effects, jingles, etc. This redundancy is used in order to improve the performance of the speaker diarization system.

The main idea of our system is to compare the show to be segmented with the same show broadcasted before in order to find the common audio parts, represented by speech sentences, silence, noise, jingles, music and advertisements. This operation is performed via audio fingerprinting which involves the extraction of the ALISP symbols, which constitutes a compact audio fingerprint, for each audio document stored in a reference database. An unlabeled test audio excerpt is identified by comparing its ALISP fingerprint with those of the reference database using our approximate matching of ALISP units. Then, these common segments are labeled according to their nature and the output pre-labeled signal is processed with a speech activity detection, GLR-BIC speaker segmentation, BIC clustering, Viterbi refinement and Normalized Cross Likelihood Ratio (NCLR) clustering.

This section is organized as follows. In the next part, the general architecture of the

proposed ALISP-based speaker diarization system is presented. Then each module of the proposed system is described individually.

8.3.1 System Architecture

The general architecture of the proposed system is illustrated in figure 8.4. The system is composed of the following steps:

1. ALISP-based audio sequencing and identification.
2. Voice activity detection.
3. GLR-BIC segmentation.
4. BIC clustering.
5. Viterbi refinement.
6. Normalized Cross Likelihood Ratio clustering.

The principle module in this architecture is ALISP-based audio sequencing and identification. The main contribution of this module is to help the diarization process by labeling the audio parts that were broadcasted before.

8.3.2 ALISP-based Audio Sequencing and Identification

The proposed system uses the transcriptions provided by ALISP tools to search for recurrent segments in TV and radio shows. As a reminder, the generic audio indexing system consists of three main modules: ALISP unit acquisition and modeling, ALISP transcription and approximate matching to find recurrent segments. The set of ALISP models is automatically acquired through parameterization, spectral stability segmentation, vector quantization, and Hidden Markov Modeling. This set of HMM ALISP models is used to transform a new incoming audio data into a sequence of ALISP symbols. And the approximate string matching algorithm is based on the Basic Local Alignment Search Tool (BLAST) [3] (Altschul et al., 1990), widely used in bioinformatics.

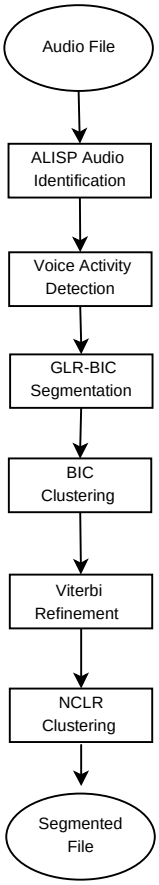


Figure 8.4: General architecture of the proposed ALISP-based system.

Figure 8.5: Example of an output file provided by ALISP-based audio sequencing and identification.

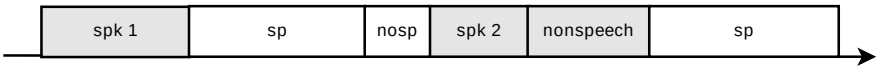


Figure 8.6: Example of an output file provided by the voice activity detection system.

This audio indexing system is applied on speaker diarization as follows. First, a reference database is built from audio parts provided from previously annotated emissions. These parts represent speech sentence excerpts, silence, noise, jingles, music and advertisements. Then, the ALISP module for transcriptions of reference segments are computed using ALISP HMM models and compared to the transcriptions of the TV and radio shows stream using our approximate matching module.

This module could be seen as a pre-processing step that helps the diarization process. In fact, instead of treating an unlabeled data, the proposed system is supposed to identify all the audio segments that were streamed before. This will provide a first segmentation of the audio signal with three types of labels:

- "spk" label: represents a sentence or an excerpt of sentences of a particular speaker that was seen before. These segments are, generally, relative to TV and radio presenters, reporters, politicians, artists, etc.
- "non-speech" label: represents the non-speech audio segments which could be: noise, silence, advertisement, jingles and music.
- "unknown" label: represents the signal parts that are not recognized by the ALISP module. These parts will be treated by the speech detection module and the GLR-BIC segmentation modules.

An example of the output file provided by the ALISP module is shown in figure 8.5. The "spk" label is related to a speech sentence detected in the reference database, while the "unknown" label is relative to the signal parts which are not detected in the reference database.

It is important to note that ALISP tools are speaker-dependent. The ALISP transcriptions of identical sentences spoken by different speakers are very different, while the ALISP transcriptions of identical sentences spoken by the same speaker are very similar.

Three main contributions of the ALISP-based audio sequencing and identification module are proposed to improve the performances of the speaker diarization system:

- Discarding the non-speech segments limits the errors caused by the false alarms and missed speech detection.
- Assigning an audio segment to a single speaker improves the purification of their models in the clustering step.
- When dealing with a long audio file (such as one day of a radio broadcast), the processing time is reduced using the approximate matching process.

8.3.3 Speech Activity Detection

The next step in the system is the voice activity detection. The goal is to remove the non-speech segments whose duration is above a predefined threshold. Our voice activity detection system operates only on the portions of the signal labeled as "unknown" by the ALISP recognizer. It relies on a two-class detector, with Gaussian Mixture Model trained on speech and non-speech data. The parameterization is done with MFCCs, calculated on 20 ms windows, with a 10 ms shift. For each frame, a cepstral vector of dimension 12 is computed and appended with first and second order deltas and the Zero Cross Ratio. A minimum duration of 0.5 s is defined for speech and non-speech segments. In fact each class is modeled as a concatenation of 50 one-state HMM models.

An example of the output file provided by the voice activity detection module is shown in figure 8.6. The "nosp" label is relative to a non-speech segment, while the "sp" is relative

to speech segment.

8.3.4 GLR-BIC Segmentation

The GLR-BIC segmentation is a two-step algorithm that consists of a first pass to determine the speaker change point candidates using the GLR criteria and a second pass based on a BIC distance to validate or discard these candidates.

This step is only performed on signal parts that were labeled as "sp" by the voice activity detection module. On the other hand, the segments that were identified by the ALISP module and labeled as "spk" are not processed by this module.

The GLR-BIC segmentation consists of two main steps:

1. GLR segmentation: The audio signal is split into equal size sliding windows. Then the speaker change candidates are determined for each window. These candidates correspond to the local maxima of the GLR measure. Therefore there is no need to define a threshold as for the conventional GLR segmentation described in section 8.2.3.1.

2. BIC segmentation: In the previous step, all the local maxima of GLR criterion are considered as speaker change points, which leads to many false alarms. The BIC measure is used to validate the real change points and to discard the false alarms. For each candidate, the ΔBIC distance is computed between the Gaussian distributions of the two adjacent windows. If the maximum of the ΔBIC is positive the change point is confirmed, otherwise the two segments are merged.

8.3.5 BIC Clustering

Whereas the BIC segmentation operates on neighboring segments in order to detect whether or not they correspond to the same speaker, BIC clustering is performed to group together all the segments that belong to the same speaker. As for the segmentation process, at each iteration the closest clusters are merged until $\Delta BIC > 0$. At this point, all the labeled segments as "spk" whether by the ALISP module or by the GLR-BIC segmentation module are processed by the BIC clustering.

8.3.6 Viterbi Refinement

The cluster boundaries produced by the BIC clustering are not perfect. Thus, a Viterbi decoding is performed to adjust these boundaries. Each cluster is modeled by a single-state HMM with an 8-component GMM trained using the EM algorithm. The speaker change points are represented by the transitions between HMMs.

8.3.7 NCLR Clustering

As was mentioned in section 8.2.4.3, a final step of clustering is performed in order to remove the background environment effects. The MFCCs features are normalized using the warping technique. Then, due to the great length of clusters, a robust speaker model is modeled using an universal background model. Unlike the system described in section 8.2.4.3, a normalized version of the CLR metric is used [74] (Le, 2007). This metric demonstrates a better performance than its original version. The Normalized Cross Likelihood Ratio (NCLR) is given by:

$$NCLR(C_1, C_2) = \frac{1}{N_1} \log\left(\frac{L(C_1, M_1)}{L(C_1, M_2)}\right) \times \frac{1}{N_2} \log\left(\frac{L(C_2, M_2)}{L(C_2, M_1)}\right) \quad (8.15)$$

Where N_1 and N_2 are the sizes of the clusters C_1 and C_2. M_1 and M 2 are respectively the GMM adapted models of the clusters C_1 and C_2 and L(.) is the likelihood function. The clustering stops when $NCLR(C_1, C_2)$ gets higher than a predefined threshold. At the end of this step, the final diarization of the audio file is provided. The next section deals with the experimental results obtained for the proposed system.

8.4 Experiments and Results

In this section, the contributions of the ALISP-based module to speaker diarization are evaluated. Two main evaluations are carried out. The first one deals with the ETAPE (Evaluations en Traitement Automatique de la ParolE) evaluation campaign 2011 in order to measure our contributions within a publicly available framework. While in the second

evaluation, the ALISP-based speaker diarization system is combined with a speaker verification system to measure the speech time of politicians in radio streams.

8.4.1 ETAPE Evaluation Campaign

ETAPE is an evaluation campaigns for automatic speech processing [55] (Gravier el al., 2012). It was held in spring 2012 and considered four tasks:

1. Multiple speaker detection: It is the task of overlapping speech detection, where the goal is to provide for each audio file the start and end times of segments containing speech from multiple speech
2. Speaker turn segmentation: It is the speaker diarization task. Two subtasks are considered, depending on whether the diarization process is performed on each audio file independently or on all audio files together.
3. Lexical transcription: It is related to the automatic speech recognition. The system should provide a start and end times for each word associated with its speaker.
4. Named entity detection: It consists in detecting all direct mentions of named entities and in categorizing the entity type.

This works addresses the speaker diarization task where each audio file is processed independently. The ETAPE evaluation campaign targets the TV and radio shows with various level of spontaneous speech and multiple speaker speech. Unlike the ESTER evaluation campaigns [48] (Galliano et al., 2005) and [49] (Galliano et al., 2009) the ETAPE evaluation set did not focus in a particular type of show.

