Une empreinte audio à base d'ALISP appliquée à l'identification audio dans un flux radiophonique

H. Khemiri[1,2] D. Petrovska-Delacrétaz[1] G. Chollet[2]

[1] Département Electronique et Physique

TELECOM SudParis, CNRS-SAMOVAR

[2] Département Traitement du Signal et des Images

TELECOM ParisTech, CNRS-LTCI

{khemiri, chollet}@telecom-paristech.fr {dijana.petrovska}@it-sudparis.eu

Résumé

Cet article présente un système d'identification audio pour détecter et identifier des publicités et des morceaux de musique dans les flux radiophoniques en utilisant des unités acoustiques. Ces unités, nommées ALISP (Automatic Language Independent Speech Processing), sont apprises de manière entièrement automatique grâce à la décomposition temporelle, la quantification vectorielle et des modèles HMM. L'originalité de l'approche est qu'aucune transcription n'est utilisée pour apprendre les modèles HMM. Pour identifier des morceaux de musique et les publicités, les transcriptions ALISP des morceaux de référence sont comparées aux transcriptions du flux radiophonique de test en utilisant la distance de Levenshtein. Pour l'identification des publicités, nous obtenons un taux de précision de 99% et un taux de rappel de 94% pour un flux de test contenant 4401 publicités. Pour l'identification de morceaux de musique nous obtenons un taux de précision de 100% et un taux de rappel de 95% sur un flux de test contenant 505 morceaux de musique.

Mots clefs

Outils ALISP, Identification Audio, Modèles HMM, Unités segmentales.

1 Introduction

L'identification audio par le contenu consiste à retrouver des métadonnées (artiste, nom de l'album, nom de la chanson, nom de la publicité, etc) à partir d'un extrait audio inconnu. Il y a de nombreuses applications potentielles de l'identification audio dont les plus populaires sont la surveillance automatique des flux radiophonique et l'identification d'un extrait audio inconnu capturé par dispositif mobile. La réalisation manuelle de la tâche de l'identification audio est assez fastidieuse et lente. Pour traiter ce problème, il existe deux grandes approches : le tatouage audio et l'extraction d'empreinte.

Le tatouage audio consiste à cacher l'information à identifier (nom de l'artiste, album,...) dans le document audio. L'enjeu de cette approche consiste à injecter les informations souhaitées sans altérer la qualité audio du document. Dans la seconde approche, une empreinte (ou signature) est extraite à partir du contenu audio inconnu et comparée aux empreintes des références stockées dans une base de données. Une empreinte audio est une représentation compacte du contenu audio.

Nous sommes intéressés par des méthodes basées sur l'extraction d'empreintes audio, qui sont plus appropriées pour la surveillance automatique des flux radiophoniques. L'identification 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'identification audio basé sur l'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 [1]. Le principal défi de ces systèmes est de calculer une empreinte audio robuste contre 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 référence.

Dans cet article, nous présentons notre système d'identification des publicités et des morceaux de musique dans un flux radiophonique. En se basant sur les outils ALISP (Automatic Language Independant Speech Processing) [2], notre technique permet d'atteindre une robustesse accrue aux distorsions présentes dans les diffusions radiophoniques.

Cet article est structuré de la façon suivante : la section 2 présente un état de l'art des principales méthodes de l'iden-

tification audio par extraction d'empreinte. La section 3 présente notre système d'identification audio basé sur les outils ALISP. La section 4 décrit les protocoles expérimentaux adoptés et expose les résultats obtenus pour l'identification des publicités et des morceaux de musique dans un flux radiophonique.

2 État de l'art des systèmes d'identification audio par extraction d'empreinte

Comme le montre la Figure 1, l'architecture générale d'un système d'identification par extraction d'empreinte est basée sur un module d'apprentissage qui consiste à transcrire les références en empreintes dans une base, et un module d'identification d'un extrait audio inconnu à travers la comparaison de son empreinte avec les empreintes des références.
Plusieurs méthodes d'identification audio par extraction d'empreinte ont été proposées [1]. 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 est de traiter le spectrogramme de chaque document audio comme une image 2-D et de transformer l'identification audio en un problème d'identification d'image. 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.