8.4.1.1 Corpus

The ETAPE 2011 evaluation campaign provides the participants with 13.5 hours of radio data and 29 hours of TV data. This corpus was selected to contain spontaneous speech and a reasonable proportion of multiple speaker data. A detailed description of this corpus is given in section 4.3

8.4.1.2 Evaluation Measure

In order to evaluate the speaker diarization system, the Diarization Error Rate (DER) is used. The DER is the sum of three errors: missed detection rate, false alarm rate and the speaker error rate:

- The missed detection rate is expressed as:

$$MD = \frac{\sum_{s=1}^{S} dur(s) \times (N_{Ref}(s) - N_{Sys}(s))}{\sum_{s=1}^{S} dur(s) \times N_{Ref}(s)} \tag{8.16}$$

- The false alarm rate is given by:

$$FA = \frac{\sum_{s=1}^{S} dur(s) \times (N_{Sys}(s) - N_{Ref}(s))}{\sum_{s=1}^{S} dur(s) \times N_{Ref}(s)} \tag{8.17}$$

- The speaker error rate is computed as follows:

$$SER = \frac{\sum_{s=1}^{S} dur(s) \times (\min(N_{Ref}(s), N_{Sys}(s)) - N_{Correct}(s))}{\sum_{s=1}^{S} dur(s) \times N_{Ref}(s)} \tag{8.18}$$

where S is the total number of speaker segments, dur(s) denotes the duration of speaker s. $N_{Ref}(s)$ and $N_{Sys}(s)$ indicate the number of speakers present in segment s provided, respectively, by the ground truth and the diarization system. $N_{Correct}(s)$ is the number of speakers in segment s that have been correctly matched between the ground truth and the proposed system. The DER is obtained by a one-to-one mapping of all the labeled speakers between the system and reference files. It could directly be computed as follows:

$$SER = \frac{\sum_{s=1}^{S} dur(s) \times (\max(N_{Ref}(s), N_{Sys}(s)) - N_{Correct}(s))}{\sum_{s=1}^{S} dur(s) \times N_{Ref}(s)} \tag{8.19}$$

8.4.1.3 Threshold Setting

The proposed speaker diarization system contains four thresholds values which need to be fixed. These thresholds are related to the Levenshtein distance, the BIC segmentation, the BIC clustering and the NCLR clustering.

	# files	Avg spk	Avg turn duration (sec)	% silence	% ovlp
BFMTV	1	21	10	32	2
LCP	6	7.8	2	28	4
TV8	2	9	8	22	7
EST2BC	6	12.5	5	35	3

Table 8.1: Number of audio files (# files), average number of speaker (Avg spk), average duration of turns in seconds (Avg turn duration), percentage of silence (% silence) and the percentage of overlapping speech (% ovlp) of the evaluation corpus.

As explained in section 6.3, some experiences are conducted in order to fix the Levenshtein distance threshold in the context of audio identification where the goal is to identify advertisements and songs in radio streams. These experiences consist of computing the Levenshtein distance between ALISP transcriptions of the reference advertisements and their broadcasted occurrences in the radios and between ALISP transcriptions of the reference advertisements and data that does not contain advertisements. This study leads to a Levenshtein distance threshold of 0.55.

In order to fix the other three thresholds, an automatically tuning, by trying various combinations of thresholds, is performed on the ETAPE development corpus. Each generated segmentation is scored against the reference segmentation and the thresholds that gave the lowest DER are chosen in the evaluation.

8.4.1.4 Results

The evaluation dataset provided by ETAPE is composed of 9 TV shows and 6 radio shows. Table 8.1 gives some statistics about the evaluation corpus.

BFMTV, LCP and TV8 are relative to TV shows while EST2BC is relative to radio shows. In addition, table 8.1 shows the diversity of the evaluation corpus which make the task of speaker diarization more complicated. In order to evaluate the contributions of the ALISP-based module to the diarization results, a second experience is performed without that module.

Table 8.2 gives the DER values for the baseline system (without the ALISP module) and the ALISP-based system.

Show name	Baseline	ALISP
BFMTV-BFMStory-175900	19.30	15.87 (-17.77%)
LCP-CaVousRegarde-235900	20.70	12.60 (-39.13%)
LCP-EntreLesLignes-192800-1	24.77	17.31 (-30.11%)
LCP-EntreLesLignes-192800-2	27.19	18.48 (-32.03%)
LCP-PilesEtFace-192800	28.42	19.76 (-30.04%)
LCP-TopQuestions-000400	35.46	29.55 (-16.66%)
LCP-TopQuestions-213800	15.87	2.44 (-84.62%)
TV8-LaPlaceDuVillage-201300	37.86	22.27 (-41.22%)
TV8-LaPlaceDuVillage-172800	35.82	20.40 (-43.04%)
EST2BC-FRE-FR-1000	14.55	13.75 (-5.49%)
EST2BC-FRE-FR-1750	39.41	22.93 (-41.81%)
EST2BC-FRE-FR-2152-1	41.83	27.34 (-34.64%)
EST2BC-FRE-FR-2152-2	29.91	23.93 (-19.99%)
EST2BC-FRE-FR-0910	8.73	8.26 (-5.38%)
EST2BC-FRE-FR-2004	21.13	15.48 (-26.73%)
ETAPE-2011 (whole data)	24.73	16.23 (-34.37%)

Table 8.2: Diarization Error Rate for the baseline and ALISP system on the ETAPE 2011 evaluation set.

Note that the ALISP-based module improves the diarization results for all TV and radio shows. However, these improvements are not significant for all audio files. For "LCP-TopQuestions-213800" TV show the relative improvement of the DER is 84.62% while for the "EST2BC-FRE-FR-0910" radio show it is only is 5.38%. This is essentially related to the structure of the radio or TV show, and whether there are repeating audio parts that can be detected by the ALISP-based module.

The main contribution of the ALISP module is essentially the purification of the clusters, which leads to more robust speaker models. Moreover, the ALISP method is able to detect recurrent audio excerpts such as commercials and jingles, decreasing the missed detection rate and the false alarms. Overall, the introduction of the ALISP module in the speaker diarization system has relatively decreased the DER by 34.37%, while the absolute improvement is 8.5%.

Since the proposed system did not deal with overlapping speech, many errors have occurred especially in TV8 shows and radio shows (EST2BC). By using the ground truth to label the overlapping speech segments, the DER decreases from 16.23% to 12.02%.

Participant	DER
Our system	16.23
System 1	19.01
System 2	21.18
system 3	22.45
system 4	22.73
System 5	27.27
system 6	29.32

Table 8.3: Diarization Error Rate for the all the participants in the ETAPE 2011 evaluation campaign.

The global DER value for each submitted system are presented in table 8.5. Seven participants have submitted results for the speaker diarization task in ETAPE 2011, which are:

- Institut Mine Télécom-Télécom ParisTech-Télécom SudParis (Our System)
- Centre de Recherche Informatique de Montréal (CRIM) [56] (Gupta et al., 2008)
- Eurecom [20] (bozonnet et al., 2010)
- Laboratoire d'Informatique de l'Université du Maine (LIUM) [113] (Rouvier and Meignier, 2012)
- Laboratoire Informatique d'Avignon (LIA) [87] (Meignier et al., 2001)
- Orange Labs [27] (Charlet et al., 2013)
- Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) [13] (Barras et al., 2006)

As shown in the table 8.3, the proposed ALISP-based speaker diarization system has obtained the best results in the ETAPE 2011 evaluation campaign among 7 participants. These results attest that the exploitation of the common structure of the radio and TV shows by the ALISP techniques, leads to great improvements of the speaker diarization process.

Related to the processing time, the system without the ALISP-based module runs at a speed of 10 seconds per minute on a 3.00GHz Intel Core 2 Duo 4GB RAM. When the ALISP-based module is added, the runtime increased to 40 seconds per minute of speech processed.

8.4.2 Speech Time Measure of Politicians

The task of measuring the speech time of politicians involves two disciplines of speaker-based processing: speaker identification and speaker diarization. Given one day of radio stream, the goal is to identify all the politicians and to measure the time of their speech. In order to achieve this goal, first the ALISP-based speaker diarization system is applied to segment the audio data into homogenous clusters according to speaker's identities. Then, a speaker identification system is performed to determine whether a cluster belongs to a politician.

The speaker identification system is an UBM-GMM system [111]. The Gaussian Mixture Model-GMM approach is used to build models from the speaker data. An Universal gender-dependent Background Model-UBM is trained with the Expectation-Maximization algorithm. Then, each speaker model is built by adapting the parameters of the UBM using the speaker's training feature vectors and the Maximum A Posteriori criterion. The similarity score is the estimation of the log-likelihood ratio between the target (politician) and UBM model.

We use the open source speaker verification system described in [103] (Petrovska-Delacrétaz et al., 2009) and available at [40] (ElHannani et al., 2009). This system was originally developed for speaker verification and adapted for speaker identification. In fact, the major difference between verification and identification lies in the decision process. In verification, the decision is accepting or rejecting the identity claim of a speaker. In identification, the goal is to determine which registered speaker provides a given utterance. Thus, the same algorithms and techniques are used for speaker verification and speaker identification.

The performances of the speaker identification system are evaluated during the com-

petition on speaker recognition in mobile environment using the MOBIO database [85] (McCool et al., 2012). Moreover, a second evaluation is carried out on the YACAST database to measure the speech time of politicians in radio streams. Unlike the first evaluation, the second one involves the speaker diarization and identification systems.

8.4.2.1 MOBIO Evaluation Campaign

Following the same spirit as the NIST SRE, the Biometric Group at the Idiap Research Institute organized the evaluation on text independent speaker recognition. It is performed on the MOBIO database, which consists of videos of talking faces that were filmed with mobile devices. A detailed description of the MOBIO database is given in section 4.4.

The proposed primary system is an UBM-GMM system. It is based on the reproducible BioSecure Speaker reference system described in [103] (Petrovska-Delacrétaz et al., 2009) and available at [40] (ElHannani et al., 2009). The main parameters of the proposed system are: 32 MFCC coefficients + deltas + delta energy, energy based voice activity detector, feature warping normalization and 512 Gaussians. The particularity of the system is to join the MOBIO training dataset and the Voxforge [1] dataset to train the UBM model.

Two additional systems were also submitted in this evaluation. The secondary system has the same configuration as the primary one, except that the UMB-GMM training was performed only MOBIO training data. Its performances are slightly worse than the primary system. For the third submission, the sampling frequency as fixed to 8 Khz and the NIST 2003-04 along with MOBIO training data are used to build the UBM model. This configuration changes do not seem to degrade drastically the performances of the system.

In total, 12 institutions participated in the speaker verification evaluation, and provided 21 valid submissions (12 primary and 9 secondary submissions). These institutions are illustrated in table 8.5.

In order to evaluate the performances of the speaker verification systems, two measure are used: Equal Error Rate (ERR) and Half Total Error Rate (HTER).