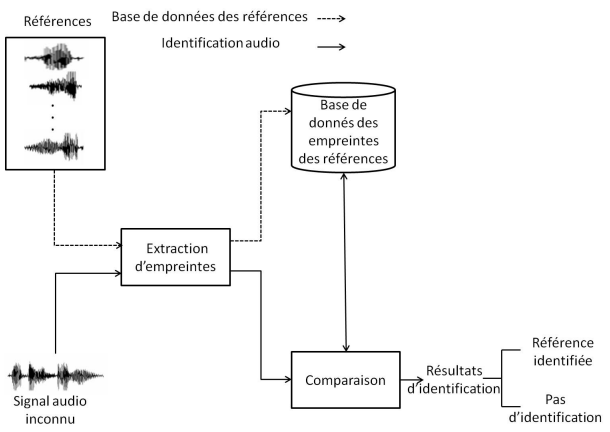

Figure 1 – *Architecture générale d'un système d'identification audio par extraction d'empreinte*

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. [3] ont développé un système d'identification audio pour la reconnaissance des morceaux de musique (méthode proposée par Philips). Ils ont utilisé une échelle de 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 recherche (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 la lookup table aura alors 2^{N_b} entrées. Chaque trame binaire de l'empreinte sert de clé dans la 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 et al. supposent 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. Y. Liu [4] a modifié l'algorithme pour contourner l'hypothèse de présence d'une trame binaire non distordue alors que les auteurs eux mêmes ont essayé d'améliorer la méthode d'extraction d'empreinte de façon à rendre plus robuste le système face aux distorsions comme l'étirement temporel (pitching) [5].
Un autre système commercial (Shazam) qui se base sur la représentation spectrale du système a été proposé par Wang [6] 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 qu'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), cette technique s'avère inefficace vue le nombre insuffisant des pics extraits. De plus Fenet et al. [7] 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).

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 l'extraction d'empreintes. L'idée principale est de traiter le spectrogramme de chaque document audio comme une image 2-D.

Baluja et al. [8] 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 l'extraction d'empreinte. Cette technique consiste à générer un spectrogramme à partir d'un signal audio avec les mêmes procédures que [3], 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. Le signe du premiers 200 amplitudes des ondelettes sont retenus dans l'empreinte. 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. [9] ont proposé un système d'identification de morceaux de musique basé sur l'algorithme de Viola-Jones [10]. 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 [11] est appliqué pour aligner le candidat avec la requête et une mesure de vraisemblance est calculée entre les deux morceaux.

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.

Allamanche [12] propose une approche essentiellement basée 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 (appelées 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. [13] 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'informations temporelles (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 [14] a été utilisé. Ce système a été évalué sur la tâche de l'identification des morceaux de musique dans un flux radio.

3 Système d'identification audio basé sur ALISP

Peu de travaux ont porté jusqu'à ce jour sur des techniques d'identification audio qui apprennent automatiquement à associer des données et des étiquettes [13]. C'est une originalité de l'approche ALISP (Automatic Language Independent Speech Processing), développée initialement pour le codage de la parole à très bas débit [2] et exploitée avec succès pour d'autres tâches, telles que la reconnaissance du locuteur [15] ou de la langue [16].

L'avantage majeur du système d'identification audio basé sur ALISP réside dans son déploiement facile sur de nouvelles applications. En effet, les systèmes d'identification audio par extraction d'empreinte présentés dans la Section 2 ont été évalués sur la tâche d'identification des morceaux de musique, alors que notre système basé sur ALISP a été tout d'abord adapté et évalué sur l'identification des morceaux de publicité dans un flux radiophonique puis appliqué directement à l'identification des morceaux de musique sans réapprendre de nouveaux modèles et sans changer la configuration initiale.