Table 8.5 shows the EER on the development set and the HTER on the evaluation

[1] http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/16kHz_16bit

Institution	System Identifier
Alpineon Ltd., Slovenia	Alpineon
ATVS Biometric Recognition Group Universidad Autónoma de Madrid, Spain	ATVS
Centre de Développement des Technologies Avancées, Algeria	CDTA
CpqD, Brazil	CPqD
GIAPSI, Universidad Politécnica de Madrid, Spain	GIAPSI
GTTS - University of Basque Country (UPV/EHU), Spain	EHU
Idiap Research Institute, Switzerland	IDIAP
L2F/INESC-ID, Portugal	L2F
Joint submission of L2F/INESC-ID and UPV/EHU	L2F-EHU
Institut Mines-Télécom (Télécom ParisTech-Télécom SudParis), France	Mines-Telecom
Phonexia s.r.o. , Czech Republic	Phonexia
Radboud University Nijmegen, The Netherlands	RUN

Table 8.4: The institutions and the identifiers of their submitted primary system (by alphabetic order).

set for both genders.

On the evaluation set, our proposed system obtains the best simple system performance on Female. Obviously, the use of additional suitable data (Voxforge database) for training the UBM is helpful. Additional experiment that combines NIST SRE data (03 and 04) and MOBIO data are carried out to train the UBM model. The EER on the DEV set are 14.80% for Female and 13.62% for Male, respectively.

8.4.2.2 YACAST Evaluation

As previously mentioned, the measuring of speech time of politicians is divided into two subtasks: speaker diarization and speaker identification. YACAST database contains the record of 26 days of radio streams from three different French radio: France Culture France Info France Inter. This database contains 283 politicians with a total duration of 42h46min. To ensure a good training of models, a politician is considered as a target if he spoke more than 10 minutes which leads to a set of 72 target speakers.

In order to ensure an objective evaluation of the speaker diarization and identification systems, the YACAST database was divided into 5 subsets:

System	Female		Male	
	DEV	EVAL	DEV	EVAL
Alpineon*	7.982	10.678	5.040	7.076
ATVS	16.836	17.858	14.881	15.429
CPqD*	14.348	15.987	11.824	10.214
CDTA	19.471	22.640	12.738	19.404
GIAPSI	11.590	12.813	9.683	8.865
EHU	17.937	19.511	11.310	10.058
IDIAP	12.011	14.269	9.960	10.032
L2F*	13.484	22.140	10.599	11.129
L2F-EHU*	11.005	17.266	7.889	8.191
Our System	11.429	11.633	10.198	9.109
Phonexia	8.364	14.181	9.601	10.779
RUN	25.405	23.112	24.643	22.524

Table 8.5: Equal error rate (EER %) on the development (DEV) set and half total error rate (HTER %) on the MOBIO evaluation (EVAL) set.

- Training corpus: It is used to train the UBM model used for both speaker diarization and speaker identification. It contains speech segments of non-target speakers. The total duration of the training set is 7h28min. The number of speakers in this corpus is 182.
- Adaptation corpus: It serves to adapt the UBM model to the target speakers. This corpus contains 1min30s speech for each of the 72 target speaker.
- Evaluation corpus: Two days of each radio are chosen to evaluate the systems. These two days contain the maximum time of the target speakers.
- Development corpus I: It is used to set the thresholds of different metrics of the speaker diarization system. For this corpus, two days of each radio are selected.
- Development corpus II: It is used to set the threshold of the speaker identification system. The average length of speech for each target speaker in this corpus is 4min.

The estimation of the speech time of politicians is evaluated on two levels. The first level is relative to speaker diarization and uses the DER as described in the section 8.4.1.2. The second level is related to the speaker identification. The performances of the speaker identification system is evaluated as the sum of three errors:

Radio Day	DER
France-Inter-30-06-2010	18.98
France-Inter-21-04-2010	16.03
France-Info-29-06-2010	18.33
France-Info-21-04-2010	18.12
France-Culture-27-06-2010	13.11
France-Culture-20-04-2010	16.32
All data	17.01

Table 8.6: Diarization Error Rate for each day of the YACAST evaluation corpus.

- Substitution error (E_{Sub}): It occurs when the system assigns a speech segment to a target speaker X when it is pronounced by a different target speaker Y.
- False alarm (E_{FA}): It is the error due to the detection of a target speaker segment when it really belongs to the non-target speaker set.
- False rejection (E_{FR}): It is the error due to the detection of a nontarget speaker segment when it really belongs to the target speaker set.

Table 8.6 shows the DER values for each day of the radio stream in the evaluation set. It indicates an overall DER of 17.01%, close to that obtained in ETAPE evaluation campaign, which proves the robustness of the proposed system. As for the audio identification evaluation described in section 6.4.2, many annotation errors are found in the ground truth. These errors are related to segments' boundaries that were not precise and to some confusions in the labels (speaker names) of speech segments.

Table 8.7 reports the performances of the speaker identification system in terms of substitution errors, false alarms and false rejections. This table reports an overall error rate of 22.55%. These errors are essentially due to the errors caused by the diarization process. In fact, by using a perfect diarization system (DER= 0%), the global error rate decreases to 13.25%. Moreover, as for the speaker diarization system, the errors found in the ground truth have a direct impact on the performances of the speaker identification system. In fact these errors lead to impure speaker models, which causes some degradations in the results of the speaker identification system.

Radio day	E_{sub}(%)	E_{FA}(%)	E_{FR}(%)	Global(%)
France-Inter-30-06-2010	10.23	9.56	11.22	31.01
France-Inter-21-04-2010	6.33	9.02	10.33	25.68
France-Info-29-06-2010	2.88	2.9	6.78	12.56
France-Info-21-04-2010	13.74	7.22	6.51	27.47
France-Culture-27-06-2010	9.63	4.51	7.28	21.42
France-Culture-20-04-2010	4.99	11.11	1.08	17.18
All data (Mean)	7.97	7.38	7.2	22.55

Table 8.7: Substitution error (E_{sub}), false alarm (E_{FA}) and false rejection (E_{FR}) for the speaker identification system computed on the YACAST evaluation corpus.

Once the speaker diarization and identification process are performed, the speech duration of politicians is measured. The ratio between the speech duration detected by the proposed system and the speech duration extracted from the ground truth is 79%. This result can be improved by correcting the ground to purify the speaker models.

8.5 Conclusion

In this chapter, the state of the art of speaker diarization is reviewed. Speaker diarization process is generally composed of speech activity detection, speaker segmentation and speaker clustering. A new module based on data-driven segmentation using ALISP techniques is added in order to improve the performance of the diarization process. This module compares the show to be segmented with the same show broadcasted before in order to find the common audio parts.

The system is evaluated during the ETAPE 2011 evaluation campaign and obtained a DER of 16.23%, which is the best result among all participants. We also demonstrate that by adding the ALISP module to the speaker diarization system the DER decreased by 8.5%.

A second evaluation relative to the estimation of the speech time of politicians is performed. First, the speaker recognition system is evaluated during the MOBIO 2013 evaluation campaign an obtained a HTER of 11.633% for female and 9.109% for male speakers. Then the speaker diarization and identification systems are evaluated using the

YACAST database. For the speaker diarization system the obtained DER value is 17.01%. While the global error rate for the speaker identification system is 22.55%.

In the next chapter, a different category of audio events, denoted as nonlinguistic vocalization, will be studied. The generic audio indexing system will be applied to laughter detection.

Chapter 9

Nonlinguistic Vocalizations Detection

9.1 Introduction

As pointed before, one of the contributions of this thesis is to identify the majority of audio items that could be present in a radio broadcast streams using the same audio indexing principles. In previous chapters, the ALISP-based audio indexing system was applied to audio identification and audio motif discovery, to detect songs and advertisements in radio streams, and speaker diarization to segment the audio data into homogeneous segments according to speaker identities. In this chapter, a different category of audio events will be studied, which is referred as nonlinguistic vocalization.

Despite the best efforts made over past two decades in speech recognition systems, detection of nonlinguistic vocalizations such as laughter, sighs, breathing, hesitation sounds is still a challenging task [135] (Weninger et al., 2011). Such vocalizations are more frequent in radio and TV shows, meetings or our daily conversational speech.

Detection of the presence of these vocalizations is useful in several disciplines. In Automatic Speech Recognition the detection of nonlinguistic vocalizations could give relevant information to decide which parts of audio data should be treated for recognition, thereby improving the performance of speech recognition systems. Traditional speech recognition

frameworks have not been adequately focused on detecting nonlinguistic vocalizations under a common and generic framework. One of the main reasons could be the complexity behind obtaining phonetic representations or a pronunciation dictionaries (i.e. phonetic lexicon) for such vocalizations.

One of the most obvious nonlinguistic sounds is laughter. Laughter is one of the complex nonlinguistic vocalizations that communicates a wide range of messages with different meanings [23] (Campbell et al., 2005). Moreover, it was shown in [9] (Bachorowski et al., 2001) [126] (Trouvain, 2003), that laugher sound is a highly variable signal whose characteristics are not yet revealed.

In this thesis, the use of ALISP-based indexing framework is proposed to detect nonlinguistic vocalizations. Given the high variability of this category of sounds both between and within speakers, we decided to use a different approach to search for these sounds from the one used in the previous systems. Our method first adapts ALISP models, previously trained on 288 hours of radio broadcast, using Maximum Likelihood Linear Regression-MLLR [75] (Leggetter and Woodland, 1995) and Maximum A Posterior-MAP [50] (Gauvain and Lee, 1994) techniques. The resulting adapted models can then be used to detect local regions of nonlinguistic vocalizations, using the standard Viterbi algorithm [137] (Young et al., 1989). Experiments on a laughter-annotated audio corpus show the usefulness of the proposed method.

This chapter is organized as follows. Section 9.2 presents a literature of nonlinguistic vocalizations sound detection. Then, the proposed methodology to detect any type of nonlinguistic vocalizations is explained in section 9.3. In Section 9.4, empirical evaluation of the proposed method, on an laughter-annotated corpus, is exposed.

9.2 Related Work

Most of the previous efforts on automatic laughter detection from audio exploit frame level acoustic features as parameters to train machine learning techniques, such as Gaussian Mixture Models and Support Vector Machines. These systems are composed of two steps: feature extraction and modeling.

9.2.1 Feature Extraction

Two categories of feature are used to represent the laughter signal: frame-level and utterance-level features.

Frame-level features refer to features extracted from each frame of the audio signal, which leads to a variable length of feature vector that depends on the length of the processed audio file. The most popular feature belonging to this category is MFCC. These features are used in [69] (Kennedy and Ellis, 2004) [73] (Laskowski and Schultz, 2008) to represent the laughter signal. Moreover, Perceptual Linear Prediction Coding features [62] (Hermansky, 1990) are exploited to model the spectral properties of laughter [127] (Truong and Van Leeuwen, 2005) [128] (Truong and Van Leeuwen, 2007). Prosodic information are also used to discriminate between laughter and no laughter sounds. In [9] (Bachorowski et al., 2001), it is found that the mean pitch for laughter is considerably higher than in speech. Therefore, pitch feature, associated with energy are used in many systems to locate regions of laughter in audio files [71] (Knox and Mirghafori, 2007) [127] (Truong and Van Leeuwen, 2005).

Utterance-level features are relative to global features computed on the whole utterance, which leads to fixed length feature vectors. In addition to the pitch computed for each frame, some authors propose to extract the pitch from the whole sentence and compute some statistics such as, the standard deviation, the mean, or the maximum and the minimum [128] (Truong and Van Leeuwen, 2007). Furthermore, it is shown in [17] (Bickley and Hunnicutt, 1992), that the ratio between unvoiced and voiced frames is higher for laughter than for speech. Thus, statistics such as the number of unvoiced frames divided by the number of total frames are introduced in [129] (Truong and Van Leeuwen, 2007) to detect laughter. Moreover, the modulation spectrum feature is chosen to exploit the fact that syllable rates are greater for laughter than for speech [17] (Bickley and Hunnicutt, 1992).

In addition to frame-level and utterance-level features, other parameterization methods are recently introduced. In [135] (Weninger et al., 2011) [117] (Schuller and Weninger, 2010), the authors show that integrating likelihood features derived from Nonnegative Matrix Factorization into Bidirectional Long Short-Term Memory Recurrent Neural Networks provide better results in terms of discriminating nonlinguistic vocalizations from speech.

In addition, phonetic transcription of laughter could be used to extract useful features to model the laughter sounds [131] (Urbain et al., 2011).

9.2.2 Machine Learning Techniques

The features described in the previous section are exploited as an input for the different machine learning techniques to model laughter and speech. Mainly, four modeling techniques are used which are: Gaussian Mixture Model-GMM, Hidden Markov Model-HMM, Neural Network, and Support Vector Machine-SVM.

Generative modeling methods such as GMM and HMM are trained on laughter and non-laughter data as explained in [73] (Laskowski and Schultz, 2008), then they are used to label an unknown audio files according to the likelihood score of each frame. On the other hand, discriminative classifiers are also exploited to segment the audio data [135] (Weninger et al., 2011) [69] (Kennedy and Ellis, 2004).

In [128] (Truong and Van Leeuwen, 2007), the authors investigate fusion of GMM and SVM methods to improve the performance of the laughter detection system. The reason of this fusion is to exploit the strength of each approach. The fusion is performed on the score level by summing and weighting the output score obtained from each classifier.

Recent works [101] (Petridis and Pantic, 2008) [116] (Scherer et al., 2009) [110] (Reuderink et al., 2008), exploit visual information to detect laughter in videos. These methods are used to build a multimodal system to locate nonlinguistic vocalizations within audiovisual data.

In this section, nonlinguistic vocalizations detection system were described. These systems are generally based on machine learning techniques using frame-level and utterance-level features. However, segmental approaches that capture higher-level information have not been adequately focused due to the nonlinguistic nature of laughter.

9.3 ALISP-based Laughter Detection System

This section describes our generic framework to detect nonlinguistic vocalizations using ALISP sequencing. The main purpose behind the proposed methodology is to adapt

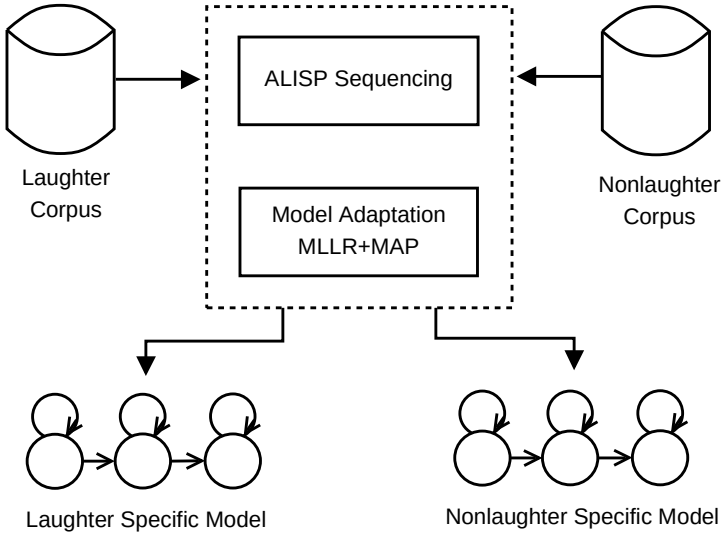


Figure 9.1: Workflow of the proposed methodology for ALISP-based acoustic model adaptation to detect nonlinguistic vocalizations ('Laughter' is used as an example for a specific set of nonlinguistic vocalizations).

ALISP HMMs in order to facilitate Viterbi decoding algorithm to detect similar regions from audio. The proposed framework is illustrated in figure 9.1, which shows the workflow of the proposed methodology for the specific example of detecting laughter vocalizations from audio. Laughter vocalizations are used as adaptation data to model laughter specific HMMs, while non-laughter audio (i.e. audio excluding laughter vocalizations) is used for getting non-laughter specific HMMs. Finally, a combined set of HMMs are used to discriminate laughter from audio with the help of Viterbi decoding algorithm.

9.3.1 ALISP Segmentation and Model Adaptation

As pointed out before, the acquired ALISP models can be used for pseudo-phonetic sequencing. In the current step, ALISP models are adapted to detect local regions of nonlinguistic vocalizations by providing some supervised adaptation data. Firstly, ALISP models segment the adaptation data and acquire segment labels as shown in figure 9.1. Next, using the segment labels and adaptation data, MLLR adaptation approach is applied to estimate a set of linear transformations for the mean and variance parameters for reducing mismatch

between the initial ALISP models and the adaptation set. Finally, the model is further adapted using MAP approach considering MLLR adapted model as a prior knowledge. Therefore, adaptation of ALISP models uses MLLR followed by MAP approaches.

We propose to adapt ALISP models for specific nonlinguistic vocalizations that need to be detected as well as for the remaining data excluding the vocalizations. In this way, the models are expected to deviate from each other in discriminating nonlinguistic vocalizations from speech. Figure 9.1 considers laughter as one of the nonlinguistic vocalizations. As shown in the figure, the adaptation is performed on the annotated laughter vocalizations as well as on the non-laughter part of audio corpora excluding laughter vocalizations.

9.3.2 Viterbi Decoding and Symbolic-level Smoothing

The Viterbi algorithm, a well-established technique for decoding an HMM sequence of states, is used in order to transform an observed sequence of speech features into a string of recognized ALISP units. In this work, a combined set of adapted ALISP models are used to discriminate nonlinguistic vocalizations from speech. Therefore, the labels of ALISP sequences that are generated from the Viterbi decoding are expected to follow a naming convention in order to support symbolic level post processing.

The other main advantage of ALISP HMM models is the possibility to operate on the level of symbols and sequences. The outliers in the Viterbi decoded sequence can be post-processed using contextual label information. This method proposes a simple voting scheme that uses a sliding window on the ALISP sequence to eliminate outliers in Viterbi-predicted sequence automatically. The sliding window counts 'yes/ no' votes depending on whether or not a symbol belongs to target vocalization. The window length is always expected to be an odd number and the result of majority votes decides if the middle segment is a part of nonlinguistic vocalization.

9.4 Experiments and Results

In this section, the experimental evaluation of the proposed method is compared to global acoustic models in discriminating laughter from speech. Firstly, the laughter-

annotated experimental corpus is described. Secondly, global HMMs (i.e. laughter versus non-laughter models) are modeled and ALISP HMM models are adapted, as described in section 9.3.1, on laughter and non-laughter training datasets. In addition, a combined set of laughter and non-laughter ALISP HMM models are used together to segment the test data set using the Viterbi algorithm. Consequently, the symbolic-level smoothing is applied to eliminate outliers from the predicted ALISP sequences. Finally, the results of our method are analyzed.

9.4.1 Experimental Corpus

The proposed laughter detection system requires supervised training material for non-linguistic vocalizations that has manual annotation. A combined audio corpus is used, it contains laughter annotations from three publicly available sources SEMAINE-DB [86] (McLeown et al., 2012), AVLaughterCycle [130] (Urbain et al., 2010), and Mahnob laughter databases [100] (Petridis et al., 2013). More details about these database are given in section 4.5. The corpus is an appropriate mix of hilarious and conversational laughter vocalizations. The data is uniformly divided into approximately 80% for training and 20% for testing. Table 9.1 shows the size of laughter and non-laughter audio (in seconds) used for training and testing.

9.4.2 Laughter Modeling

In order to detect laughter vocalizations from speech, we have trained global acoustic models such as GMMs, serial HMMs and ergodic HMMs with different HMM topologies, as shown in figure 9.2. All of the above global acoustic models include an additional silence

	Laughter [sec]	non-laughter [sec]
Training	3943	4957
Test set	853	1206
Total	4796	6163

Table 9.1: Training and test data sets used to train the specific HMM models and to evaluate the ALISP-based system.

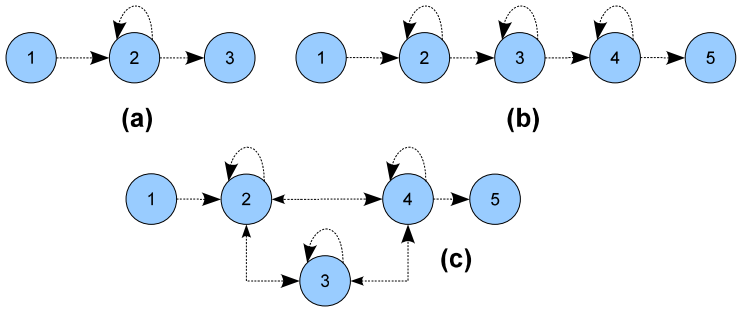


Figure 9.2: Global HMM topologies: (a) Simple GMM; (b) Serial (left-to-right) HMM; (c) Ergodic (fully-connected) HMM.

model.

In this work, the unlabeled audio corpus is modeled by the set of 32 ALISP HMM model (i.e. pseudo-phonetic HMMs) along with a silence model (The same ALISP model used on previous chapter). This set can be considered as an universal acoustic model because of its training database includes all possible sounds like music, laughter, advertisements etc. It can be used not only for segmenting any audio, but also for getting pseudo-phonetic (symbolic) transcription.

In order to represent ALISP segments, the segmentation system uses 32 ALISP symbols (such as `HA`, `HB` and `H4`), referring each to an ALISP HMM model, in addition to a silence label (`Hsil`). Figure 9.3 shows an example of the segmentation task performed by the ALISP segmental HMMs on an unseen laughter vocalization.

In the next step, we adapt the generic ALISP HMM models into:

- Laughter specific ALISP HMMs by using laughter vocalizations as adaptation data.
- Non-laughter specific ALISP HMMs considering non-laughter vocalizations (audio excluding laughter vocalizations) as adaptation data.