Notre système d'identification audio utilise les unités segmentales fournies par les outils ALISP pour identifier les publicités et les morceaux de musique dans les flux radio. En effet, les transcriptions ALISP des publicités et des morceaux de musique sont calculées en utilisant les modèles HMM fournis par les outils ALISP et l'algorithme de Viterbi et comparées aux transcriptions du flux radio en utilisant la distance de Levenshtein [17].

3.1 Acquisition et modélisation des unités ALISP

Comme expliqué dans [2], l'ensemble des unités ALISP est automatiquement acquis par la paramétrisation MFCC (Mel Frequency Cepstral Coefficients), la décomposition temporelle, la quantification vectorielle, et les modèles de Markov cachés.

La première étape du traitement est l'analyse spectrale des signaux audio. La paramétrisation des données audio se fait avec les coefficients MFCC, calculés sur une fenêtre de 20 ms, avec un décalage de 10 ms. Pour chaque trame, une fe-

nêtre de Hamming est appliquée et un vecteur cepstral de dimension 15 est calculé et ajouté à sa dérivée de premier ordre.
Dans la deuxième étape, une segmentation initiale du corpus à l'aide de la décomposition temporelle est effectuée [18]. Cette technique introduite par Atal [19] décompose la séquence des vecteurs MFCCs en cibles spectrales reliées entre elles par des fonction d'interpolation.
La prochaine étape dans le processus ALISP est la classification non supervisé effectuée via la quantification vectorielle. La quantification vectorielle cherche à regrouper les segments issus de la décomposition temporelle en un nombre limité de classes. Le nombre de classes est défini par la taille du dictionnaire de quantification vectorielle déterminé par l'algorithme LBG [20]. Nous n'avons pas utilisé les cibles spectrales comme ensemble d'apprentissage, mais les vecteurs cepstraux réels situés au centre de gravité des fonctions d'interpolation.
Les classes obtenues par décomposition temporelle et quantification vectorielle sont modélisées par des modèles de Markov cachés (HMM). Cette modélisation facilite leur utilisation dans un système de reconnaissance et est utilisée pour affiner le jeu d'unités acoustiques. Le nombre des modèles HMM est déterminé par le nombre de classes de la quantification vectorielle. Un affinement du jeu d'unités acoustiques est effectué en répétant quelques itérations comprenant un apprentissage des HMM suivi d'une resegmentation de la base de données avec les nouveaux modèles HMM. Le nombre des gaussiennes de chaque modèle ALISP est fixé en utilisant un algorithme de scission dynamique. Pour l'apprentissage des modèles HMM le logiciel HTK [21] est utilisé.

3.2 Reconnaissance des unités ALISP

La reconnaissance des unités ALISP travaille comme un système de reconnaissance de parole continue et utilise les mêmes techniques pour reconnaître une suite de segments acoustiques caractéristiques dans le signal à indexer.

3.3 Mesure de similarité et méthode de recherche

Après la reconnaissance d'unités ALISP, la prochaine étape du système proposé est le processus d'appariement. Ce module compare les séquences ALISP extraites de flux radio continu contre les transcriptions ALISP stockées dans la base de données de référence. Tout d'abord, les transcriptions ALISP de chaque publicité et morceau de musique 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 [17]. 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.
A ce stade, 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.

4 Expériences et résultats

Dans cette section, nous présentons les résultats de l'identification des publicités et morceaux de musique avec notre système.

4.1 Protocole expérimental

Pour l'expérience d'identification de publicité, 7 jours de diffusion radio fournis par Yacast (http ://www.yacast.fr) ont été utilisés. Les 7 jours sont répartis comme suit :
– Données de développement : pour apprendre les modèles ALISP 1 jour de flux audio de 12 radios est utilisé (288 heures) ; 3 jours de flux audio sont utilisés pour étudier la stabilité des transcriptions ALISP des publicités et fixer le seuil de décision pour la distance de Levenshtein.
– Données de référence : elles contiennent 2172 publicités qui représentent les publicités à détecter dans le flux radio. Ces publicités correspondent à des segments audio précédemment diffusés et annotées manuellement.
– Données d'évaluation : notre système est évalué sur 7 jours de flux audio de 11 radios françaises (l'équivalent de 77jours). Cette base de données contient 753 publicités différentes qui sont répétées entre 1 et 12 fois. Le nombre total des publicités est 4401.