In order to facilitate combining the two sets, laughter-specific adapted models are renamed such that `HA` to `LHA`, `H4` to `LH4`, and so on. On the other hand, non-laughter specific adapted models keeps the same names such as `HA`, `H4`, `HB`, etc. The combined set of the models (referred as ALISP-adapt) are used to discriminate local regions of laughter. As

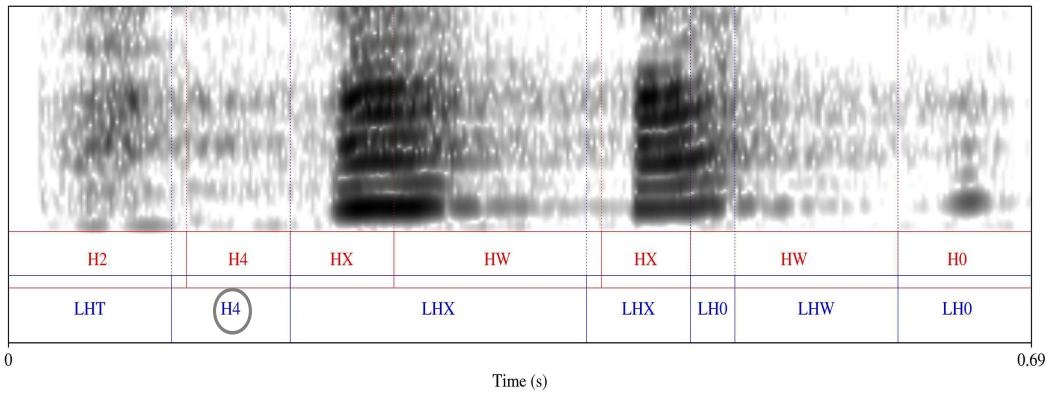


Figure 9.3: Segmentation task performed on an unseen laughter vocalization by: (i) generic ALISP HMMs before model adaptation (top row labels that are in Red); (ii) Combined set of specific (or adapted) ALISP HMMs after MLLR+MAP adaptation (i.e. ALISP-adapt) (bottom row labels that are in Blue). The marked symbol with a circle is an outlier which can be automatically found using proposed smoothing scheme on ALISP sequences.

shown in Figure 9.3, laughter specific regions seemed to be detected by the model except some outliers. In order to eliminate these outliers a majority voting scheme has been proposed in section 7.4.2. The smoothing scheme is experimented using sliding window size 3 (referred as ALISP-adapt-sm3) and 5 (referred as ALISP-adapt-sm5). According to the scheme, for example, the outlier (H4) in figure 9.3 obtains majority 'yes' votes in case of laughter detection if sliding window size is either 3 or 5. Such a way, we can automatically detect and eliminate the outliers.

9.4.3 Results

Table 9.2 shows the precision, recall and F-measures obtained from different approaches to detect laughter on test set. The F-measures is computed as follows:

$$F = \frac{2(precision \times rappel)}{precision + rappel} \tag{9.1}$$

Among the global acoustic models, ergodic HMMs perform better than GMMs and serial (left-to-right) HMMs. Ergodic HMMs show high precision (92.8%) in locating laughter

regions, whereas serial HMMs are relatively good in recall (86.3%) rates. When compared with adapted ALISP segmental HMMs (ALISP-adapt), global ergodic HMMs are still 4.2% better in precision. However, the ALISP HMM models (ALISP-adapt) perform better in terms of overall accuracy (F-measure) when compared to global HMMs.

Adapted ALISP HMM models provide an additional flexibility to find outliers with the help of a simple majority voting scheme. Therefore, ALISP-adapt-sm3 and ALISP-adapt-sm5 show improvement in terms of F-measure when compared to ALISP-adapt by 2.9% and 4.4% of respectively. Overall, ALISP-adapt-sm5 show 94.3% precision and 93.9% recall rates and perform relatively better than all other approaches experimented in this work.

[%]	Precision	Recall	F-measure
GMMs	70.8	78.6	74.5
Serial HMMs	85.7	86.3	86.0
Ergodic HMMs	92.8	84.5	88.5
ALISP-adapt	88.6	90.9	89.7
ALISP-adapt-sm3	92.4	92.7	92.6
ALISP-adapt-sm5	94.3	93.9	94.1

Table 9.2: Precision, Recall and F-measure values computed on the evaluation set for the different systems of laughter detection.

9.5 Conclusion

In this chapter, we proposed a generic approach for detecting nonlinguistic vocalizations using ALISP sequencing. In fact, this is the first time that a data-driven approach is applied for the detection of nonlinguistic vocalizations.

The proposed methodology was evaluated against global acoustic models such as GMMs, left-to-right HMMs and ergodic HMMs on a laughter-annotated audio corpus. The results show that the proposed methodology yields an increase of 19.6%, 8.1% and 5.6% on F-measure against the three methods compared respectively.

With this work, we argue that the adaptation of ALISP HMM models is useful in detecting local regions of nonlinguistic vocalizations. This method has further facilitated us

to improve the performance using symbolic-level smoothing such as majority voting scheme with sliding window approach.

Chapter 10

Conclusions, Discussions and Perspectives

10.1 Conclusions

In this thesis, we propose a generic audio indexing system to retrieve and recognize the majority of the audio items present in a radio streams. These items are usually: music, commercial, jingle, speech and nonlinguistic vocalization (such as laughter, cough, sight,...). To his end, an audio indexing system based on data-driven ALISP technique is exploited for radio streams indexing and used for different fields to cover all the items that could be present in a radio stream. The proposed audio indexing system is composed of three modules:

- Automated acquisition (with an unsupervised machine learning methods) and Hidden Markov Modeling (HMM) of ALISP acoustic models.
- Segmentation module (also referred as sequencing) that transforms the audio data into a sequence of symbols (using the previously acquired ALISP Hidden Markov Models).
- Comparison and decision module, including approximate matching algorithms inspired from the Basic Local Alignment Search (BLAST) tool widely used in bioinformatics and the Levenshtein distance, to search for a sequence of ALISP symbols of

unknown audio data in the reference database (related to different audio items).

Throughout this thesis we have shown that the proposed ALISP data-driven approach can be used to extract high-level information for audio indexing. Our major contributions can be summarized as follows:

1. Improving the ALISP tools by introducing a simple method to find stable segments within the audio data. This technique, referred as spectral stability segmentation, is replacing the temporal decomposition used before for speech processing. The main advantage of this method is its computation requirements which are very low compared to those of temporal decomposition.

2. Proposing an efficient technique to retrieve relevant information from ALISP sequences using BLAST algorithm and Levenshtein distance. This method speeds up the retrieval process without affecting the accuracy of the audio indexing process.

3. Proposing a generic audio indexing system, based on data-driven ALISP sequencing, for radio streams indexing. This system is applied and evaluated for different fields of audio indexing to cover the majority of audio items that could be present in a radio stream:

   - audio identification: detection of occurrences of a specific audio content (music, advertisement, jingle) in a radio stream;
   - audio motif discovery: detection of repeating objects in audio streams. (music, advertisement, and jingle);
   - speaker diarization: segmentation of an input audio stream into homogenous regions according to speaker's identities in order to answer the question "Who spoke when?";
   - nonlinguistic vocalization detection: detection of nonlinguistic sounds such as laughter, sighs, cough, or hesitation;

The evaluation of the proposed audio identification system, in the 2010 QUAERO evaluation campaign, shows the relevance of our ALISP-based fingerprint compared to the

other systems. Moreover, it was shown that the best configuration of ALISP HMM models is the one using the multi-Gaussian configuration with 33 ALISP units and the spectral stability method for the initial segmentation. The choice of spectral stability segmentation is motivated by its simplicity compared to the temporal decomposition.

For the audio motif discovery the experimental results shows that the proposed system performs as well as the systems using audio fingerprinting to detect repeating objects in radio streams. Furthermore, our system shows its ability to detect long repeating objects, such as songs and short repeating objects, such as advertisements, using the same configuration.

The ALISP-based speaker diarization system was evaluated during the 2011 ETAPE evaluation campaign and has obtained the best results in the ETAPE 2011 evaluation campaign among 7 participants. Moreover, a speaker identification system was developed to measure the speech time of politicians in radio streams. This systems has obtained the best simple system performance on Female gender in the MOBIO 2010 evaluation campaign.

Finally, for the nonlinguistic vocalization detection, the segmental HMMs provided by ALISP tools outperformed the global acoustic models (GMM, serial HMM, ergodic HMM). Actually, the proposed system showed a 94.3% precision and 93.9% recall rates and performed relatively better than all other approaches experimented in laughter detection.

10.2 Discussions

This thesis opened the way of exploiting high-level information for audio indexing using data-driven approaches. Nevertheless, there are still many points to discuss.

First, the proposed audio identification system was not able to recognize different versions of a song (such as live and studio versions). This problem raises the question of how we could improve the ALISP HMM models to take into account this variability and to use the proposed framework for cover song detection.

Second, we showed that the proposed audio motif discovery system performed very well in the case of repeating songs and advertisements. But what about detecting repeating worlds or sentences in speech data? It was shown in our work that ALISP segmenter is speaker-dependent. It means that, ALISP transcriptions of identical sentences spoken by

different speakers are very different, while the ALISP transcription of identical sentences spoken by the same speaker is very similar. This could be an interesting point to explore in order to train a new set of ALISP HMM models that are speaker-independent.

Third, our speaker diarization system is based on the assumption that we dispose of an annotated previous broadcasted copy of the show to be segmented, which is not always the case. Moreover our system is not appropriate for meetings and other type of spoken document that we cannot have an annotated previous copy.

Finally, we should evaluate the performances of the proposed framework to retrieve all the audio items simultaneously. This evaluation is requiring an audio corpus with a detailed annotations of music, jingle, advertisement, speaker turn, nonlinguistic vocalizations. However disposing of such a corpus is not obvious which make the realization of this evaluation very complicated.

10.3 Perspectives

Many perspectives could result from this work:

- An extension of our work to the visual context. The main idea is to train an audiovisual data driven model and exploit them to build a generic audiovisual indexing system. To this end, a coupled data-driven HMM models will be used to characterize the state asynchrony of the audio and visual observations features while their natural correlation over time is preserved. This technique was used before for audio-visual speech recognition and showed better results than multistream HMM [93] (Nefian et al., 2002).

- Improving the speaker diarization system by using the semantic information derived from an automatic speech recognition system. The resulting transcriptions will be used to locate the current, previous or next speaker. Transcripts such as, "bonjour et bienvenue, Romain" and "bienvenue à toutes et à tous, c'est Christophe Ruaults" could be exploited to correct the output segmentation of the diarization system.