Pour la tâche d'identification de morceaux de musique, le protocole expérimental du projet Quaero (http ://www.quaero.org/) a été adopté. Le projet Quaero comprend un sous-projet axé sur l'identification audio. Le protocole expérimental est décrit comme suit :
– Données de développement : Les mêmes que celles utilisées dans la tâche d'identification des publicités.
– Données de référence : elles contiennent 7309 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 : notre système est évalué sur 7 jours de la radio française RTL. Par conséquent, la durée totale est de 168 heures. Ces enregistrements contiennent 551 morceaux de musique.

La paramétrisation est faite avec les outils HTK. Pour la quantification vectorielle, la taille du dictionnaire est de 64, alors que la topologie utilisée pour la modélisation des HMM est de type gauche-droite avec 3 états émetteurs et 2 états non émetteurs. Dans nos travaux nous avons utilisé la fonction HVite de HTK pour la reconnaissance des unités ALISP. Cette fonction utilise une alternative de l'algorithme de Viterbi appelée méthode de passage du jeton [21].

4.2 Caractéristiques des unités ALISP

Le nombre actuel des unités ALISP est 65 (64 + modèle de silence) et la durée moyenne par unité est d'environ 100 ms. Par rapport à la méthode d'identification audio décrite dans [13] qui extrait 800 gènes par minute, notre approche propose aussi une représentation compacte des données audio avec 600 unités ALISP par minute.

4.3 Identification des publicités

Pour identifier les publicités et les morceaux de musique dans un flux radio, les transcriptions ALISP des références sont comparées aux transcriptions du flux radio en utilisant la distance de Levenshtein. Mais avant cela, la stabilité des transcriptions ALISP des publicités est étudiée. Cette étude, décrite dans [22], nous a permis de fixer le seuil de décision de la distance de Levenshtein sur les données de développement.
Le seuil de décision de la distance Levenshtein a été fixé à 0,65 pour les modèles multi-Gaussiens pour être sûr d'identifier toutes les publicités [22]. Afin d'évaluer les performances de notre système d'identification, les mesure de précision (P%) et rappel (R%) sont utilisées et exposées dans le Tableau 1 :

- Précision : Le nombre de publicités correctement détectées / nombre total de publicités détectées.
- Rappel : Le nombre de publicités correctement détectées / Le nombre de publicités qui doivent être détectées.

Le Tableau 1 montre que le système n'était pas en mesure d'identifier 288 publicités. En analysant ces erreurs, nous avons trouvé deux types d'erreurs. Le premier concerne les publicités très courtes (3 à 5 secondes) ayant des références décalées par deux ou trois secondes. Le second type d'erreur est lié à des publicités qui étaient différentes de leurs références, les annotations manuelles de ces publicités étaient incorrectes.
En corrigeant les références décalées, notre système a pu identifier 4177 publicités sur 4401. Les 224 publicités non détectées étaient différentes de leurs références. Nous notons la présence de 18 fausses alarmes. En fait, le système a confondu certaines publicités des mêmes produits mais avec un léger changement dans leur contenu et qui ont été annotées différemment par des annotateurs humains.

Nb de pubs	Pubs non identifiées	Fausses alarmes	P%	R%
4401	288	18	99	93

Tableau 1 – *Précision (P%), rappel (R%), nombre des publicités non identifiées et nombre des fausses alarmes calculés pour 7 jours de diffusion radiophonique pour 11 radios*

4.4 Application à l'identification des morceaux de musique

Comme l'on a mentionné auparavant, notre approche d'identification audio est une approche générique qui pourrait être appliquée à la fois à l'identification des publicités et des morceaux de musique. Par conséquent, nous avons décidé de garder la même configuration et le même seuil de la distance de Levenshtein utilisés pour l'identification des publicités.
Afin d'évaluer les performances de notre système d'identification, les mesure de précision (P%) et rappel (R%) sont utilisées. Ces mesures sont exposées dans le Tableau 2.
Le Tableau 2 montre que le système basé sur ALISP n'était pas capable de détecter 46 morceaux de musique. Ces morceaux sont liés à des chansons qui ont une version différente de celle présente dans la base de données de référence. Par exemple, nous avons trouvé 32 morceaux de musique interprétés en direct dans le flux radio, tandis que les références associées sont interprétées en version studio. 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.

Nb de morceaux	Morceaux non identifiés	Fausses alarmes	P%	R%
505	46	0	100	91

Tableau 2 – *Precision (P%), rappel (R%), nombre des publicités non identifiées et nombre des fausses alarmes calculés pour 7 jours de diffusion radiophonique pour la radio RTL*

Dans Fenet et al. [7], qui utilisent le même protocole, la reconnaissance d'interprétations différentes du même titre est considérée comme hors du périmètre de l'identification audio. Par conséquent, en éliminant les 46 morceaux de musique qui ont une version différente de celle de la référence, le système décrit dans [7] obtient un taux de rappel de 97,4% avec 0 fausses alarmes, alors qu'avec notre système basé sur ALISP le taux de rappel et le taux de précision obtenus sont de 100%.
Pour le temps du traitement, l'acquisition et la modélisation des unités ALISP se fait hors ligne. Lors du traitement des données de test, le temps de traitement nécessaire pour

transcrire les flux audio avec des modèles ALISP est négligeable. La complexité de calcul du système est actuellement limitée par la recherche de la plus proche séquence ALISP avec la distance de Levenshtein. Avec l'implémentation actuelle, le temps de traitement requis par le système pour la recherche de nos morceaux de musique de référence (7309 morceaux) dans une heure de flux radio est de quatre heures avec une machine 3.00GHz Intel Core 2 Duo 4 Go de RAM.

5 Conclusions et perspectives

Dans cet article nous avons présenté un système d'identification audio générique pour trouver et détecter des publicités et des morceaux de musique dans les flux radiophoniques en utilisant des unités acoustiques. Ces unités sont déterminées de manière entièrement automatique grâce aux outils ALISP. Pour la tâche de l'identification des publicités nous avons eu un taux de précision de 99% et un taux de rappel de 95%, sachant que les publicités non détectées sont différentes de leurs références. Pour la tâche de l'identification des morceaux de musique nous avons obtenu un taux de précision et de rappel de 100% en suivant le protocole Quaero.

Les travaux futurs seront consacrés à améliorer la méthode de recherche des transcriptions ALISP, qui est très élémentaire, en adaptant l'algorithme BLAST (Basic Local Alignment Search Tools) [23] qui est souvent utilisé pour comparer des séquences biologiques. Cette technique trouve des régions de similarité locale entre deux séquences. Cette technique compare des séquences de nucléotides ou de protéines aux bases de données de référence. Cette méthode pourrait être utilisée pour accélérer la comparaison entre les séquences ALISP extraites du flux radio et transcriptions ALISP stockées dans la base de référence.

De plus d'autres travaux seront menés pour déterminer le nombre optimal des modèles ALISP, pour le moment le nombre de symboles ALISP utilisées est de 64 mais d'autres modèles HMM seront appris avec 4, 16 et 32 symboles ALISP.

Remerciements

Ces travaux sont menés dans le cadre du projet ANR-SurfOnHertz. Certains des programmes utilisés dans ces travaux ont été fournit par F. Bimbot, J. Cernocky, A. El Hannani, G. Aversano et les membres du projet SYMPATEX.