- A parallel processing can be done in order to speed up the processing computation. The proposed audio indexing system will be integrated in a car radio engine which require a simultaneous treatment of several radio stations. Moreover, MFCCs computation and Viterbi algorithms along with the approximate matching of ALISP units will be studied in order to detect the part that could be parallelized and implemented using Graphic Processor Unit (GPU) architecture.

- Exploiting the framework used in the nonlinguistic vocalizations detection to domestic sounds such as door closure, impact transients, machinery that could be used to provide specific voice controlled home care and communication services people suffering from chronic diseases and persons suffering from (fine) motor skills impairments[1].

[1] http://vassist.cure.at/project_overview/

Personal Bibliography

.

1. P. Perrot, J. Razik, M. Morel, H. Khemiri and G. Chollet. Techniques de conversion de voix appliquées à l'imposture. Traitement ET Analyse dE l'Information: Méthodes et Applications (TAIMA), 2009.

2. L. Zouari, H. Khemiri, J. Razik, A. Amehraye and G. Chollet. Reconnaissance de la parole en temps réel pour le dialogue oral. Traitement ET Analyse dE l'Information: Méthodes et Applications (TAIMA), 2009.

3. G. Chollet, A. Amehraye, J. Razik, L. Zouari, H. Khemiri and C. Mokbel. Spoken Dialogue in Virtual Worlds Chap. Development of Multimodal Interfaces: Active Listening and Synchrony, Springer Verlag, LNCS, 5967:423-443, 2010.

4. H. Khemiri, G. Chollet and D. Petrovska-Delacrétaz. Automatic Detection of Known Advertisements in Radio Broadcast with Data-driven ALISP Transcriptions. International Workshop on Content-Based Multimedia Indexing (CBMI), pages 223-228, 2011.

5. B. Happi-Tietche, O. Romain, B. Denby, F. de Dieuleveult, B. Granado, M. Karabernou, H. Khemiri, G. Chollet, D. Petrovska, R. Blouet, K. Hachicha, S. Viateur. SurfOnHertz: un navigateur hertzien en radio logicielle pour l'indexation des bandes de radiodiffusion FM. Colloque GRETSI, 2011.

6. H. Khemiri, D. Petrovska-Delacrétaz and G. Chollet. Une empreinte audio à base d'ALISP appliquée à l'identification audio dans un flux radiophonique. Colloque en COmpression et REprésentation des Signaux Audiovisuels (CORESA), 2012.

7. B. Happi-Tietche, O. Romain, B. Denby, L. Benaroya, F. De Dieuleveult, B. Granado, H. Khemiri, G. Chollet, D. Petrovska-Delacrétaz, R. Blouet, K. Hachicha and S. Viateur. Software Radio FM Broadcast Receiver for Audio Indexing Applications. IEEE International Conference on Industrial Technology (ICIT), pages 585-590, 2012.

8. B. Happi-Tietche, O. Romain, B. Denby, L. Benaroya, F. De Dieuleveult, B. Granado, G. Wassi H. Khemiri, G. Chollet, D. Petrovska-Delacrétaz, R. Blouet, K. Hachicha and S. Viateur. Prototype of a radio-on-demand broadcast receiver with real time musical genre classification. Conference on Design and Architectures for Signal and Image Processing (DASIP), pages 1-2, 2012.

9. H. Khemiri, D. Petrovska-Delacrétaz and G. Chollet. A Generic Audio Identification System for Radio Broadcast Monitoring Based on Data-driven Segmentation. IEEE International Symposium on Multimedia (ISM), pages 427-432, 2012.

10. H. Khemiri, G. Chollet and D. Petrovska-Delacrétaz. Automatic Detection of Known Advertisements in Radio Broadcast with Data-driven ALISP Transcriptions. Multimedia Tools And Applications (MTAP), 62(1):35-49, 2013.

11. H. Khemiri, D. Petrovska-Delacrétaz and G. Chollet. Speaker Diarization Using Data-driven Audio Sequencing. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

12. S. Pammi, H. Khemiri, D. Petrovska-Delacrétaz and G. Chollet. Detection Of Non-linguistic Vocalizations Using ALISP Sequencing. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

13. E. ElKhoury, B. Vesnicer, J. Franco-Pedroso, R. Violato, Z. Boulkenafet, L.M. Mazaira-Fernandez, M. Diez, J. Kosmala, H. Khemiri, T. Cipr, R. Saeidi, M. Gunther, J. Zganec-Gros, R. Zazo-Candil, F. Simoes, M. Bengherabi, A. Alvarez-Marquina, M. Penagarikano, A. Abad, M. Boulayemen, P. Schwarz, D. Van-Leeuwen, J. Gonzalez-Dominguez, M. Uliani-

Neto, E. Boutellaa, P. Gomez-Vilda, A. Varona, D. Petrovska-Delacrétaz, P. Matejka, J. Gonzalez-Rodriguez, T. Pereira, F. Harizi, L. J. Rodriguez-Fuentes, L. ElShafey, M. Angeloni, G. Bordel, G. Chollet and S. Marcel. The 2013 Speaker Recognition Evaluation in Mobile Environment. The International Conference on Biometrics (ICB), 2013.

Bibliography

[1] Cambridge University Engineering Department. HTK: Hidden Markov Model ToolKit, http://htk.eng.cam.ac.uk.

[2] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan. Unknown-Multiple Speaker Clustering Using Hmm. In International Conference on Spoken Language Processing, pages 573–576, 2002.

[3] S. F. Altschul, W. Gish, and W. Miller. Basic Local Alignment Search Tool. Journal of Molecular Biology, 215:403–410, October 1990.

[4] X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando. Hybrid Speech/non-speech detector applied to Speaker Diarization of Meetings. In Speaker and Language Recognition Workshop, Odyssey, pages 1–6, 2006.

[5] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker Diarization: A Review of Recent Research. IEEE Transactions on Audio, Speech, and Language Processing, 20(2):356–370, 2012.

[6] X. Anguera, C. Wooters, B. Peskin, and M. Aguiló. Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. In International conference on Machine Learning for Multimodal Interaction, pages 26–38, 2005.

[7] A. Apostolico, M. Comin, and L. Parida. Bridging Lossy and Lossless Compression by Motif Pattern Discovery. In Rudolf Ahlswede, Lars Bäumer, Ning Cai, Harout Aydinian, and Vladimir Blinovsky, editors, General Theory of Information Transfer

and Combinatorics, chapter Bridging lossy and lossless compression by motif pattern discovery, pages 793–813. Springer-Verlag, 2006.

[8] B. Atal. Efficient Coding of LPC Parameters by Temporal Decomposition. ICASSP, pages 81–84, 1983.

[9] J.A. Bachorowski, M.J. Smoski, and M.J. Owren. The acoustic features of human laughter. Journal of the Acoustical Society of America, 110(3):1581–1597, 2001.

[10] I. Badr, I. McGraw, and J. Glass. Pronunciation Learning from Continuous Speech. In Interspeech, pages 549–552, 2011.

[11] J. Baker, J. Li Deng, J. Glass, S. Khudanpur, S. Chin-hui Lee, N. Morgan, and D. O'Shaughnessy. Research Developments and Directions in Speech Recognition and Understanding. IEEE Signal Processing Magazine, 26(3):75–80, 2009.

[12] S. Baluja and M. Covell. Waveprint: Efficient wavelet-based audio fingerprinting. Pattern Recognition, 41(11):3467–3480, 2008.

[13] C. Barras, Xuan Zhu, S. Meignier, and J. Gauvain. Multistage speaker diarization of broadcast news. IEEE Transactions on Audio, Speech, and Language Processing, 14(5):1505–1512, 2006.

[14] M.A. Bartsch and G.H. Wakefield. To catch a chorus: using chroma-based representations for audio thumbnailing. In Workshop on the Applications of Signal Processing to Audio and Acoustics, pages 15–18, 2001.

[15] M. Betser. Décomposition harmonique des signaux audio appliqué à l'indexation audio. PhD thesis, Telecom ParisTech, 2008.

[16] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz, and A. Sixtus. Large Vocabulary Continuous Speech Recognition of Broadcast News – The Philips/RWTH Approach. Speech Communication, 37(1/2):109–131, 2002.

[17] C.A. Bickley and S. Hunnicutt. Acoustic analysis of laughter. In International Conference on Spoken Language Processing, pages 927–930, 1992.

[18] F. Bimbot and B. Atal. An evaluation of temporal decomposition. In EUROSPEECH, 1991.

[19] M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. Speech Communications, 50(5):434–451, 2008.

[20] S. Bozonnet, N.W.D. Evans, and C. Fredouille. The lia-eurecom RT'09 speaker diarization system: Enhancements in speaker modelling and cluster purification. In IEEE International Conference on Acoustics Speech and Signal Processing, pages 4958–4961, 2010.

[21] C.J.C. Burges, D. Plastina, J.C. Platt, E. Renshaw, and H.S. Malvar. Using audio fingerprinting for duplicate detection and thumbnail generation. In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 3, pages 9–12, 2005.

[22] C.J.C Burges, J.C. Platt, and S. Jana. Distortion discriminant analysis for audio fingerprinting. IEEE Transactions on Speech and Audio Processing, 11(3):165–174, 2003.

[23] N. Campbell, R. Kashioka, H., and R. Ohara. No laughing matter. In Interspeech, pages 465–468, 2005.

[24] P. Cano, E. Battla, H. Mayer, and H. Neuschmied. Robust Sound Modeling for Song Detection in Broadcast Audio. Audio engineering society, 2002.

[25] P. Cano, E. Battle, T. Kalker, and J. Haitsma. A review of audio fingerprinting. Journal of VLSI Signal Proc., 41(3):271–284, November 2005.

[26] J. Černocký. Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification. PhD thesis, Université Paris XI Orsay, 1998.

[27] D. Charlet, C. Barras, and J.S. Liénard. Impact of Overlapping Speech Detection on Speaker Diarization for Broadcast News and Debates. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[28] G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot. Towards ALISP: a proposal for Automatic Language Independent Speech Processing, pages 375–388. NATO ASI Series. Springer Verlag, 1999.

[29] G. Chollet, K. McTait, and D. Petrovska-Delacrétaz. Data driven approaches to speech and language processing. Lecture notes in computer science, pages 164–198, 2005.

[30] I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamoon. A Secure, Robust Watermark for Multimedia. In International Workshop on Information Hiding, pages 185–206, 1996.

[31] M. Cremer, B. Froba, J. Hellmuth, O. Herre, and E. Allamanche. AudioID: Towards Content-Based Identification of Audio Material. In Audio Engineering Society Convention 110, 2001.

[32] R.B Dannenberg. Listening to "Naima": An Automated Structural Analysis from Recorded Audio. In International Computer Music Conference, pages 28–34, 2002.

[33] R.B. Dannenberg and N. Hu. Pattern discovery techniques for music audio. In International Conference on Music Information Retrieval, pages 63–70, 2002.

[34] A. De Cheveigné. Computational Auditory Scene Analysis: Principles, Algorithms and Applications, chapter Multiple F0 estimation, pages 65–70. Wiley/IEEE Press, 2006.

[35] P. Delacourt, D. Kryze, and C. Wellekens. Detection of speaker changes in an audio document. In EUROSPEECH, 1999.

[36] P. Delacourt and C. Wellekens. DISTBIC : A speaker-based segmentation for audio data indexing. Speech Communication, 32:111–126, 2000.

[37] S. Deligne and F. Bimbot. Inference of variable-length linguistic and acoustic units by multigrams. Speech Commun., 23(3):223–241, 1997.

[38] S. Deligne, S. Dharanipragada, R. A. Gopinath, B. Maison, P.A. Olsen, and H. Printz. A robust high accuracy speech recognition system for mobile applications. IEEE Transactions on Speech and Audio Processing, 10(8):551–561, 2002.

[39] A. El Hannani. Text-Independant Speaker Verification Based On High-Level Information Extracted With Data-Driven Methods. PhD thesis, University of Fribourg (Switzerland) and INT/SITEVRY (France), 2007.

[40] A. El Hannani, D. Petrovska-Delacrétaz, B. Fauve, A. Mayoue, J. Mason, J.F. Bonastre, and G. Chollet. Text independent Speaker Verification. In Guide to Biometric Reference Systems and Performance Evaluation. Springer Verlag, 2009.

[41] E. El-Khoury. Unsupervised Video Indexing based on Audiovisual Characterization of Persons. PhD thesis, University of Toulouse, 2010.

[42] E. El-Khoury, C. Senac, and J. Pinquier. Improved speaker diarization system for meetings. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4097–4100, 2009.

[43] S. Fenet, M. Moussallam, Y. Grenier, G. Richard, and L. Daudet. A framework for fingerprint-based detection of repeating objects in multimedia streams. In EUSIPCO, pages 1464–1468, 2012.

[44] S. Fenet, G. Richard, and Y. Grenier. A Scalable Audio Fingerprint Method with Robustness to Pitch-Shifting. In International Symposium on Music Information Retrieval, pages 121–126, 2011.

[45] F. Fischler and R. Bolles. Readings in computer vision: issues, problems, principles, and paradigms. chapter Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, pages 726–740. 1987.

[46] C. Fredouille, S. Bozonnet, and N. W D Evans. The LIA-EURECOM RT'09 Speaker Diarization System. In NIST Rich Transcription Workshop, 2009.

[47] G. Friedland, O. Vinyals, Y. Huang, and C. Muller. Prosodic and other Long-Term Features for Speaker Diarization. IEEE Transactions on Audio, Speech, and Language Processing, 17(5):985–993, 2009.

[48] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier. The ESTER Phase II Evaluation Campaign for the Rich Transcription o French Broadcast News. In EUROSPEECH, 2005.

[49] S. Galliano, G. Gravier, and L. Chaubard. The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In Interspeech, pages 2583–2586, 2009.

[50] J. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. Transactions on Speech and Audio Processing, 2(2):291–298, 1994.

[51] H. Gish, M. M Siu, A. Chan, and W. Belfield. Unsupervised Training of an HMM-based Speech Rec. System for Topic Class. In Interspeech, 2009.

[52] H. Gish, Man-Hung S., and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In IEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 873–876, 1991.

[53] J. Glass. Towards unsupervised speech processing. In International Conference on Information Science, Signal Processing and their Applications, pages 1–4, 2012.

[54] C. Gollan, S. Hahn, R. Schluter, and H. Ney. An Improved Method for Unsupervised Training of LVCSR Systems. In Interspeech, pages 2101–2104, 2007.

[55] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert. The ETAPE corpus for the evaluation of speech-based TV content processing in the French

language. In International Conference on Language Resources, Evaluation and Corpora, 2012.

[56] V. Gupta, G. Boulianne, P. Kenny, P. Ouellet, and P. Dumouchel. Speaker diarization of French broadcast news. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4365–4368, 2008.

[57] J. Haitsma and T. Kalker. A Highly Robust Audio Fingerprinting System. In International Society for Music Information Retrieval, pages 107–115, 2002.

[58] J. Haitsma and T. Kalker. Speed-change resistant audio fingerprinting using autocorrelation. In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 4, pages 728–31, 2003.

[59] K.J. Han and S.S. Narayanan. Agglomerative hierarchical speaker clustering using incremental Gaussian mixture cluster modeling. In Interspeech, pages 20–23, 2008.

[60] D.F. Harwath, H. J. Timothy, and J.R. Glass. Zero Resource Spoken Audio Corpus Analysis. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[61] C. Herley. ARGOS: automatically extracting repeating objects from multimedia streams. IEEE Transactions on Multimedia, 8(1):115–129, 2006.

[62] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America, 87(4):1738–1752, 1990.

[63] M. Huijbregts, D.A. Van Leeuwen, and C. Wooters. Speaker Diarization Error Analysis Using Oracle Components. IEEE Transactions on Audio, Speech, and Language Processing, 20(2):393–403, 2012.

[64] A. Jansen and K. Church. Towards Unsupervised Training of Speaker Independent Acoustic Models. In INTERSPEECH, pages 1693–1692, 2011.

[65] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel. Speaker segmentation and clustering in meetings. In International Conference on Spoken Language Processing, 2004.

[66] P. Joly, J. Benois-Pineau, E. Kijak, and G. Quenot. The Argos Campaign: Evaluation of Video Analysis Tools. In International Workshop on Content-Based Multimedia Indexing, pages 130–137, 2007.

[67] S. Kanthak and H. Ney. Multilingual Acoustic Modeling Using Graphemes. In EUROSPEECH, volume 1, pages 1145–1148, 2003.

[68] Y. Ke, D. Hoiem, and R. Sukthankar. Computer vision for music identification. In IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 597–604, 2005.

[69] L. Kennedy and D. P. W. Ellis. Laughter Detection in Meetings. In NIST Meeting Recognition Workshop, pages 118–121, 2004.

[70] M. Killer, S. Stuker, and T. Schultz. Grapheme Based Speech Recognition. In EUROSPEECH, pages 3141–3144, 2003.

[71] M. Knox and N. Mirghafori. Automatic laughter detection using neural networks. In Interspeech, pages 2973–2976, 2007.

[72] S. Kullback and R.A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics, 22(1):79–86, 1951.

[73] K. Laskowski and T. Schultz. Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings. In workshop on Machine Learning for Multimodal Interaction, pages 149–160, 2008.

[74] V.B. Le, O. Mella, and D. Fohr. Speaker diarization using normalized cross likelihood ratio. In Interspeech, pages 1869–1872, 2007.

[75] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer Speech and Language, 9(2):171–185, 1995.

[76] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and control theory, 10:707–710, 1966.

[77] Y. Linde, A. Buzo, and R.M. Gray. An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications, 28(1):84–95, 1980.

[78] Y. Liu, K. Cho, H.S. Yun, J.W. Shin, and N.S. Kim. DCT based multiple hashing technique for robust audio fingerprinting. In ICASSP, pages 61–64, 2009.

[79] J. Loof, C. Christian Gollan, and H. Ney. Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System. In Interspeech, pages 88–91, 2009.

[80] J.F. Lopez and D.P.W Ellis. Using Acoustic Condition Clustering To Improve Acoustic Change Detection On Broadcast News. In International Conference on Speech and Language Processing, pages 568–571, 2000.

[81] J.M. Makhoul, S. Roucos, and H. Gish. Vector Quantization in Speech Coding. Proceedings of the IEEE, 73(11):1551–1588, 1985.

[82] I. Malioutov, A. Park, R. Barzilay, and J. Glass. Making Sense of Sound: Unsupervised Topic Segmentation over Acoustic Input. In Annual Meeting of the Association of Computational Linguistics, pages 504–511, 2007.

[83] M. Marolt. SONIC: Transcription of Polyphonic Piano Music with Neural Networks. In Workshop on Current Research Directions in Computer Music, pages 217–224, 2001.

[84] A.F. Martin and J. S. Garofolo. NIST Speech Processing Evaluations: LVCSR, Speaker Recognition, Language Recognition. In IEEE Workshop on Signal Processing Applications for Public Security and Forensics, pages 1–7, 2007.

[85] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.F. Bonastre, P. Tresadern, and T. Cootes. Bi-Modal Person Recognition on a Mobile Phone: Using Mobile Phone Data. In IEEE International Conference on Multimedia and Expo Workshops, pages 635–640, 2012.

[86] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. Transactions on Affective Computing, 3(1):5–17, 2012.

[87] S. Meignier, J.F. Bonastre, and S. Igounet. E-HMM approach for learning and adapting sound models. In Speaker and Language Recognition Workshop, Odyssey, pages 175–180, 2001.

[88] P. Mermelstein. Distance Measures for Speech Recognition–Psychological and Instrumental. In Joint Workshop on Pattern Recognition and Artificial Intelligence, 1976.

[89] D. Moraru, M. Ben, and G. Gravier. Experiments on speaker tracking and segmentation in radio broadcast news. In Interspeech, 2005.

[90] A. Muscariello, G. Gravier, and F. Bimbot. An efficient method for the unsupervised discovery of signalling motifs in large audio streams. In Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on, pages 145–150, 2011.

[91] A. Muscariello, G. Gravier, and F. Bimbot. Zero-resource audio-only spoken term detection based on a combination of template matching techniques. In Interspeech, 2011.

[92] A. Muscariello, G. Gravier, and F. Bimbot. Unsupervised Motif Acquisition in Speech via Seeded Discovery and Template Matching Combination. IEEE Transactions on Audio, Speech, and Language Processing, 20(7):2031–2044, 2012.

[93] A.V. Nefian, L. Luhong, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled HMM for audio-visual speech recognition. In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 2013–2016, 2002.

[94] S. Novotney, R. Schwartz, and J. Ma. Unsupervised acoustic and language model training with small amounts of labelled data. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4297–4300, 2009.

[95] J.P. Ogle and D.P.W. Ellis. Fingerprinting to Identify Repeated Sound Events in Long-Duration Personal Audio Recordings. In IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 233–236, 2007.

[96] M. Padellini, F. Capman, and G. Baudoin. Very Low Bit Rate speech coding in Noisy Environments. In Speech and Computer (SPECOM), 2005.