Références

[1] Pedro Cano, Eloi Batlle, Ton Kalker, et Jaap Haitsma. A review of audio fingerprinting. *J. VLSI Signal Process. Syst.*, 41(2) :271–284, Novembre 2005.

[2] Gérard Chollet, Jan Cernockỳ, Andrei Constantinescu, Sabine Deligne, et Frederic Bimbot. *Towards ALISP : a proposal for Automatic Language Independent Speech Processing*, pages 375–387. 1999.

[3] Jaap Haitsma et Ton Kalker. A highly robust audio fingerprinting system. Dans *ISMIR*, pages 107–115, 2002.

[4] Yu Liu, Kiho Cho, Hwan Sik Yun, Jong Won Shin, et Nam Soo Kim. Dct based multiple hashing technique for robust audio fingerprinting. Dans *ICASSP*, pages 61–64, 2009.

[5] Jaap Haitsma et Ton Kalker. Speed-change resistant audio fingerprinting using auto-correlation. Dans *ICASSP*, pages 728–731, 2003.

[6] Avery Wang. The shazam music recognition service. *Commun. ACM*, 49(8) :44–48, Août 2006.

[7] Sébastien Fenet, Yves Grenier, et Gael Richard. Une empreinte audio à base de cqt appliquée à la surveillance de flux radiophoniques. Dans *GRETSI*, 2011.

[8] Shumeet Baluja et Michele Covell. Content fingerprinting using wavelets. Dans *CVMP*, pages 198–207, 2006.

[9] Yan Ke, Derek Hoiem, et Rahul Sukthankar. Computer vision for music identification. Dans *CVPR*, pages 597–604, 2005.

[10] Paul Viola et Michael Jones. Rapid object detection using a boosted cascade of simple features. Dans *CVPR*, pages 511–518, 2001.

[11] Martin Fischler et Robert Bolles. Readings in computer vision : issues, problems, principles, and paradigms. chapitre Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography, pages 726–740. 1987.

[12] Eric Allamanche, Jurgen Herre, Oliver Hellmuth, Bernhard Froba, et Markus Cremer. Audioid : Towards content-based identification of audio material. Dans *Audio Engineering Society Convention 110*, 2001.

[13] Pedro Cano, Eloi Batlle, Harald Mayer, et Helmut Neuschmied. Robust sound modeling for song detection in broadcast audio. Dans *in Proc. AES 112th Int. Conv*, pages 1–7, 2002.

[14] William Pearson et David Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8) :2444–2448, Avril 1988.

[15] Asmaa ElHannani, Dijana Petrovska-Delacrétaz, Benoît Fauve, Aurélien Mayoue, John Mason, Jean-François Bonastre, et Gérard Chollet. Text-independent speaker verification. Dans *Guide to Biometric Reference Systems and Performance Evaluation*, pages 167–211. 2009.

[16] Gérard Chollet, Kevin McTait, et Dijana Petrovska-Delacrétaz. Data driven approaches to speech and language processing. Dans *Nonlinear Speech Modeling and Applications*, volume 3445 de *Lecture Notes in Computer Science*, pages 164–198. 2005.

[17] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Dans *Cybernetics and control theory*, pages 707–710, 1966.

[18] Frederic Bimbot. An evaluation of temporal decomposition. Rapport technique, Acoustic Research Department AT&T Bell Labs, 1990.

[19] Bishnu Atal. Efficient coding of lpc parameters by temporal decomposition. Dans *ICASSP*, pages 81–84, 1983.

[20] Yoseph Linde, Andres Buzo, et Robert M. Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1) :84–95, Janvier 2003.

[21] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, et Phil Woodland. The htk book. Rapport technique, Entropics Cambridge Research Lab, 1996.

[22] Houssemeddine Khemiri, Gérard Chollet, et Dijana Petrovska-Delacrétaz. Automatic detection of known advertisements in radio broadcast with data-driven alisp transcriptions. *Multimedia Tools and Applications*, pages 1–15, 2012.

[23] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, et David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215 :403–410, Mai 1990.