[97] A.S. Park and J.R. Glass. Unsupervised Pattern Discovery in Speech. IEEE Transactions on Audio, Speech, and Language Processing, 16(1):186–197, 2008.

[98] W.R. Pearson and D.J. Lipman. Improved tools for Biological Sequence Comparison. In Proceedings of the National Academy of Sciences, pages 2444–2448, 1988.

[99] P. Perrot, G. Aversano, Raphael Blouet, M. Charbit, and G. Chollet. Voice Forgery Using ALISP: Indexation in a Client Memory. In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, pages 17–20, 2005.

[100] S. Petridis, B. Martinez, and M. Pantic. The MAHNOB Laughter database. Image Vision Comput., 31(2):186–202, 2013.

[101] S. Petridis and M. Pantic. Fusion of audio and visual cues for laughter detection. In International Conference on Image and Video Retrieval, pages 329–337, 2008.

[102] D. Petrovska-Delacrétaz, C. Černocký, J. Hennebert, and G. Chollet. Segmental Approaches for Automatic Speaker Verification. Digital Signal Processing, 10(3):198–212, 2000.

[103] Dijana Petrovska-Delacrétaz, Gérard Chollet, and Bernadette Dorizzi. Guide to Biometric Reference Systems and Performance Evaluation. Springer Verlag, 2009.

[104] J. Pinquier and R. André-Obrecht. Jingle detection and identification in audio documents. IEEE International Conference on Acoustics, Speech, and Signal Processing, 4:329–322, 2004.

[105] J. Pinquier, J.L. Rouas, and R. André-Obrecht. A fusion study in speech/music classification. In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 17–20, 2003.

[106] J. Ramírez, J. M. Gorriz, and J. C. Segura. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, pages 1–22. InTech, 2007.

[107] J.C. Ramírez, J.; Segura, C. Benítez, A. De La Torre, and A. Rubio. A new adaptive longterm spectral estimation voice activity detector. In EUROSPEECH, pages 3041–3044, 2003.

[108] M. Ramona, S. Fenet, R. Blouet, H. Bredin, T. Fillon, and G. Peeters. Audio Fingerprinting: a Public Evaluation Framework Based on a Broadcast Scenario. Applied Artificial Intelligence,, 26(1-2):119–136, 2011.

[109] M. Ramona and G. Peeters. Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection. In IEEE International Conference on Acoustics, Speech and Signal Processing, pages 477–480, 2011.

[110] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic. Decision-Level Fusion for Audio-Visual Laughter Detection. In International Workshop on Machine Learning for Multimodal Interaction, pages 137–148, 2008.

[111] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using Adapted Gaussian mixture models. In Digital Signal Processing, pages 19–41, 2000.

[112] J. Rissanen. Stochastic Complexity in Statistical Inquiry Theory. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1989.

[113] M. Rouvier and S. Meignier. A Global Optimization Framework For Speaker Diarization. In Speaker and Language Recognition Workshop, Odyssey, 2012.

[114] D. Roy and A. Pentland. Learning Words from Sights and Sounds: A Computational Model. Cognitive Science, 26:113–146, 2000.

[115] G. K. Sandve and F. Drablos. A survey of motif discovery methods in an integrated framework. Biology Direct, 1:1, 2006.

[116] S. Scherer, F. Schwenker, N. Campbell, and G. Palm. Multimodal Laughter Detection in Natural Discourses. Cognitive Systems Monographs, 6:111–120, 2009.

[117] B. Schuller and F. Weninger. Discrimination of speech and non-linguistic vocalizations by Non-Negative Matrix Factorization. In International Conference on Acoustics Speech and Signal Processing, pages 5054–5057, 2010.

[118] T. Schultz and K. Kirchhoff. Multilingual Speech Processing. Elsevier, 2006.

[119] M.A. Siegler, U. Jain, B. Raj, and R.M. Stern. Automatic Segmentation, Classification and Clustering of Broadcast News Audio. In DARPA Speech Recognition Workshop, pages 97–99, 1997.

[120] M. Sinclair and S. King. Where Are The Challenges in Speaker Diarization? In IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[121] A. Sinitsyn. Duplicate Song Detection using Audio Fingerprinting for Consumer Electronics Devices. In IEEE International Symposium on Consumer Electronics, pages 1–6, 2006.

[122] M. Siu, H. Gish, S. Lowe, and A. Chan. Unsupervised Audio Pattern Discovery using HMM-based Self-Organized Units. In Interspeech, 2011.

[123] I. Stylianou. Modèles Harmoniques plus Bruit combinés avec des Méthodes Statistiques, pour la Modification de la Parole et du Locuteur. PhD thesis, ENST, 1996.

[124] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech, and Language Processing, 14(5):1557–1565, 2006.

[125] A. Tritschler and R. Gopinath. Improved speaker segmentation and segments clustering using the bayesian information criterion. In EUROSPEECH, volume 2, pages 679–682, 1999.

[126] J. Trouvain. Segmenting phonetic units in laughter. In International Conference of the Phonetic Sciences, pages 2793–2796, 2003.

[127] K.P. Truong and D.A. Van Leeuwen. Automatic detection of laughter. In Interspeech, pages 485–488, 2005.

[128] K.P. Truong and D.A. Van Leeuwen. Automatic discrimination between laughter and speech. Speech Communication, 49(2):144–158, 2007.

[129] K.P. Truong and D.A. Van Leeuwen. Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features. In Interdisciplinary Workshop on the Phonetics of Laughter, pages 49–53, 2007.

[130] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. The AVLaughterCycle Database. In International Conference on Language Resources and Evaluation (LREC'10), pages 2996–3001, 2010.

[131] J. Urbain and T. Dutoit. A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In International Conference on Affective Computing and Intelligent Interaction, pages 397–406, 2011.

[132] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In IEEE Conference on Computer Vision and Pattern Recognition, volume 1, pages 511–518, 2001.

[133] A. Wang. The shazam music recognition service. Communications of the ACM, 49(5):44–48, 2006.

[134] D. Wang, R. Vipperla, and N Evans. Online pattern learning for non-negative convolutive sparse coding. In Interspeech, 2011.

[135] F. Weninger, B. Schuller, M. Wollmer, and G. Rigoll. Localization of non-linguistic events in spontaneous speech by Non-Negative Matrix Factorization and Long Short-

Term Memory. In International Conference on Acoustics, Speech and Signal Processing, pages 5840–5843, 2011.

[136] Z. Yaodong and J.R. Glass. Towards multi-speaker unsupervised speech pattern discovery. In IEEE International Conference on Acoustics Speech and Signal Processing, pages 4366–4369, 2010.

[137] S. Young, N.H. Russell, and J.H.S Thornton. Token Passing: a Conceptual Model for Connected Speech Recognition Systems. Technical report, Cambridge University, 1989.

[138] B. Zhu, W. Li, Z. Wang, and X. Xue. A novel audio fingerprinting method robust to time scale modification and pitch shifting. In Proceedings of the international conference on Multimedia, pages 987–990, 2010.

[139] X. Zhu, C. Barras, L. Lamel, and J.L. Gauvain. Speaker Diarization: From Broadcast News to Lectures. In Machine Learning for Multimodal Interaction, volume 4299 of Lecture Notes in Computer Science, pages 396–406. Springer Berlin Heidelberg, 2006.

[140] X. Zhu, C. Barras, S. Meignier, and J.L. Gauvain. Combining Speaker Identification and BIC for Speaker Diarization. In Interspeech, 2005.

Unified Data-Driven Approach for Audio Indexing Retrieval and Recognition

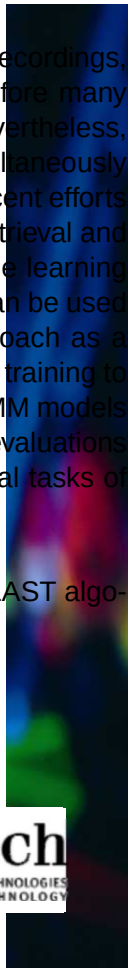
Houssemeddine KHEMIRI

RESUME : La quantité de données audio disponibles, telles que les enregistrements radio, la musique, les podcasts et les publicités est en augmentation constance. Par contre, il n'y a pas beaucoup d'outils de classification et d'indexation, qui permettent aux utilisateurs de naviguer et retrouver des documents audio. Dans ces systèmes, les données audio sont traitées différemment en fonction des applications. La diversité de ces techniques d'indexation rend inadéquat le traitement simultané de flux audio où différents types de contenu audio coexistent. Dans cette thèse, nous présentons nos travaux sur l'extension de l'approche ALISP, développé initialement pour la parole, comme une méthode générique pour l'indexation et l'identification audio. La particularité des outils ALISP est qu'aucune transcription textuelle ou annotation manuelle est nécessaire lors de l'étape d'apprentissage. Le principe de cet outil est de transformer les données audio en une séquence de symboles. Ces symboles peuvent être utilisés à des fins d'indexation. La principale contribution de cette thèse est l'exploitation de l'approche ALISP comme une méthode générique pour l'indexation audio. Ce système est composé de trois modules : acquisition et modélisation des unités ALISP d'une manière non supervisée, transcription ALISP des données audio et comparaison des symboles ALISP avec la technique BLAST et la distance de Levenshtein. Les évaluations du système proposé pour les différentes applications sont effectuées avec la base de données YACAST et avec d'autres corpus disponibles publiquement pour différentes tâche de l'indexation audio.

MOTS-CLEFS : indexation audio, modélisation HMM, segmentation ALISP, apprentissage non supervisée, algorithme BLAST.

ABSTRACT : The amount of available audio data, such as broadcast news archives, radio recordings, music and songs collections, podcasts or various internet media is constantly increasing. Therefore many audio indexing techniques are proposed in order to help users to browse audio documents. Nevertheless, these methods are developed for a specific audio content which makes them unsuitable to simultaneously treat audio streams where different types of audio document coexist. In this thesis we report our recent efforts in extending the ALISP approach developed for speech as a generic method for audio indexing, retrieval and recognition. The particularity of ALISP tools is that no textual transcriptions are needed during the learning step. Any input speech data is transformed into a sequence of arbitrary symbols. These symbols can be used for indexing purposes. The main contribution of this thesis is the exploitation of the ALISP approach as a generic method for audio indexing. The proposed system consists of three steps ; an unsupervised training to model and acquire the ALISP HMM models, ALISP segmentation of audio data using the ALISP HMM models and a comparison of ALISP symbols using the BLAST algorithm and Levenshtein distance. The evaluations of the proposed systems are done on the YACAST and other publicly available corpora for several tasks of audio indexing.

KEY-WORDS : audio indexing, HMM modeling, ALISP sequencing, unsupervised training, BLAST algorithm.



TELECOM ParisTech

EDITE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY