THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité
Electronique

Présentée par
Thomas HUEBER

Pour obtenir le grade de
DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse
Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal : vers une communication parlée silencieuse

soutenue le 9 décembre 2009

devant le jury composé de

M. Gérard Bailly	Directeur de recherche au CNRS	Rapporteurs
M. Thierry Dutoit	Professeur	
M. Thierry Artières	Professeur	Examinateurs
M. Olivier Boëffard	Professeur	
Mme. Lise Crevier-Buchman	Médecin ORL - Phoniatre Chargée de Recherche au CNRS	
M. Gérard Dreyfus	Professeur	
M. Bruce Denby	Professeur	Directeurs de thèse
M. Gérard Chollet	Directeur de recherche au CNRS	

Résumé

L'objectif poursuivi dans ce travail de thèse est la réalisation d'un dispositif capable d'interpréter une parole normalement articulée mais non vocalisée, permettant ainsi la « communication parlée silencieuse ». Destiné, à terme, à être léger et portatif, ce dispositif pourrait être utilisé d'une part, par une personne ayant subi une laryngectomie (ablation du larynx suite à un cancer), et d'autre part, pour toute communication, soit dans un milieu où le silence est requis (transport en commun, opération militaire, etc.), soit dans un environnement extrêmement bruité. Le dispositif proposé combine deux systèmes d'imagerie pour capturer l'activité de l'appareil vocal pendant « l'articulation silencieuse » : l'imagerie ultrasonore, qui donne accès aux articulateurs internes de la cavité buccale (comme la langue), et la vidéo, utilisée pour capturer le mouvement des lèvres. Le problème traité dans cette étude est celui de la synthèse d'un signal de parole « acoustique », uniquement à partir d'un flux de données « visuelles » (images ultrasonores et vidéo). Cette conversion qualifiée ici de « visuo-acoustique », s'effectue par apprentissage artificiel et fait intervenir quatre étapes principales : l'acquisition des données audiovisuelles, leur caractérisation, l'inférence d'une cible acoustique à partir de l'observation du geste articulatoire et la synthèse du signal.

Dans le cadre de la réalisation du dispositif expérimental d'acquisition des données, un système de positionnement de la sonde ultrasonore par rapport à la tête du locuteur, basé sur l'utilisation combinée de deux capteurs inertiaux a tout d'abord été conçu. Un système permettant l'enregistrement simultané des flux visuels et du flux acoustique, basé sur la synchronisation des capteurs ultrasonore, vidéo et audio par voie logicielle, a ensuite été développé. Deux bases de données associant observations articulatoires et réalisations acoustiques, contenant chacune environ une heure de parole (continue), en langue anglaise, ont été construites. Pour la caractérisation des images ultrasonores et vidéo, deux approches ont été mises en œuvre. La première est basée sur l'utilisation de la transformée en cosinus discrète, la seconde, sur l'analyse en composantes principales (approche *EigenTongues/EigenLips*). La première approche proposée pour l'inférence des paramètres acoustiques, qualifiée de « directe », est basée sur la construction d'une « fonction de conversion » à l'aide d'un réseau de neurones et d'un modèle par mélange de gaussiennes. Dans une seconde approche, qualifiée cette fois « d'indirecte », une étape de décodage des flux visuels au niveau phonétique est introduite en amont du processus de synthèse. Cette étape intermédiaire permet notamment l'introduction de connaissances linguistiques *a priori* sur la séquence observée. Elle s'appuie sur la modélisation des gestes articulatoires par des modèles de Markov cachés (MMC). Deux méthodes sont enfin proposées pour la synthèse du signal à partir de la suite phonétique décodée. La première est basée sur une approche par concaténation d'unités ; la seconde utilise la technique dite de « synthèse par MMC ». Pour permettre notamment la réalisation d'adaptations prosodiques, ces deux méthodes de synthèse s'appuient sur une description paramétrique du signal de parole du type «Harmonique plus Bruit» (HNM).

Mots clés

parole silencieuse, communication parlée augmentée, imagerie ultrasonore, vidéo, capteurs, traitement du signal appliqué, modélisation par apprentissage, fusion de données, multimodalité, reconnaissance, synthèse, modèle de Markov caché, mélange de gaussiennes, réseau de neurones, image, systèmes homme-machine, laryngectomie, cancer, handicap, langue, lèvres, articulateurs.

Abstract

The aim of the thesis is the design of a "silent speech interface", or system permitting voice communication without vocalization. Two main applications are targeted: assistance to laryngectomized persons; and voice communication when silence must be maintained (public transport, military situation) or in extremely noisy environments. The system developed is based on capturing articulatory activity via ultrasound and video imaging. The problem addressed in this work is that of transforming multimodal observations of articulatory gestures into an audio speech signal. This "visuo-acoustic" conversion is achieved using machine learning methods requiring the construction of audiovisual training databases. To this end, in order to monitor the position of the ultrasound probe relative to the speaker's head during data acquisition, a procedure based on the use of two inertial sensors is first proposed. Subsequently, a system allowing to synchronously acquire high-speed ultrasound and video images of the vocal tract together with the uttered acoustic speech signal is presented. Two databases containing approximately one-hour of multimodal continuous speech data (in English) were recorded. Discrete cosine transform (DCT) and principal component analysis (*EigenTongues/EigenLips* approach) are then compared as techniques for visual feature extraction. A first approach to visuo-acoustic conversion is based on a direct mapping between visual and acoustic features using neural networks and Gaussian mixture models (GMM). In a second approach, an intermediate HMM-based phonetic decoding step is introduced, in order to take advantage of *a priori* linguistic information. Finally, two methods are compared for the inference of the acoustic features used in the speech synthesis step, one based on a unit selection procedure, and the second invoking HMMs (HMM-based synthesis system HTS), with the "Harmonic plus Noise" model (HNM) of the speech signal being used in both approaches.

Keywords

silent speech, speech recognition, speech synthesis, concatenative synthesis, machine learning, human-computer interface, inertial sensors, ultrasound, video, multimodal, data fusion, laryngectomy, tongue, lips, *Ultraspeech*, *Ouisper*, *EigenTongues*, HMM, GMM, HTS, HNM.

Remerciements

Je tiens en tout premier lieu à remercier Monsieur Bruce Denby, acteur majeur de la recherche sur les « interfaces de communication en parole silencieuse » et directeur de cette thèse. Grâce à lui, j'ai eu la chance d'évoluer, pendant ces trois dernières années, dans un environnement scientifique stimulant tout en profitant d'une grande liberté dans l'orientation de mon travail. Sur le plan scientifique comme humain, je le remercie pour son écoute attentive, ses conseils avisés et sa grande disponibilité.

Je souhaite ensuite remercier chaleureusement Monsieur Gérard Chollet, également directeur de ce travail de thèse. Ses nombreux conseils et sa très grande expérience dans le domaine du traitement de la parole m'ont beaucoup apporté. Je le remercie vivement pour la confiance qu'il m'a témoignée et le regard attentif qu'il n'a cessé de porter sur mon travail.

Dans le cadre ma thèse, j'ai eu la grande chance de travailler avec le Professeur Maureen Stone, directrice du *Vocal Tract Visualization Laboratory* de l'Université du Maryland (USA) et pionnière de l'étude du conduit vocal par imagerie ultrasonore. Nos échanges réguliers ainsi que mon séjour au sein de son équipe m'ont beaucoup apporté et je souhaite la remercier vivement pour cela.

Je tiens ensuite à exprimer ma gratitude aux différents membres du jury. Je remercie Messieurs Gérard Bailly et Thierry Dutoit pour avoir accepté de rapporter sur ce travail, ainsi que Madame Lise Crevier-Buchman, Monsieur Thierry Artières et Monsieur Olivier Boëffard pour l'avoir examiné.

Je tiens également à remercier Monsieur Gérard Dreyfus, directeur du Laboratoire d'Electronique de l'ESPCI ParisTech, pour m'avoir accueilli dans son laboratoire ainsi que pour son accompagnement et ses conseils.

Je souhaite remercier la Délégation Générale pour l'Armement (DGA) qui a financé cette thèse et tout particulièrement Mme Couesnon pour sa grande disponibilité et son efficacité.

Je tiens à remercier la direction des relations industrielles et du transfert technologique (DRITT) de l'Université Pierre et Marie Curie (UPMC) et plus particulièrement Monsieur Arnaud Boissière, pour avoir organisé la valorisation d'une partie de ce travail de thèse.

J'ai eu la chance d'effectuer en parallèle de cette thèse un monitorat en Electronique à l'UPMC. Cette activité d'enseignement a contribué à rendre ces trois années de thèse très épanouissantes. Je souhaite remercier vivement tous les étudiants à qui j'ai eu la chance d'enseigner, ainsi que différents collègues, qui m'ont donné le goût de leur métier. Un grand merci donc à Olivier Romain, Nicolas Bovo, Christophe Martin, et Benoît Fabre.

Je tiens à remercier très sincèrement les différents locuteurs qui ont participé aux campagnes d'acquisition de données. Un grand merci donc à Stacy, Greg et Sophie.

Je tiens à remercier Madame Lise Crevier-Buchan (LPP, Paris), Monsieur Yves Laprie (LORIA Nancy), Monsieur Gérard Bailly (GIPSA, Grenoble) et Monsieur Laurent Miclet

(IRISA, Lannion) de m'avoir donné l'opportunité de présenter mon travail à l'occasion de séminaires dans leurs laboratoires respectifs.

Je souhaite également remercier Monsieur Alain Marchal et Monsieur Christian Cavé pour m'avoir invité à participer à la rédaction de leur ouvrage (intitulé « Imagerie médicale pour l'étude de la parole »).

Durant la seconde moitié de ma thèse, j'ai eu la chance de travailler au coté de Elie-Laurent Benaroya, post-doctorant sur le projet *Ouisper*. Je tiens tout particulièrement à le remercier pour la qualité de nos échanges, sa rigueur scientifique, et son aide si précieuse.

Pour leur soutien constant, leur écoute et leur bonne humeur quotidienne, je souhaite également remercier très sincèrement tous les autres membres du laboratoire d'Electronique de l'ESPCI ParisTech et en particulier : Pierre Roussel dont la disponibilité, la compétence et la qualité d'écoute m'ont tant apporté, Rémi Dubois pour son dynamisme, sa rigueur et son aide précieuse, Arthur Duprat, chimiste-informaticien de talent, pour son soutien logistique sans faille, Paola Milpied Bouchet pour nos mémorables pauses-café qui ont rythmées nos après-midi de thésards, et enfin Iness Ahriz, ma « colocatrice » d'exception pour sa gentillesse et son sourire quotidien.

Parce qu'ils représentent tant pour moi et en raison de leur soutien qui, pour certains d'entres-eux, dure déjà depuis de nombreuses années, je profite de cet espace quelque peu « solennel » pour remercier quelques amis. Un grand merci donc à Edgar, Gilbert, FX, Jef, Vincent, Tonio, Ben, Matthieu, Jean-Marie, Bati, Jean, Florent, Pierre, Amandine, Baptiste, Julie, Gérald, Joseph, Jeanne, Thomas, Valérie, Philipe, Anaïs et Greg.

Je tiens évidemment à remercier ma famille qui m'a accompagné et soutenu pendant tout mon parcours d'études dont ce travail est, d'une certaine manière, une forme d'aboutissement. Parce que je puise en eux et en leurs attentes une grande partie de ma motivation, je tiens à associer tout particulièrement ma mère et mon frère à ce travail.

Enfin, parce que quelques phrases ne suffiraient pas pour lui montrer tout ce qu'elle m'apporte, je terminerai simplement cette page en la remerciant tendrement. Merci Laure …

Tables des matières

Table des figures

Table des tableaux

Acronymes

Français

ICPS : Interface de communication en parole silencieuse
MMC : Modèle de Markov Caché
EMG : Electromyographie
TCD : Transformée en Cosinus Discrète
TFT : Transformée de Fourier Discrète

Anglais

LPC : *Linear Predictive Coding*
LSF : *Line Spectrum Frequencies*
GMM : *Gaussian Mixture Model*
HMM : *Hidden Markov Model*
HNM : *Harmonic plus Noise Model*
MFCC : *Mel Frequency Cepstral Coefficients*
NAM : *Non-Audible Murmur*

Introduction générale

Motivation

Support pour le transport de l'information, vecteur de notre identité et témoin de nos émotions, la communication parlée est l'interface majeure des interactions humaines. Plus qu'un simple moyen de porter un message, la voix est un outil privilégié pour l'expression et l'affirmation de soi, pour convaincre, séduire ou entraîner l'adhésion de l'auditeur. Ainsi, les troubles et maladies de la voix sont à l'origine d'importants handicaps, sur le plan physique, comme sur le plan social. Les pathologies de l'appareil vocal sont multiples. Certaines, parmi les plus répandues, sont relatives à l'appareil vibrateur (ou excitateur), constitué principalement du larynx et des cordes vocales (plis vocaux). D'autres concernent les articulateurs, ces éléments mobiles comme la langue et les lèvres qui, en modifiant la géométrie de notre « caisse de résonnance », déterminent le son qui sera prononcé. Une des pathologies les plus connues est la laryngite aiguë (inflammation des cordes vocales), qui peut notamment engendrer ce que l'on nomme communément « l'extinction de voix », tant redoutée par l'enseignant ou le chanteur professionnel car pouvant parfois aller jusqu'à l'aphonie totale. Si dans ce cas le trouble ne persiste généralement que quelques jours puis disparaît totalement, d'autres pathologies, plus lourdes, sont à l'origine de dommages bien plus importants sur notre appareil vibrateur. C'est le cas de certaines formes du cancer du larynx dont le traitement, qui doit s'effectuer par voie chirurgicale, peut comporter une laryngectomie, c'est-à-dire l'ablation partielle ou totale de l'organe et donc des cordes vocales qu'il abrite. Dépourvu de vibrateur laryngé, le patient ainsi traité, bien que toujours capable d'articuler, ne peut plus parler normalement. Il doit apprendre à maitriser une des techniques de vocalisation palliatives aujourd'hui disponibles, qui pour la plupart, tentent de recréer, par voie chirurgicale ou à l'aide d'une prothèse mécanique externe, un vibrateur de substitution. Si ces techniques permettent de retrouver la capacité de parler, leur mise en œuvre peut s'avérer difficile, leur apprentissage long, et la parole produite, bien qu'intelligible, d'une qualité limitée (voix rauque et irrégulière). Ce constat est le premier point de départ d'une réflexion autour de la recherche de nouvelles solutions et la mise au point de « prothèses » plus performantes. C'est dans cette première perspective que s'inscrit cette thèse.

Ce travail s'inscrit par ailleurs dans le récent domaine de recherche sur la conception d'une « interface de communication en parole silencieuse » (ICPS). Il s'agit ici de réaliser un système de communication parlée, capable de saisir et d'interpréter la production d'une parole normalement articulée mais non vocalisée (articulation silencieuse). Destiné à terme à être portatif, ce type de dispositif pourrait permettre la réalisation d'un « téléphone silencieux ». Cet outil serait utilisable par une personne handicapée de la parole ou non, dans le monde civil comme militaire, pour communiquer sans perturber son environnement (transport en commun, réunion, etc.), ou dans un environnement « sensible » où la discrétion est primordiale et le silence requis. Le champ applicatif de ce type d'interface s'étend également à la communication

dans un environnement extrêmement bruyant (hélicoptère, char d'assaut, concert) pour lequel la modalité acoustique, trop bruitée, n'est pas un support exploitable pour un système de traitement automatique de la parole.

Contexte technologique et objectif poursuivi

Une interface de communication en parole silencieuse est généralement basée sur l'analyse de l'activité de l'appareil vocal pendant l'articulation. Si l'activité des articulateurs « externes » que sont les lèvres et la mâchoire peut être simplement captée à l'aide d'une caméra vidéo, celle des articulateurs internes à la cavité buccale, est plus difficile. L'imagerie médicale est aujourd'hui régulièrement utilisée en phonétique expérimentale pour la visualisation des mouvements articulatoires et l'étude des mécanismes qui régissent la production de la parole. Une des techniques privilégiées est l'imagerie ultrasonore, ou échographie. Non invasive, totalement inoffensive et intégrable dans un dispositif de taille restreint, elle permet notamment, lorsque le transducteur (la sonde) est placé sous la mâchoire, la visualisation de l'un des articulateurs majeurs, la langue. De plus, l'intégration de la modalité visuelle dans le traitement automatique de la parole est, ces dernières années, au cœur de nombreuses études sur les systèmes de dialogue « homme-machine ». Les nombreux travaux en reconnaissance « audio-visuelle » de la parole ont montré que la robustesse au bruit d'un système de reconnaissance automatique de la parole (transcription) peut être améliorée en s'appuyant, en complément du signal acoustique, sur l'interprétation du mouvement des lèvres. Par ailleurs, les systèmes de dialogue « homme-machine » bénéficient aujourd'hui de techniques de synthèse de parole performantes qui produisent un signal d'une intelligibilité et d'une qualité très acceptable.

Ainsi, c'est dans ce contexte technologique qui combine analyse des mouvements articulatoires par imagerie, approche multimodale du traitement automatique de la parole et techniques robustes de synthèse, que s'inscrit ce travail de recherche sur la conception d'un « système de reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal ». L'objectif qui est poursuivi dans cette thèse, est d'évaluer la faisabilité d'un système capable de synthétiser un signal de parole intelligible, uniquement à partir de l'observation des mouvements articulatoires par imagerie ultrasonore et vidéo.

Une parole silencieuse

En articulation silencieuse et *a fortiori* chez une personne laryngectomisée, l'appareil vibrateur (et donc les cordes vocales) n'est pas actif ; les éléments constitutifs du résonateur, à savoir les articulateurs de la cavité bucco-pharyngale (langue, lèvres, vélum, etc.), sont donc les seuls « témoins » de la production de la parole. Des informations majeures telles que la caractéristique de voisement, les variations segmentales et suprasegmentales de la fréquence fondamentale (accentuation et intonation globale) et l'énergie, ne sont pas « directement » disponibles en articulation silencieuse. Une parole articulée mais non vocalisée est donc une

parole « incomplète ». Ce constat sera la motivation principale d'une approche d'interprétation des mouvements articulatoires pouvant intégrer des informations sur le contenu linguistique *a priori* de la séquence observée.

Une approche par apprentissage

Au sein de la communauté de la recherche sur la parole, la synthèse du signal acoustique à partir des positions articulatoires est traditionnellement envisagée dans le cadre de la « synthèse articulatoire ». Cette approche vise à modéliser explicitement les phénomènes physiques d'écoulement de l'air et donc de propagation de l'onde de pression acoustique dans le conduit vocal. Les articulateurs tout comme l'appareil excitateur, sont alors décrits par des modèles physiologiques (physiques, géométriques ou plus récemment biomécaniques) dont les paramètres sont directement reliés à l'évolution de leur structure au cours de la production. La mise en œuvre de ces modèles est une tâche particulièrement délicate, les équations qui les régissent étant très complexes. Dans les récents travaux sur la synthèse articulatoire, la construction de modèles précis s'appuie notamment sur l'analyse du conduit vocal par IRM (imagerie par résonnance magnétique). Dans le cadre de notre étude, la mise en œuvre de modèles articulatoires à partir des données ultrasonores et vidéo n'est pas envisagée.

Par opposition à ces techniques qui modélisent la « physique » des mécanismes de production de la parole, l'approche développée dans le cadre de ce travail pourra être qualifiée d'approche « par apprentissage ». La méthodologie proposée utilise en effet l'apprentissage artificiel et la modélisation probabiliste pour lier l'observation visuelle d'un mouvement articulatoire à la séquence de paramètres acoustiques la plus adaptée. Pour réaliser cette conversion dite « visuo-acoustique », deux approches sont proposées. La première, qualifiée de directe, est basée sur l'apprentissage d'une fonction de transformation liant l'espace visuel (observations articulatoires) et l'espace acoustique (contenu spectral du signal de parole correspondant). Cette fonction est construite à l'aide d'une modélisation par réseau de neurones et par mélange de gaussiennes. La seconde approche proposée, qualifiée d'indirecte, introduit en amont du processus de synthèse, une étape de décodage segmental, qui tente d'identifier dans le flux visuel continu, une suite discrète de gestes articulatoires. Cette étape permet l'introduction éventuelle de connaissances linguistiques *a priori* sur la séquence articulatoire observée. Elle s'appuie sur la segmentation automatique des flux visuels et audio au niveau phonétique et sur la modélisation des classes de segments ainsi trouvées par des modèles de Markov cachés (MMC). Deux méthodes sont proposées pour l'inférence des paramètres acoustiques. La première utilise une approche par concaténation d'unités ; la seconde se base sur les techniques dites de « synthèse stochastique » ou « synthèse par MMC ». Pour permettre notamment la réalisation d'adaptations prosodiques, ces deux méthodes s'appuient sur une description paramétrique du signal de parole du type «Harmonique plus Bruit».

La mise en œuvre de cette « approche par apprentissage » nécessite :

- la constitution de bases de données d'apprentissage qui associent observations articulatoires et réalisations acoustiques. Ceci requiert notamment l'élaboration d'un dispositif expérimental permettant l'acquisition synchrone de données ultrasonores, vidéo et audio.
- le traitement et le codage des données audiovisuelles acquises : il s'agit de l'étape d'extraction des caractéristiques visuelles et acoustiques.
- la conversion dite « visuo-acoustique » directe ou indirecte, qui permet le passage de la modalité visuelle (données articulatoires ultrasonores et vidéo) à la modalité audio, rendant ainsi possible la synthèse du signal de parole.

Ces différentes étapes, qui ont toutes été abordées dans le cadre de ce travail, structurent l'organisation de ce document.

Organisation du document

En s'appuyant sur une brève description du fonctionnement de l'appareil vocal, le premier chapitre aborde tout d'abord le problème de la production vocale chez le patient laryngectomisé. Une synthèse des différentes approches proposées dans la littérature, pour la réalisation « d'interfaces de communication en parole silencieuse , est ensuite effectuée. L'approche basée sur l'imagerie ultrasonore et vidéo, c'est-à-dire celle adoptée dans ce travail de thèse, est enfin introduite.

Le second chapitre est dédié à l'acquisition des données audiovisuelles. Après une brève introduction à l'imagerie ultrasonore, nous décrivons les spécificités de l'utilisation de cette technique pour l'étude de l'appareil vocal. Les dispositifs expérimentaux mis en œuvre pour l'acquisition synchrone des flux de données ultrasonores, vidéo, et audio ainsi que les bases de données construites, sont ensuite présentés.

Les différentes techniques mises en œuvre pour le traitement des données acquises font l'objet du troisième chapitre. Une technique de réduction du bruit dans les images ultrasonores est tout d'abord présentée. Différentes procédures d'extraction de caractéristiques visuelles sont ensuite discutées. Enfin, deux procédures d'analyse-synthèse du signal de parole sont décrites : la décomposition cepstrale et la décomposition « Harmonique plus Bruit ».

Les deux approches proposées pour effectuer la conversion visuo-acoustique, qualifiées respectivement d'approche directe et d'approche indirecte, font l'objet du quatrième et du cinquième chapitre.

Notations phonétiques

Dans le cadre des expérimentations effectuées en langue anglaise, le format TIMIT est utilisé dans ce document pour la transcription des symboles de l'alphabet phonétique

internationale (IPA). Une table de conversion « IPA – TIMIT » est disponible dans (Hieronymus, 1993).

Exemples sonores et vidéo

A ce manuscrit est associée une page Internet sur laquelle sont accessibles divers exemples sonores et vidéo. Cette page est disponible à l'adresse suivante : http://hueber.thomas.free.fr/these/.

Contexte

Ce travail de doctorat, financé par la Délégation Générale pour l'Armement (DGA), s'inscrit dans le cadre du projet de recherche *Ouisper*[1], soutenu par l'Agence Nationale de la Recherche (ANR-06-BLAN-0166). Il a été réalisé au sein du « Laboratoire d'Electronique » (CNRS, UMR 7084) de l'École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI ParisTech) et au Laboratoire Traitement et Communication de l'Information (LTCI, CNRS, UMR 5141) de Telecom ParisTech, du 1er octobre 2006 au 1er décembre 2009. Ce travail fait également l'objet d'une collaboration avec le Dr. Maureen Stone, directrice du *Vocal Tract Visualization Laboratory* de l'Université de Maryland (USA).

[1] *OUISPER* : *Oral Ultrasound synthetIc SPEech souRce*

Chapitre 1. Vers une communication parlée silencieuse, état de l'art

1.1. Architecture et fonctionnement de l'appareil vocal

Cette section, qui rappelle l'architecture et les principes généraux de fonctionnement de notre appareil vocal s'appuie sur les ouvrages suivants : (Le Huche, 2001) et (Boite *et al.*, 2000). Une vue schématique de notre appareil vocal est proposée à la Figure 1.1.

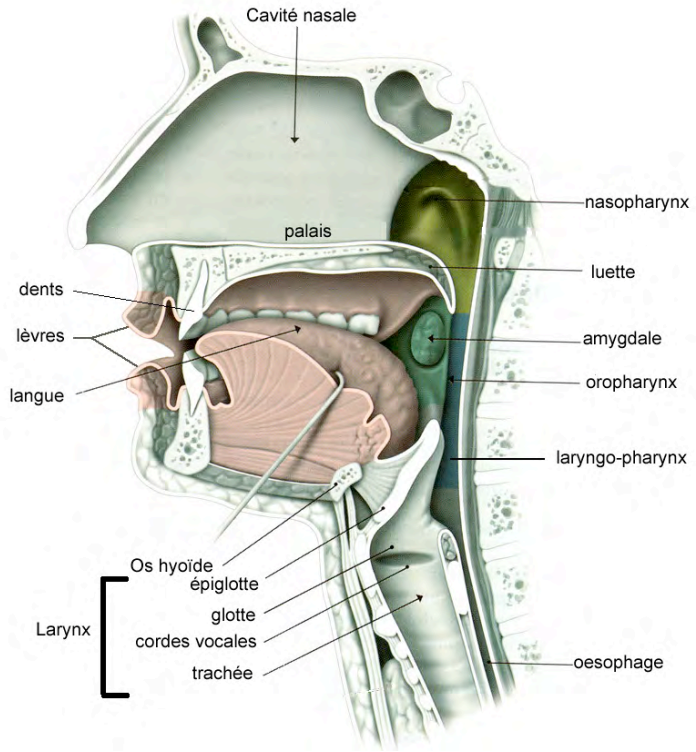


Figure 1.1 : Vue schématique de l'appareil vocal[2], dans le plan sagittal médian

1.1.1. L'appareil vibrateur

L'air est la matière première de la voix. Si le fonctionnement de notre appareil vocal est souvent comparé à celui d'un instrument de musique, il doit être décrit comme celui d'un instrument à vent. En effet, en expulsant l'air pulmonaire à travers la trachée, le système respiratoire joue le rôle d'une soufflerie. Il s'agit du « souffle phonatoire » produit, soit par l'abaissement de la cage thoracique, soit dans le cadre de la projection vocale[3] par l'action des muscles abdominaux.

L'extrémité supérieure de la trachée est entourée par un ensemble de muscles et de cartilages mobiles qui constituent le larynx. Le plus important est le cartilage thyroïde qui

[2] Illustration extraite de http://lecerveau.mcgill.ca (sous *copyleft)*

[3] Définie comme étant une production vocale destinée à agir efficacement sur autrui (appeler quelqu'un, ordonner, etc.)

forme le relief de la pomme d'Adam. Le larynx se trouve au carrefour des voies aériennes et digestives, entre le pharynx et la trachée, et en avant de l'œsophage. Les plis vocaux, communément nommés « cordes vocales », sont deux lèvres symétriques (structures fibreuses) placées en travers du larynx. Ces lèvres se rejoignent en avant et sont plus au moins écartées l'une de l'autre sur leur partie arrière (structure en forme de V) ; l'ouverture triangulaire résultante est nommée glotte. Les structures du larynx et des plis vocaux sont illustrées à la Figure 1.2.

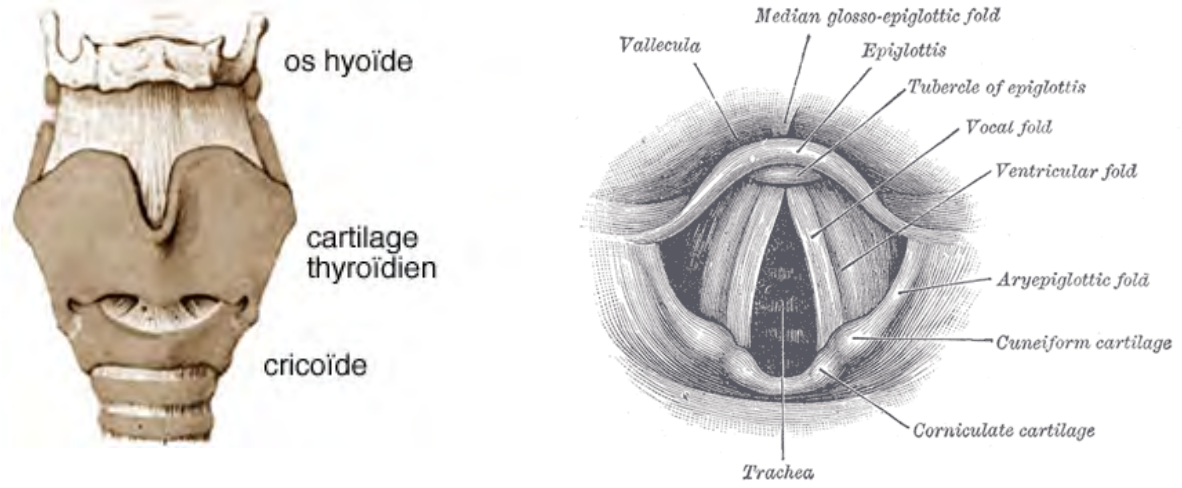


Figure 1.2 : Vue schématique antérieure du larynx[4] (à gauche). Vue laryngoscopique des cordes vocales[5] (à droite).

Le larynx et les plis vocaux forment notre « appareil vibrateur ». Lors de la production d'un son qualifié de « non-voisé » (ou sourd), comme c'est le cas, par exemple, pour les phonèmes [s] ou [f], les plis vocaux sont écartés et l'air pulmonaire circule librement en direction des structures en aval. En revanche, lors de la production d'un son voisé (ou sonore), comme c'est le cas, par exemple, pour les phonèmes [z], [v] et pour les voyelles, les plis vocaux s'ouvrent et se ferment périodiquement, obstruant puis libérant par intermittence le passage de l'air dans le larynx. Le flux continu d'air pulmonaire prend ainsi la forme d'un train d'impulsions de pression ; nos « cordes vocales vibrent ». Le dernier élément principal de notre appareil vibrateur est l'épiglotte. Lors de la déglutition, cette dernière agit comme un clapet qui se rabat sur le larynx, conduisant les aliments vers l'œsophage en empêchant leur passage dans la trachée et les poumons (« fausse route »).

1.1.2. Le résonateur

L'air pulmonaire, ainsi modulé par l'appareil vibrateur, est ensuite appliqué à l'entrée du conduit vocal. Ce dernier est principalement constitué des cavités pharyngiennes (laryngophaynx

[4] Extrait du site Internet http://www.infocancer.org

[5] Extrait de (Gray, 1973). Malgré la présence d'une légende en langue anglaise, cette illustration est choisie car libre de droit.

et oropharynx situés en arrière-gorge) et de la cavité buccale (espace qui s'étend du larynx jusqu'aux lèvres). Pour la réalisation de certains phonèmes, le voile du palais (le velum) et la luette qui s'y rattache, s'abaissent, permettant ainsi le passage de l'air dans les cavités nasales (fosses nasales et rhinopharynx ou nasopharynx). Ces différentes cavités forment un ensemble que nous qualifierons ici de « résonateur ». Si l'appareil vibrateur peut être décrit comme le lieu de production de « la voix », le résonateur apparaît alors comme le lieu de naissance de « la parole ». Il abrite en effet des organes mobiles, nommés articulateurs, qui en modifiant sa géométrie et donc ses propriétés acoustiques, mettent en forme le son laryngé (ou son glottique) en une séquence de sons élémentaires. Ces derniers peuvent être interprétés comme la réalisation acoustique d'une série de phonèmes, unités linguistiques élémentaires propres à une langue. Les articulateurs principaux sont la langue, les lèvres, le voile du palais et la mâchoire (maxillaire inférieur)[6].

L'articulateur principal de la cavité buccale est la langue. Intervenant dans la mastication et la déglutition, la langue est également l'organe du goût. S'étendant sur une longueur d'une dizaine de centimètres environ, cet organe complexe et hautement vascularisé est composé d'un squelette, de muscles et d'une muqueuse. Son squelette est qualifié d'ostéofibreux ; il est constitué de l'os hyoïde, situé au dessus du larynx, sur lequel se fixe la membrane hyo-glossienne, d'une hauteur d'un centimètre environ, et le septum lingual, lame fibreuse à l'origine de la dépression visible sur toute la longueur de la langue. Son mouvement est contrôlé par dix sept muscles, dont huit paires de muscles agonistes/antagonistes. Quatre paires de muscles extrinsèques (muscles qui prennent naissance à l'extérieur de la langue) servent notamment à sa protrusion, sa rétraction, sa dépression ou son élévation. La langue est usuellement décrite comme un ensemble de deux structures au comportement distinct, la racine (ou base), fixée à l'os hyoïde, et le corps, plus mobile. Ce dernier se décompose également en deux parties, le dos et la pointe de la langue, nommée apex. L'organisation du système musculaire de la langue ainsi que ses principales structures sont illustrées à la Figure 1.3.

[6] Notons cependant que le larynx n'est pas totalement à exclure de l'ensemble des articulateurs, puisqu'il peut être amené à se déplacer, notamment lors de l'articulation des voyelles.

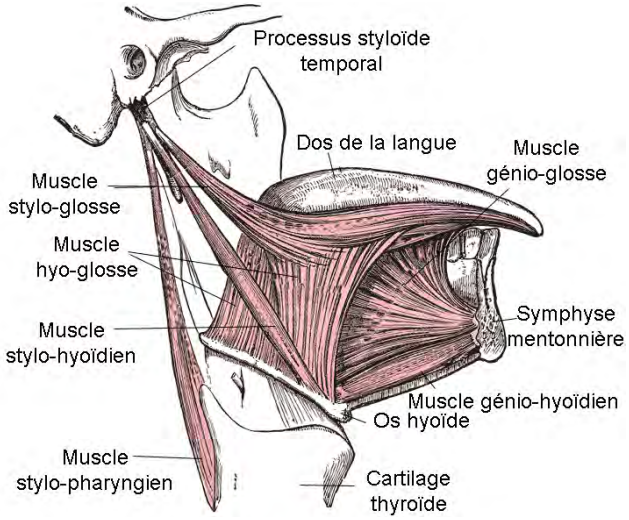


Figure 1.3 : Structures de la langue, détails des muscles extrinsèques[7] (plan sagittal médian, vue de droite)

Le rôle de la langue dans la phonation est déterminant, notamment pour la production des voyelles, caractérisée par le libre passage de l'air dans le résonateur. La phonétique articulatoire[8] décrit le système vocalique d'une langue (classification des voyelles) précisément à l'aide de deux critères qui décrivent la configuration de la langue dans la cavité buccale. Le premier est le « lieu d'articulation » ; « avant » ou « arrière », il localise la masse de la langue et qualifie ainsi les voyelles produites d' « antérieures », de « centrales » ou de « postérieures ». Le second critère est « l'aperture » ; il décrit l'espace de résonnance ménagé entre la langue et le palais (fermé ou ouvert), qualifiant ainsi les voyelles produites de « hautes » ou « basses ». La langue joue également un rôle important pour l'articulation des consonnes, dont le mode de production est, à l'inverse des voyelles, caractérisé par l'obstruction du passage de l'air dans le résonateur. Dans ce cas, le « lieu d'articulation » localise cette obstruction. Pour produire une consonne dite « dentale » ([t], [d], [n]), la pointe de la langue crée cette obstruction en se rapprochant des dents. Dans le cas des consonnes « inter-dentales » ([th] comme *thin*, [dh] comme *then*), la langue dépasse les dents et vient s'appuyer directement sur les incisives. Pour les consonnes dites « alvéolaires ([s], [z] ou la consonne liquide [l] mais également [t], [d], [n]), et « post-alvéolaires » ([ch] comme ***ch**urch*, [jh] comme *judge*, [sh] comme ***sh**e*, [zh] comme *a**z**ure*), elle se déplace respectivement vers les alvéoles (creux de l'os alvéolaire dans lequel est enchâssée une dent) et vers la partie antérieure du palais (à la juxtaposition avec le palais dur). Pour une consonne dite « palatale » ([j] comme *ye*, catégorisée également comme une semi-voyelle), l'organe articulateur est le dos de la langue, l'obstruction ayant lieu au niveau du palais dur. Pour

[7] Illustrations publiées sous la licence de documentation libre GNU FDL, d'après (Gray, 1973)

[8] Cette section n'a pas pour but de dresser une description exhaustive des configurations articulatoires possibles. Elle présente plutôt les phonèmes pour lesquels la langue et les lèvres, les principaux articulateurs visés par le système envisagé, sont fortement impliqués.

une consonne vélaire ([k], [g], [ng] comme *parking*), la partie postérieure du dos de la langue se bombe et se rapproche du palais mou. Enfin, pour une consonne uvulaire ([r] comme *Paris* en français), le lieu d'articulation se situe au niveau de la luette.

Les lèvres constituent l'autre articulateur majeur de la cavité buccale. Elles permettent la production des consonnes « bilabiales » (rapprochement des lèvres inférieures et supérieures, [p], [b], [m]) et des consonnes « labio-dentales » ([f], [v], rapprochement de la lèvre inférieure avec les dents). Elles interviennent également dans le cadre de la production vocalique en apportant la notion d'arrondissement des voyelles. Enfin, la réalisation acoustique de certains phonèmes nécessite parfois deux lieux d'articulation, impliquant à la fois la langue et les lèvres ; c'est le cas notamment de la consonne « labio-velaire » [w] (comme *who*).

Le dernier articulateur du résonateur est le voile du palais qui permet, lorsqu'il s'abaisse, de mettre en parallèle les cavités buccale et nasale. Il intervient notamment dans la production des consonnes nasales [m], [n] et [ng] en les différenciant respectivement des groupes de consonnes ([p], [b]), ([t], [d]), et ([k], [g]), qui présentent la même configuration linguale et labiale. Enfin, l'abaissement du voile du palais permet, en langue française notamment, la formation des voyelles nasales [ɔ] (*on*), [ɛ] (*hein*), [œ] (*un*), [ɑ] (*an*).

Au regard de ces principaux résultats issus de la phonétique articulatoire, la réalisation acoustique d'un phonème dépend principalement des configurations de la langue, des lèvres et du voile du palais mais également de l'activité des cordes vocales. Lorsque ces dernières doivent être retirées, dans le cadre notamment du traitement chirurgical du cancer du larynx, les mécanismes de la phonation sont profondément modifiés.

1.2. Une voix sans larynx

Les cancers de la sphère ORL, et du larynx en particulier, sont des maladies relativement fréquentes[9]. D'après les dernières statistiques publiées par l'Institut de Veille Sanitaire[10], datant de 2005, ils représentent, en France, environ 12 % de l'ensemble des cancers diagnostiqués. En 2005, l'incidence annuelle, ajustée pour l'âge, pour 100 000 personnes, était de 3 800 cas de cancers du larynx. Cette affection touche essentiellement les hommes (1 femme pour 7 hommes). Le tabagisme actif en est la cause principale, aggravée par la consommation conjointe d'alcool et la respiration de matières cancérigènes telle que l'amiante.

Le cancer du larynx peut naître dans n'importe quelle partie de cet organe. La tumeur prend généralement la forme d'une ulcération anormale d'une des deux cordes vocales. Le traitement consiste alors en une radiothérapie et une chimiothérapie, associée à l'ablation de la corde vocale atteinte (cordectomie). Cependant, lorsque l'étendue du cancer est trop importante et touche la quasi totalité de l'organe, l'ablation complète du larynx est nécessaire

[9] Les informations de nature médicale énoncée dans cette section s'appuient notamment sur le Groupe Coopérateur Multidisciplinaire en Oncologie GERCOR (http://www.canceronet.com).

[10] Statistiques disponibles sur le site Internet de l'Institut : http://www.invs.sante.fr

(laryngectomie totale[11]). Le larynx jouant le rôle d'aiguilleur entre les voies respiratoires et digestives, son ablation nécessite l'isolement de ces deux voies. Pour pouvoir respirer, le patient subit alors une trachéostomie (trachéotomie permanente), c'est-à-dire la mise en place d'un trou au milieu du cou, relié à la trachée. La cavité buccale est alors connectée exclusivement à l'œsophage, ce qui permet une alimentation normale, comme l'illustre la Figure 1.4.

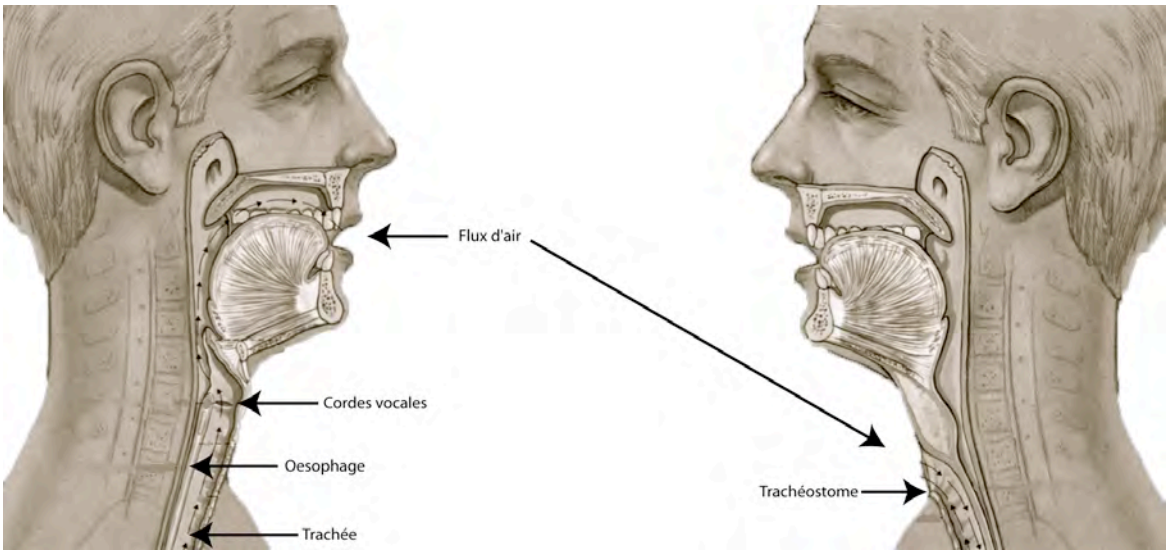


Figure 1.4 : Appareil phonatoire d'une personne laryngectomisée (à droite, avant, à gauche, après l'intervention)[12]

La laryngectomie totale a pour conséquence la perte de la voix. En effet, l'air pulmonaire passe exclusivement par le trachéostome et ne peut donc pas atteindre la cavité buccale. Sans air, la phonation est impossible. Pour la rétablir partiellement, plusieurs techniques existent. Tout d'abord, le patient peut apprendre la technique dite de la « voix œsophagienne » qui consiste à ingérer, puis à renvoyer de l'air par l'œsophage (éructation). La voix digestive remplace alors la voie respiratoire dans le rôle de soufflerie. Au passage de cet air, la bouche œsophagienne, c'est-à-dire l'orifice d'abouchement de l'œsophage dans l'hypopharynx (qui a une fonction d'un sphincter), se met à vibrer et forme ainsi un appareil vibrateur de substitution. L'air ainsi mis en forme (pulsé) peut alors « résonner » normalement dans les cavités buccales et nasales ; la phonation est rétablie. Cette technique, bien que très utilisée à ce jour, reste relativement difficile à maitriser. Chaque réjection d'air ne permettant la production que d'un nombre restreint de syllabes, la pratique de la voix œsophagienne demande une importante dépense d'énergie. De plus, la parole générée est généralement de faible volume et demande parfois une amplification à l'aide d'un dispositif électronique externe. La seconde alternative proposée au

[11] D'après la société *Heva Health Evaluation*, spécialisée dans l'épidémiologie hospitalière et l'analyse pharmaco-économique (http://www.hevaweb.com), 698 personnes ont subi une laryngectomie totale en 2007, en France.

[12] D'après des illustrations de *Inhealth Technologies,* extraites de http://inhealthtechnologies.com

laryngectomisé est la « voix trachéo-oesophagienne ». Cette technique consiste à réaliser une communication entre la trachée et l'œsophage (fistule), par la mise en place, par voie chirurgicale[13], d'un implant phonatoire. Ce dernier fonctionne en « sens unique » ; il autorise le passage de l'air pulmonaire de la trachée vers l'œsophage mais interdit celui de la salive, des aliments et des liquides de la cavité buccale vers la trachée. A la différence de la voix œsophagienne, l'air n'a plus besoin d'être ingéré. Lorsque le trachéostome est obstrué à l'aide d'un doigt ou grâce à une valve trachéale automatique, l'air pulmonaire est redirigé depuis la trachée vers l'œsophage et vient faire vibrer la bouche œsophagienne, permettant ainsi la phonation. Cette technique est plus « confortable » que la voix œsophagienne, elle est physiquement moins exigeante, la durée possible de phonation est plus longue, et la parole produite est généralement d'une intelligibilité plus satisfaisante. Néanmoins, la mise en place de la voix trachéo-oesophagienne n'est pas toujours possible[14] et la présence de l'implant phonatoire peut parfois entraîner de complications (fuites alimentaires autour de l'implant, granulomes, déplacements etc.).

Bien qu'intelligible, la parole produite par ces deux techniques reste, de plus, d'une qualité relativement limitée. Utilisée comme appareil vibrateur de substitution, la bouche œsophagienne vibre moins vite que de véritables cordes vocales et la hauteur de la parole produite est généralement très basse. La fréquence de cette vibration étant difficilement contrôlable, l'intonation de la parole produite est quasiment plate. Une comparaison de la voix trachéo-œsophagienne et de la voix laryngée est proposée à la Figure 1.5. La fréquence fondamentale estimée est basse et varie très peu[15].

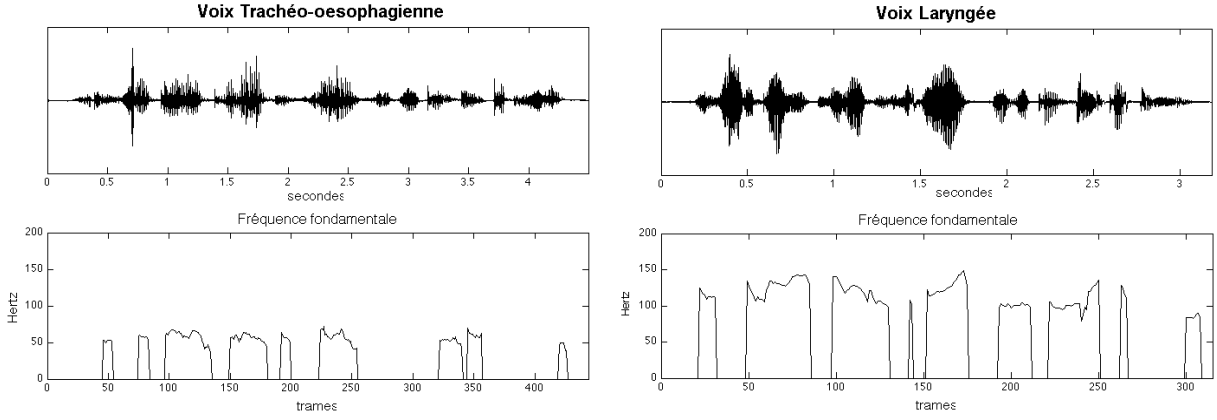


Figure 1.5 : Comparaison de la voix trachéo-oesophagienne avec la voix laryngée - Les deux locuteurs prononcent la phrase : « Il se garantira du froid avec un bon capuchon »

[13] En première intention, le jour de la laryngectomie totale ou en deuxième intention, lors d'une intervention ultérieure. Néanmoins, l'implant (ou prothèse) phonatoire doit être régulièrement remplacé.

[14] L'échec de la voix trachéo-œsophagienne est souvent lié à un spasme des muscles constricteurs du pharynx (Source : Association « Les Mutilés de la voix », www.mutiles-voix.com)

[15] Cet exemple sonore est disponible sur la page Internet associée à ce manuscrit (URL indiquée dans l'introduction générale).

La troisième et dernière possibilité proposée à une personne laryngectomisée pour rétablir la possibilité d'une communication parlée après l'intervention, est l'utilisation d'un dispositif externe nommé « electrolarynx ». Il s'agit d'un appareil portable qui joue le rôle de vibrateur de substitution. Maintenu généralement sous le menton[16], il produit une onde mono fréquence qui sera transmise puis modulée dans la cavité buccale. Cette dernière est alors constamment excitée, aussi bien pour les phonèmes voisés que non voisés, et sans variation de fréquence. Bien qu'intelligible, la parole produite reste, de ce fait, très « robotique ». Aussi, cette technique, qui ne requiert néanmoins pas d'apprentissage long, est souvent proposée juste après l'intervention, comme solution transitoire.

En complément de ces approches, anatomique, chirurgicale et mécanique, différents travaux de recherche, basés cette fois sur le traitement du signal acoustique, ont également été effectués pour améliorer la qualité de la communication parlée chez le laryngectomisé. L'objectif principal de ces travaux est le rétablissement des caractéristiques de voisement et d'intonation. Dans (Yingyoung, 1990), le codage par prédiction linaire (LPC) est utilisé dans le cadre de la voix trachéo-œsophagienne, pour améliorer la qualité des voyelles, par analyse puis « re-synthèse » du signal à l'aide d'une fonction d'activation basée sur une impulsion glottale naturelle ; la fréquence fondamentale moyenne est, de plus, lissée puis rehaussée. Dans (Matsui, 1999), une approche similaire est proposée dans le cas de la voix œsophagienne ; l'amélioration des caractéristiques spectrales du signal est ici effectuée à l'aide de la technique de synthèse par formants. Plus récemment, des approches basées sur les techniques dites de « conversion de voix » (ou *morphing* vocal) ont été proposées (Del Pozo, 2008). Initialement développées pour transformer la voix d'un locuteur source en celle d'un locuteur cible, ces techniques sont ici employées pour convertir le signal produit par la voix trachéo-œsophagienne (la source) en celui qui serait produit par le même locuteur s'il disposait encore de sa voix laryngée (la cible). Ces techniques sont basées sur l'apprentissage d'une « fonction de conversion », qui s'obtient en modélisant, par des mélanges de gaussiennes, les densités de probabilités conjointes des paramètres d'excitation (fréquence fondamentale) et de timbre (caractéristiques spectrales) des voix source et cible.

Ces différentes techniques s'appuient exclusivement sur l'analyse du signal acoustique, généré à l'aide des techniques de voix œsophagienne ou trachéo-œsophagienne, puis traité. Ces approches sont donc utiles pour la communication à distance, en transmettant par exemple le signal acoustique amélioré sur le réseau téléphonique. Néanmoins, la communication parlée « face à face » reste plus délicate car le signal perçu par l'auditeur est alors une superposition du signal transformé et du signal original. Même si certains traitements peuvent être réalisés quasiment en temps réel, la présence résiduelle de ce dernier reste inévitable.

[16] Lorsque les tissus autour du cou sont trop sensibles, par exemple durant une radiothérapie, l'électrolarynx peut être appliqué contre la joue. Il peut également être muni d'une sonde buccale, disposée à l'intérieur de la bouche, entre la gencive et la bouche.

Dans le cadre de ce travail, nous proposons une approche différente, basée non sur l'analyse du signal acoustique mais sur l'accès direct aux mouvements articulatoires qui en sont à l'origine. L'objectif est alors de synthétiser un signal acoustique, d'une qualité similaire à celle d'un signal produit en voix laryngée, à partir uniquement d'informations de nature « non acoustique » sur l'activité de l'appareil vocal. Dans cette approche, la phonation n'est plus nécessaire. Le locuteur articule normalement mais ne produit aucun son ; il ne vocalise pas. Il produit alors ce que nous appellerons par la suite, une « parole silencieuse », convertie en un signal acoustique audible par ce que nous qualifierons d' « interface de communication en parole silencieuse » ou ICPS.

1.3. Interface de communication en parole silencieuse : état de l'art

Destinée à reconstituer un signal audible à partir d'informations inaudibles, à interpréter une parole normalement articulée mais non vocalisée, une interface de communication en parole silencieuse pourrait représenter pour les patients laryngectomisés, une alternative ou un complément aux techniques existantes décrites à la section précédente. Néanmoins, son champ applicatif est plus large ; il s'étend notamment au secteur des télécommunications civiles et militaires. Une ICPS permettrait en effet la communication parlée dans un milieu où la discrétion est capitale, comme dans le cadre d'une opération de sécurité (prise d'otage), ou très appréciable, par exemple pour téléphoner dans un transport en commun sans déranger les autres voyageurs. Une ICPS constituerait également la base d'un système mains-libres de saisie silencieuse de données, utilisable notamment pour la transmission confidentielle de codes et autres mots de passe. N'étant pas basée sur l'analyse du signal acoustique recueilli en sortie de la cavité buccale, ce type d'interface pourrait également faciliter la communication dans un environnement extrêmement bruyant, où la modalité acoustique est trop bruitée pour être utilisable.

Interface homme-machine réunissant des techniques d'instrumentation et de modélisation issues de domaines divers comme le traitement automatique de la parole, les sciences du langage et la bio-ingénierie, les interfaces de communication en parole silencieuse constituent un domaine de recherche relativement récent. Le concept d'un synthétiseur de parole piloté exclusivement par la saisie de l'activité articulatoire apparaît pour la première fois en 1985 dans (Sugie et Tsunoda, 1985). A partir des signaux recueillis par trois capteurs electromyographiques[17] disposés sur le visage du locuteur, le système décrit est capable, dans 71% des cas, d'identifier correctement les cinq voyelles du japonais, puis de les restituer sur un haut parleur externe. Dans (Hasegawa et Ohtani, 1992), le même objectif de décodage des voyelles est poursuivi, mais le capteur utilisé est cette fois ci une caméra vidéo qui fournit une image des lèvres du locuteur pendant l'articulation. L'utilisation de la modalité visuelle de la parole, c'est-à-dire l'image des lèvres, en complément de la modalité acoustique (le signal audio)

[17] Technique de mesure de l'activité nerveuse et musculaire, volontaire et réflexe, basée sur le recueil du potentiel d'action électrique des cellules musculaires.

avait par ailleurs déjà été proposée par Petajan (Petajan, 1984), dans le but d'améliorer les performances d'un système de reconnaissance automatique de la parole en présence de bruit. Cependant, d'après les résultats de phonétique articulatoire présentés à la section 1.1, le pouvoir de discrimination phonémique des lèvres est relativement limité. Ainsi la prononciation de langue anglaise est généralement décrite à l'aide d'un jeu d'environ 40 phonèmes, supposé décrire autant de configurations articulatoires différentes, mais seuls 13 groupes de phonèmes présentant des configurations labiales distinctes peuvent être constitués. Ces groupes de « sosies labiaux » sont nommés visèmes (*visual phoneme* ; (Fisher, 1968)). Pour cette raison, on considère d'ailleurs que la lecture labiale ne permet de percevoir que 30 % environ du message oral émis ; le reste de la compréhension s'effectue notamment par suppléance mentale (intégration du contexte) et par l'interprétation d'informations non verbales comme certains gestes manuels, faciaux (mimiques) et corporels (Dumont et Calbour, 2002). Ainsi, l'image des lèvres ne permet pas de distinguer suffisamment de configurations articulatoires pour être utilisée, seule, dans le cadre d'une ICPS. La mise en œuvre de techniques instrumentales permettant l'accès aux mouvements des articulateurs internes est donc nécessaire. De l'étude de la littérature récente (depuis 2000) sur la réalisation d'une ICPS, se dégagent différentes approches. Ces dernières sont brièvement décrites dans les paragraphes suivants[18].

Articulographie électromagnétique

Dans (Fagan *et al.*, 2008), un ensemble de sept aimants permanents est reparti sur la langue, les lèvres et les dents du locuteur, créant ainsi un champ magnétique local autour de la cavité buccale. Pendant l'articulation, les aimants sont en mouvement et la structure de ce champ se modifie. Ces variations sont mesurées à l'aide de six capteurs magnétiques disposés sur une paire de lunettes. Les signaux ainsi recueillis sont utilisés dans le cadre d'un système de reconnaissance de mots isolés (reconnaissance par comparaison d'exemples), choisis soit parmi un vocabulaire de 9 mots (*cat*, *dog*, *on*, *off*, *up*, *down*, *yes*, *no*, *bag*), soit parmi 13 phonèmes.

Electromyographie

Dans (Jorgensen *et al.*, 2003), l'électromyographie (EMG) est utilisée pour capter l'activité nerveuse et musculaire du larynx et de la langue, à l'aide de quatre électrodes de surface placées sur la gorge du locuteur. Le système proposé est évalué dans le cadre (restreint) du décodage de 6 mots (*stop*, *go*, *left*, *right*, *alpha*, et *omega*). Dans (Maier-Hein *et al.*, 2005), (Walliczek *et al.*, 2006), (Jou *et al.*, 2006), (Jou *et al.*, 2007), l'électromyographie est également la technologie utilisée, mais les électrodes de surface, au nombre de douze, sont ici disposées sur le visage du locuteur, en haut de la gorge et sous le menton. Le placement des électrodes est déterminé afin

[18] Pour une étude approfondie des différentes approches proposées pour la réalisation d'interfaces de communication en parole silencieuse, on pourra consulter le numéro spécial de la revue « *Speech Communication* » consacré à cette thématique (parution courant 2010), et notamment l'article de synthèse (Denby *et al.*, 2009).

de capter, le plus précisément possible, les activités des muscles impliqués dans le mouvement des lèvres (muscle élévateur de l'angle de la bouche, muscle grand zygomatique, platysma ou muscle peaucier du cou, muscle orbiculaire de la bouche) et de façon moins localisée, l'activé des muscles de la langue. Si les différents systèmes proposés sont également des systèmes de « reconnaissance de la parole », la plupart de ces travaux portent cette fois sur le décodage de la parole continue. Il s'agit alors d'identifier, à partir de l'analyse des signaux recueillis, la suite des mots prononcés. Les contributions principales portent sur la description des signaux electromyographiques (extraction des caractéristiques), leur modélisation et leur décodage s'effectuant à l'aide des techniques traditionnellement utilisées pour la reconnaissance de la parole « acoustique ». En restreignant le vocabulaire autorisé à 100 mots, et en utilisant un modèle stochastique de langage (trigramme)[19], cette approche permet notamment le décodage d'une séquence de mots, prononcée de façon continue, avec une performance de l'ordre de 70 % (*Word error rate* de l'ordre de 30 %).

Microphonie stéthoscopique (NAM)

Une autre approche proposée pour la réalisation d'une ICPS se base sur la production et la saisie de murmures qualifiés d' « inaudibles » (*Non Audible Murmur* ou NAM), qui peuvent être assimilés à des chuchotements presque imperceptibles. Dans ce mode de production, les cordes vocales ne vibrent pas. Les cavités orales sont donc excitées par un flux d'air laryngé continu mais, à la différence de la voix chuchotée, d'énergie extrêmement faible. Le phénomène de radiation aux lèvres étant très limité, la parole produite est difficilement perceptible, même pour un auditeur proche du locuteur. Néanmoins, les très faibles résonnances de l'onde de pression dans les cavités orales ainsi que les chocs des articulateurs contre les parois sont transmis dans les tissus mous de la tête et peuvent être capturés à l'aide d'un microphone spécifique, dit « stéthoscopique », placé juste en dessous de l'oreille. Initialement proposé par Nakajima en 2003, les premiers travaux sur une ICPS basée sur ce type de capteur portent sur la transcription (reconnaissance) automatique de cette parole murmurée (Nakajima *et al.*, 2003), (Heracleous *et al.*, 2003), (Nakajima *et al.*, 2006). Dans (Toda et Tomoki, 2005), une autre approche est proposée. L'objectif poursuivi est alors de transformer directement le signal de parole murmurée en celui qui serait produit en voix laryngée, à l'aide des techniques de « conversion de voix ». Cette approche est également adoptée par Tran pour la conversion en voix laryngée, d'une voix qualifiée cette fois de « chuchotée » (Tran *et al.*, 2008a), (Tran *et al.*, 2008b). Ces travaux proposent également de combiner les informations acoustiques fournies par ce type de microphone avec des informations visuelles sur les mouvements faciaux, capturés à l'aide de 142 capteurs collés sur le visage du locuteur. Par ailleurs, la réalisation d'une voix murmurée ou

[19] Pour une langue donnée, un modèle stochastique de langage fournit la probabilité d'apparition d'un mot sachant le contexte lexical antérieur.

chuchotée n'est possible qu'en présence d'un flux d'air, même faible, circulant depuis la trachée vers les cavités orales. Chez une personne laryngectomisée, l'air laryngé s'échappe intégralement par le trachéostome et une ICPS basée exclusivement sur cette approche n'est *a priori* pas envisageable. Pour faire face à ce problème, il a notamment été proposé de combiner ce microphone stéthoscopique avec un vibreur mécanique, qui fournit alors l'onde de pression excitatrice manquante (Nakamura *et al.*, 2007).

Imagerie ultrasonore

Parallèlement aux travaux de Nakajima (NAM) et de Jorgensen (EMG), Denby propose de saisir l'activité articulatoire à l'aide d'un capteur ultrasonore placé sous la mâchoire du locuteur (Denby et Stone, 2004). L'utilisation des ultrasons est en effet un moyen non invasif d'accéder aux mouvements des articulateurs internes et notamment à ceux de la langue. Le travail de thèse décrit dans ce document s'inscrivant dans la lignée de ces travaux, cette approche sera décrite plus en détail à la section 1.4.

Électro-encéphalographie et magnéto-encéphalographie

Bien qu'à un stade encore très exploratoire, des travaux ont récemment été menés sur la réalisation de systèmes de communication pilotés par l'activité cérébrale. Rétablir une forme de communication parlée simplement à partir de la pensée est, en effet, un nouveau défi pour les interfaces « cerveau-ordinateur ». Les premières contributions, apportées par Suppes (Suppes *et al.*, 1997; Suppes *et al.*, 1998), portent sur la reconnaissance de 7 mots à partir d'enregistrements électro-encéphalographiques (EEG) et magnéto-encéphalographique (MEG) d'un sujet se voyant présenter le mot à « penser » sous la forme d'un stimulus visuel. Dans une approche légèrement différente décrite dans (Wester, 2006), le sujet « s'imagine » en train d'articuler un mot mais ne produit aucun des mouvements associés. Si ces travaux mettent en œuvre un dispositif expérimental relativement non-invasif comme l'EEG, d'autres s'appuient sur une instrumentation plus lourde, basée sur l'implantation d'électrodes directement dans le cortex cérébral. On citera notamment les travaux de Brumberg et Guenther (Brumberg *et al.*, 2008) effectués notamment dans le cadre du traitement du syndrome d'enfermement (*Locked-in syndrome*).

Systèmes dédiés à la communication en environnement bruyant

Toutes les approches décrites dans les paragraphes précédents pourront *a priori* être envisagées pour la mise en œuvre d'un dispositif permettant aussi bien la communication en parole silencieuse que celle en environnement extrêmement bruyant. Néanmoins, de nombreux autres travaux, visant exclusivement cette dernière application, ont été effectués et des dispositifs sont déjà commercialisés (Bos et Tack, 2005). C'est notamment le cas des « microphone de gorge » ou « laryngophone », (*Throat Microphone*, (Stuart, 1939)), utilisés depuis longtemps par

les pilotes de chasse, qui captent principalement les vibrations des cordes vocales mais également une partie des résonnances dans la cavité buccale, transmises par les tissus peauciers. L'amélioration de ce type de dispositif fait encore l'objet de travaux. Ces derniers visent principalement à améliorer le couplage air-tissu, augmentant ainsi l'immunité au bruit ambiant extérieur (Quatieri *et al.*, 2006). Egalement utilisés dans diverses situations militaires, les microphones à conduction osseuse (*Bone microphone*) exploitent le phénomène de transmission du son à travers les os du crâne. L'immunité au bruit ambiant peut également être obtenue en plaçant un microphone acoustique traditionnel au plus près des organes de production, comme c'est le cas dans les microphones intra-auriculaires (*in-ear microphone*). Enfin, plusieurs travaux, qui ne seront pas détaillés ici, se focalisent sur la saisie de l'activité glottale qu'ils proposent d'utiliser pour le débruitage du signal de parole, enregistré par un microphone standard (technologies GEMS (Burnett *et al.*, 1997) et TERC (Brown *et al.*, 2004)).

Inversion et synthèse articulatoire

Au sein de la communauté de la recherche sur la parole, la modélisation de la relation entre l'activité de l'appareil vocal et la réalisation acoustique, est traditionnellement envisagée dans le cadre de l'inversion ou de la synthèse articulatoire. Si ces recherches portent plus sur la compréhension approfondie des mécanismes qui régissent la production de la parole que sur la conception de « systèmes de communication », les techniques qui y sont développées présentent cependant un intérêt certain pour la conception d'interfaces de communication en parole silencieuse. Parmi les travaux les plus récents, on citera notamment : (Potard, 2008), (Toda *et al.*, 2008), (Kröger et Birkholz, 2007) et (Krstolovic, 2001).

1.4. Interface de communication silencieuse par imagerie ultrasonore et vidéo de l'appareil vocal

La première étude sur la réalisation d'une ICPS basée sur l'imagerie ultrasonore de la cavité buccale est décrite dans (Denby et Stone, 2004). L'objectif poursuivi dans cette étude est la synthèse du signal de parole (et non la transcription vers du texte), directement à partir d'un flux d'images ultrasonores de la cavité buccale. Le système d'acquisition utilisé est constitué d'un échographe standard, utilisé pour le diagnostic médical. Il fournit une image de la cavité buccale dans le plan sagittal médian sur laquelle apparaît distinctement la surface supérieure de la langue (une analyse qualitative du type d'image obtenu sera effectuée au chapitre suivant, à la section 2.3.2). Dans cette première étude, une modélisation par réseau de neurones est mise en œuvre pour piloter un codeur vocal GSM (*Global System for Mobile Communications*) à partir de la position de la langue observée dans l'image ultrasonore. Dans (Denby *et al.*, 2006), une approche similaire est proposée mais le dispositif expérimental est complété par une caméra vidéo qui fournit, en complément de l'image ultrasonore, une image des lèvres du locuteur. Ces travaux exploratoires, à l'origine de résultats préliminaires prometteurs, ont permis

d'appréhender certaines des problématiques sous-jacentes à la réalisation d'une ICPS basée sur l'imagerie ultrasonore et vidéo de l'appareil vocal. Ce travail de thèse, qui s'inscrit dans la continuité de ces travaux, propose de nouvelles approches. Une description schématique du dispositif envisagé est rappelée à la Figure 1.6.

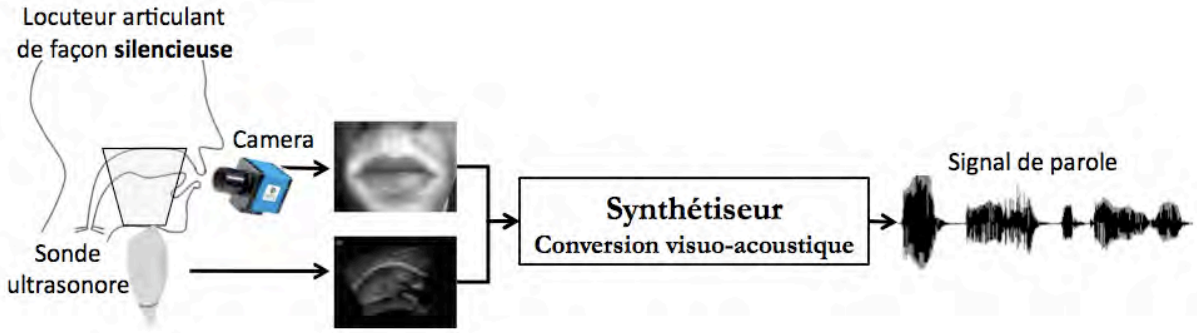


Figure 1.6 : Interface de communication en parole silencieuse par imagerie ultrasonore et vidéo de l'appareil vocal – Schéma général de fonctionnement

Le « synthétiseur » présenté à Figure 1.6 réalise une conversion « visuo-acoustique », qui permet le passage des modalités visuelles (image ultrasonore de la cavité buccale et image vidéo des lèvres) à la modalité acoustique (signal audio).

L'objectif poursuivi dans ce travail de recherche est d'évaluer la faisabilité d'un tel système et de délimiter le périmètre du champ applicatif que ce dernier pourrait occuper. Ainsi, c'est dans un contexte de « laboratoire » que seront mises en œuvres les différents protocoles expérimentaux. Les contraintes spécifiques liées à la réalisation d'un prototype ergonomique, capable d'effectuer les différents traitements en « temps réel », ne seront que très peu abordées. Néanmoins, une vue conceptuelle de différentes formes que pourraient prendre un tel prototype est proposée à la Figure 1.7.

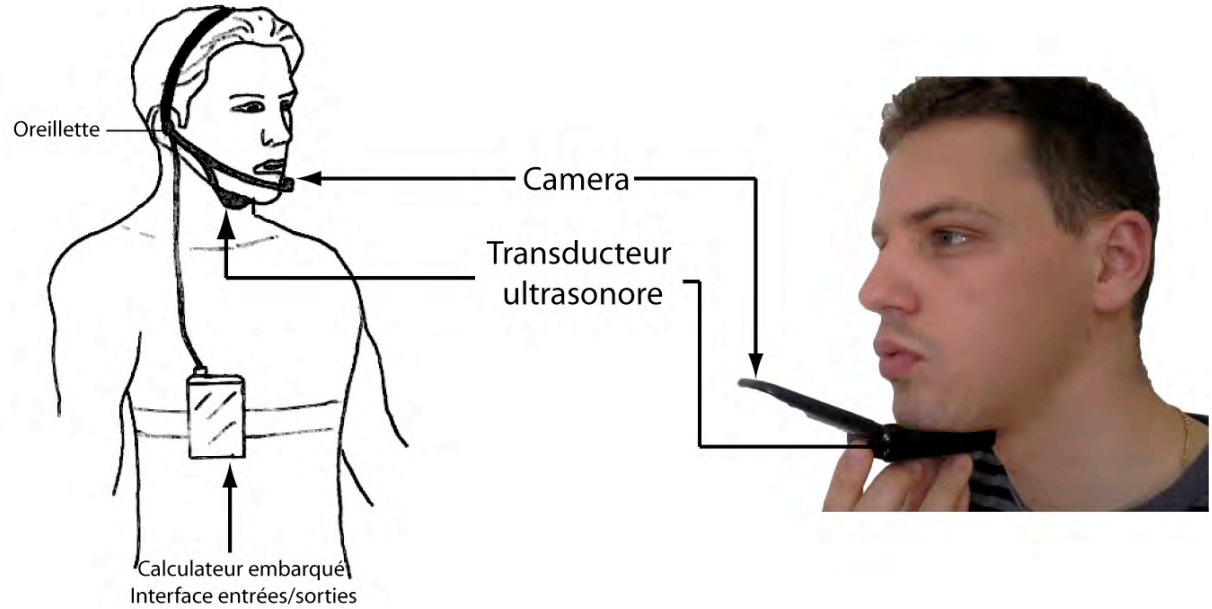


Figure 1.7 : Vue conceptuelle des prototypes envisagés

Dans l'approche proposée, la modélisation « visuo-acoustique » s'effectue par apprentissage artificiel. Cette étape nécessite notamment la constitution de bases de données qui relient observations articulatoires et réalisations acoustiques. Elle requiert la réalisation d'un dispositif expérimental permettant l'acquisition synchrone de données ultrasonores, vidéo et audio. Ceci fait l'objet du chapitre suivant.

Chapitre 2. Protocole expérimental d'acquisition des données

2.1. Avant-propos

L'élément majeur de l'instrumentation de l'ICPS envisagée est le capteur ultrasonore qui renseigne sur la configuration du conduit vocal pendant l'articulation. Comme expliqué dans les sections suivantes, un système d'analyse par ultrasons permet d'établir la « signature » d'un milieu en mesurant les réflexions d'une onde ultrasonore (échos) émise dans sa direction. Dans le domaine industriel, les applications des ultrasons sont multiples. On citera la télémétrie (mesure de distance), la télédétection (*sonar*) ou même la sonication[20]. Dans le domaine médical, les ultrasons sont la plupart du temps utilisés à des fins de diagnostic dans le cadre de l'échographie (ou imagerie ultrasonore)[21]. Dans le domaine de recherche sur la phonétique articulatoire, l'imagerie ultrasonore est devenue une des techniques privilégiées pour l'étude du mouvement de la langue, grâce aux travaux précurseurs de Maureen Stone. Un échographe permet notamment, grâce à une disposition particulière des capteurs ultrasonores (émetteur/récepteur) et à un traitement spécifique des échos reçus, d'obtenir une information sur la géométrie du milieu biologique étudié, en formant une image d'un plan de coupe de ce dernier. Dans le cadre de cette étude, c'est ce type de système d'analyse ultrasonore qui sera utilisé pour analyser la configuration de l'appareil vocal. De plus, les contributions relatives à l'instrumentation, décrites dans les sections suivantes, sont basées sur l'utilisation de systèmes « standard » d'échographie. La conception d'un capteur *ad-hoc* n'est pas envisagée.

2.2. L'imagerie ultrasonore : principe et caractéristiques générales

Les sections suivantes exposent les bases physiques de l'imagerie ultrasonore ainsi que les principes généraux de fonctionnement d'un échographe ; elles s'appuient sur plusieurs ouvrages et articles de référence comme (Bonnin *et al.*, 2004) et (Stone, 2005).

2.2.1. Bases physiques de l'analyse d'un milieu par ultrasons

Une onde ultrasonore est une onde dont la fréquence est supérieure à 20 kHz ; ce type d'onde est inaudible pour l'oreille humaine. En échographie, des ondes ultrasonores sont émises en direction du milieu à imager. Ces ondes sont des ondes de pression, et leur propagation est à l'origine d'un phénomène de compression puis de dilatation du milieu traversé, transmis de proche en proche. La vitesse de propagation de l'onde c est une caractéristique du milieu

[20] Utilisation des ultrasons pour rompre les membranes des cellules, afin de nettoyer ou désinfecter du matériel.

[21] Dans le domaine médical, les ultrasons peuvent également être utilisés à des fins thérapeutiques ; le faisceau ultrasonore est alors focalisé afin de provoquer un échauffement pouvant entraîner une destruction cellulaire.

traversé : elle dépend exclusivement de la densité et de l'élasticité de ce dernier[22]. Dans les tissus mous, la vitesse de propagation d'une onde ultrasonore, soit 1540 m/s, est proche de celle de l'eau, qui est de 1480 m/s. Dans l'air ou dans un tissu osseux, les vitesses de propagation moyennes sont très différentes (respectivement 340 m/s et 3000 m/s).

Phénomènes de réflexion et de réfraction

Lorsqu'une onde traverse différents milieux, elle est susceptible de subir, aux interfaces de ces milieux, un ensemble de réflexions et de réfractions. Dans le cadre du phénomène de réflexion, le faisceau réfléchi « repart » de l'interface avec un angle identique à l'angle d'incidence. Dans le cadre de la réfraction, le faisceau incident est dévié d'un angle qui dépend du rapport des vitesses de propagation de l'onde dans les deux milieux traversés. Ces phénomènes sont illustrés à la Figure 2.1.

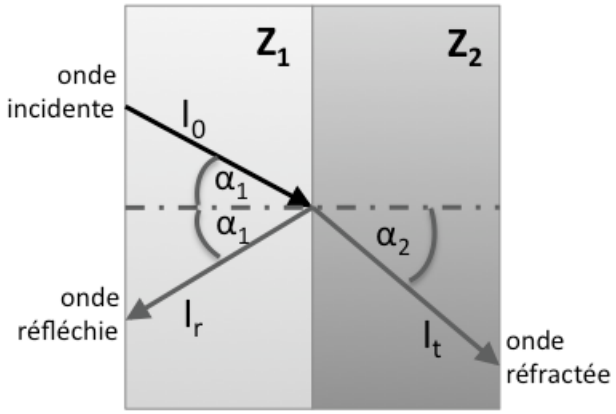


Figure 2.1 : Phénomènes de réflexion et de réfraction d'une onde ultrasonore à l'interface de deux milieux d'impédance acoustique Z1 et Z2 différentes

Les intensités des faisceaux réfléchis et réfractés (transmis) dépendent de la différence des impédances acoustiques des deux milieux ; l'impédance acoustique Z d'un milieu étant reliée à sa densité ρ et à la vitesse de propagation c par la relation $Z = \rho c$. A l'interface entre deux milieux d'impédance acoustique Z_1 et Z_2, les rapports entre intensité incidente I_0, intensité réfléchie I_r, et intensité transmise I_t, sont définis par les relations suivantes :

$$\frac{I_r}{I_0} = \left(\frac{r-1}{r+1}\right)^2 \text{ et } \frac{I_t}{I_0} = \frac{4r}{(1+r)^2} \text{ avec } r = \frac{Z_1}{Z_2} \qquad \text{(Équation 2.1)}$$

On constatera que plus le rapport des d'impédances des deux milieux est grand, plus la réflexion de l'onde incidente est importante. La mesure des caractéristiques des échos ultrasonores réfléchis est la base d'un système d'échographie. Les limites anatomiques des organes constituent en effet des interfaces entre des milieux d'impédances acoustiques différentes. Une

[22] Sous l'hypothèse de faible amplitude des ondes, en excluant les phénomènes non linéaires.

image échographique est alors la « cartographie » des échos ultrasonores réfléchis par ces interfaces.

Phénomène de diffusion

Une onde sonore émise à une fréquence *f*, contrainte dans un milieu donné à se propager à une vitesse *c*, possède une longueur d'onde λ qui est définie par $\lambda = c/f$. Lorsque cette longueur d'onde est d'un ordre de grandeur bien inférieur à l'ordre de grandeur de la structure cible, comme dans le cas d'un organe, les phénomènes de réflexion et de réfraction sont prédominants. En revanche, lors de la traversée par une onde d'un milieu très hétérogène, constitué de microstructures d'impédances acoustiques différentes (tissu spongieux ou organes parenchymateux par exemple) et de taille proche de la longueur d'onde, un nouveau phénomène intervient : la diffusion. Ce dernier consiste en une réémission isotrope d'une partie de l'onde incidente.

Atténuation de l'onde ultrasonore

Lors de son parcours, l'onde ultrasonore subit donc une série de réflexions, de réfractions et de diffusions. Cet ensemble d'interactions avec les différents milieux de propagation est à l'origine d'une atténuation de l'énergie transportée par l'onde. Les mécanismes sous-jacents à ce phénomène d'atténuation sont trop nombreux et complexes pour être décrits ici de manière exhaustive. Néanmoins, on retiendra d'une part, que l'intensité de l'onde ultrasonore décroit exponentiellement avec la profondeur de pénétration dans les tissus[23] et d'autre part, que la profondeur d'exploration maximale augmente lorsque la fréquence de l'onde diminue.

2.2.2. Fonctionnement de l'échographe

Le transducteur ultrasonore

Un transducteur est un dispositif qui convertit une grandeur physique en une autre. Un transducteur ultrasonore transforme un signal électrique en une onde ultrasonore et inversement, en utilisant l'effet piézoélectrique. La piézoélectricité est la propriété que possèdent certains corps de se polariser électriquement sous l'action d'une contrainte mécanique et réciproquement, de se déformer lorsqu'on leur applique un champ électrique. En soumettant ce type de matériau à un champ électrique alternatif, il se comprime puis se dilate périodiquement (vibration mécanique), émettant ainsi une onde sonore. Dans un transducteur échographique, un élément piézoélectrique est utilisé à la fois en émission, pour transformer une impulsion électrique de commande en une onde ultrasonore, et en réception, pour convertir en courant électrique les échos ultrasonores des ondes réfléchies. Pour permettre ce mode de fonctionnement dual, l'onde ultrasonore n'est pas émise en continu, mais elle est modulée par

[23] Le coefficient d'absorption pour les tissus mous est en moyenne de 0,9 dB/cm/MHz.

des impulsions brèves (émission pulsée). Le temps entre la fin d'une émission et le début d'une nouvelle émission d'impulsion est appelé le « temps d'attente ».

Formation de l'image ultrasonore

Le transducteur émet une onde ultrasonore en excitant un élément piézoélectrique. Cette onde est transmise au milieu à étudier et se propage dans les tissus et structures biologiques à imager. En raison des phénomènes de réflexions (et de diffusions) à l'interface des milieux d'impédance acoustique différente, des échos de cette onde font chemin arrière, en direction du transducteur qui, en mode récepteur pendant le temps d'attente, est prêt à les convertir en signal électrique. En supposant que la vitesse de propagation c de l'onde dans le milieu biologique (tissu mou) est fixe (conventionnellement fixée à 1540 m/s), la distance d_{cible} entre l'émetteur et l'interface à l'origine de l'écho est déduite du « temps de vol » t_{vol}, temps qui sépare l'émission de l'onde et la réception de cet écho, à l'aide de la relation $d_{cible} = c\frac{t_{vol}}{2}$. Le temps qui sépare deux émissions dépend de la profondeur d'exploration (tous les échos doivent être revenus avant une nouvelle émission). Il y a donc un compromis entre la fréquence d'émission des impulsions et la profondeur d'exploration.

Un seul élément émetteur-récepteur fournit une réponse monodimensionnelle dans la direction de tir du faisceau ultrasonore. Pour obtenir une image de la cible, le transducteur ultrasonore comporte un ensemble d'une centaine d'éléments piézoélectriques, disposés linéairement (transducteur à barrette linéaire) ou de façon convexe (transducteur à barrette radiale). La Figure 2.2 illustre de façon simplifiée un cycle d'émission/réception qui aboutit à la formation d'une image ultrasonore.

Un échographe fournit plusieurs modes d'affichage de la réponse bidimensionnelle ainsi déduite. Le plus commun est le mode B. Dans ce mode, la position d'un point dans l'image dépend du temps de vol de l'écho et de la position sur le transducteur de l'élément piézoélectrique. L'amplitude du signal électrique fourni par cet élément est représentée en niveaux de gris, après compensation du phénomène d'atténuation en fonction de la profondeur traversée (en anglais, *Time Gain Correction*, qui peut éventuellement être réglée manuellement) et adaptation de la dynamique de l'image (réduite par compression logarithmique). La Figure 2.3 explicite l'image formée dans le cas « d'école » exposé à la Figure 2.2.

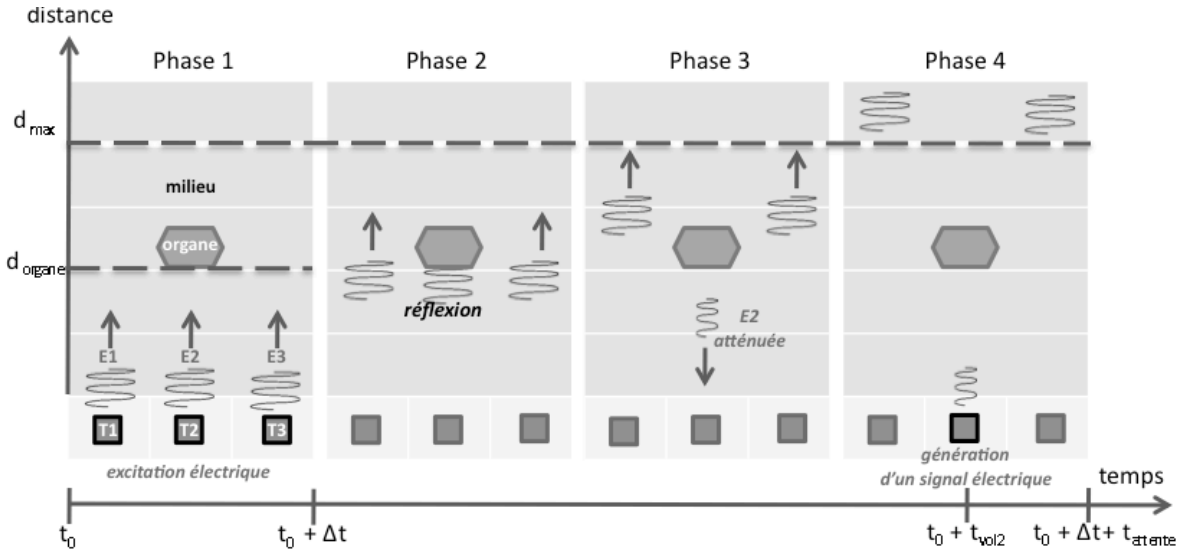


Figure 2.2 : Schématisation d'un cycle d'émission/réception. A t_0, un transducteur à trois éléments (T1, T2, T3) émet pendant un temps Δt, 3 impulsions ultrasonores[24] (E1, E2, E3) dans un milieu d'impédance acoustique Z_1 où se trouve un organe d'impédance acoustique Z_2 (avec $Z_1 \neq Z_2$) (phase 1). L'onde E2 est réfléchie à l'interface avec l'organe (phase 2), et atténuée proportionnellement à la différence Z_1-Z_2 (phase 3). A t_0+t_{vol2}, T2 reçoit E2 (phase 4). A $t_0+\Delta t+t_{attente}$, E1 et E3 ont dépassé d_{max}, la profondeur maximale d'exploration.

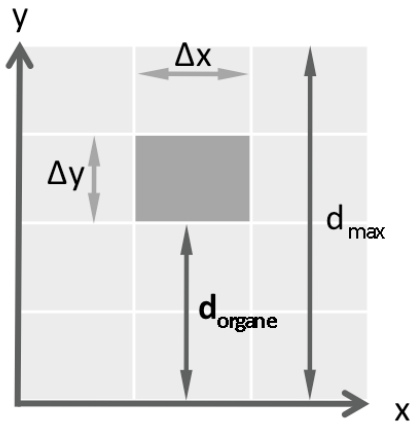


Figure 2.3 : Image ultrasonore formée dans le cadre de la Figure 2.2. La distance d_{organe} est déduite du temps de vol t_{vol2} de l'écho reçu par l'élément piézoélectrique d'abscisse x=2 par $d_{organe}=c(t_0+t_{vol2})/2$. Δy et Δx sont respectivement la résolution axiale et la résolution latérale, explicitées dans la section suivante.

Résolutions spatiale et temporelle

La résolution spatiale d'un système d'imagerie est définie par la distance minimale qui doit séparer deux objets ponctuels pour que ces derniers soient visibles de façon distincte dans l'image. Dans le cadre d'un système d'échographie, il est nécessaire de différencier résolution

[24] Comme il le sera décrit dans le paragraphe suivant, les éléments piézoélectriques ne sont en pratique, pas excités simultanément.

axiale et résolution latérale. La résolution axiale de l'image Δy est la résolution dans l'axe du faisceau ultrasonore ; elle est déterminée par la relation: $\Delta y = c\Delta t$ où Δt est la durée de l'impulsion ultrasonore (voir Figure 2.2 et Figure 2.3). Cette durée dépend principalement des caractéristiques fréquentielles de l'impulsion ultrasonore générée et donc des propriétés des éléments piézoélectriques (bande passante de la sonde). L'image apparaît d'autant plus fine que la gamme des fréquences portées par l'impulsion ultrasonore est large (plus l'impulsion est brève, plus son occupation spectrale est large). Cependant, les composantes « haute-fréquence » de l'onde sont plus atténuées, ce qui limite la profondeur d'exploration. Il s'agit ici du compromis « résolution axiale / profondeur d'exploration » qui sera discuté à la section 2.3.1, dans le cadre de l'imagerie du conduit vocal. En pratique, les sondes proposent un choix de plusieurs bandes de fréquence, ce qui autorise l'analyse par ultrasons d'un milieu à des profondeurs variées (sans avoir besoin de changer de sonde).

La résolution latérale (ou angulaire) qui mesure la capacité de l'échographe à imager distinctement deux objets ponctuels situés dans un plan perpendiculaire au faisceau d'onde ultrasonore, dépend de la largeur de ce faisceau (*i.e.* de l'occupation spatiale de l'onde ultrasonore). Pour la réduire, une procédure de focalisation électronique est utilisée ; elle consiste à donner au front d'onde une forme concave en jouant sur des retards à l'excitation de sous-groupes d'éléments piézoélectriques adjacents. Il est possible de créer plusieurs niveaux de focalisation du faisceau ultrasonore (jusqu'à 3 pour les systèmes actuels). En pratique, les sous-groupes d'éléments ne sont pas excités simultanément mais les uns après les autres, en faisant « glisser » une fenêtre d'excitation le long de la barrette d'éléments ; il s'agit donc d'un balayage électronique. Si d est la distance de focalisation, l la largeur de l'élément piézoélectrique, et λ la longueur d'onde, un ordre de grandeur de la résolution latérale Δx est donné par $\Delta x \approx 1.2(\lambda d)/l$.

La résolution temporelle d'un système échographique, ou fréquence de répétition des images (exprimée en *Hz*), dépend essentiellement de la profondeur d'exploration maximale. Comme l'illustre la Figure 2.2, la formation d'une image ne peut se faire qu'une fois le temps d'attente atteint ; il s'agit du compromis « résolution temporelle / profondeur d'exploration ». Si n est le nombre d'éléments du transducteur balayé électroniquement, d, la profondeur maximale d'exploration, et c la vitesse de propagation dans les tissus mous, alors la cadence d'images f_{fps} est de l'ordre de $f_{fps} = (2nd/c)^{-1}$. Par ailleurs, on notera également que la cadence des images diminue (fortement) lorsque le nombre de zones de focalisation augmente.

Enfin, certains systèmes permettent aujourd'hui d'acquérir des images 3D, en intégrant dans le transducteur non plus une barrette, mais une matrice d'éléments piézoélectriques. Les techniques d'échographie 3D sont actuellement en plein développement mais les systèmes existants ne peuvent fournir un flux d'images qu'à une cadence faible. L'utilisation de tels systèmes n'est donc pas envisagée dans le cadre de cette étude.

2.3. L'analyse du conduit vocal par imagerie ultrasonore

2.3.1. Configuration et positionnement du matériel

Les sections suivantes décrivent les problèmes relatifs à la configuration de l'échographe et au positionnement de la sonde ultrasonore pour l'imagerie ultrasonore du conduit vocal.

Choix du couple échographe/sonde

Il n'existe à ce jour aucun système matériel d'échographie dédié « exclusivement » à l'analyse du conduit vocal ; des systèmes standards de diagnostic médical sont traditionnellement utilisés. Néanmoins, un modèle en particulier pourra être choisi pour sa capacité d'interfaçage avec d'autres capteurs, comme un microphone par exemple. L'autre choix important est celui de la sonde ; il implique le choix de sa forme géométrique et celui de sa bande passante. Pour pouvoir imager une grande partie de la langue, idéalement de la racine à l'apex, l'utilisation d'une sonde convexe (radiale) est appropriée. Certaines sondes, dites microconvexes, permettent d'atteindre un angle d'ouverture de l'ordre de 140 degrés et sont suffisamment petites pour s'insérer de manière assez stable entre les os de la mandibule. La plupart de ces sondes offrent une large bande passante (de 3 à 8 MHz), le choix de la fréquence de travail étant ajustable.

Disposition de la sonde par rapport à la tête

Dans le système envisagé, la sonde est placée sous le menton du locuteur. Afin de maintenir la transmission de l'onde acoustique, la sonde doit rester en contact avec la mâchoire au cours de la production. Un gel, dont l'impédance acoustique est proche de celle des tissus mous, est utilisé pour éviter la formation d'une mince couche d'air entre la sonde et les tissus[25]. Pour le positionnement de la sonde par rapport à la tête, deux approches peuvent alors être envisagées.

Une première approche consiste à fixer, de manière rigide, à la fois la sonde et la tête, comme c'est le cas dans le système HATS (Stone et Davis, 1995), illustré à la Figure 2.4 (et également dans le système utilisé dans (Davidson, 2006)). Pour ne pas trop perturber la production, et permettre un léger mouvement de la mâchoire inférieure, il est possible d'utiliser un petit coussinet rempli de gel ultrasonore placé sur la sonde. Ce dernier se déforme (légèrement) en fonction de la pression imposée par la descente de la mâchoire inférieure, tout en maintenant le contact acoustique.

[25] Dans le cadre de la réalisation d'un prototype, la nécessité d'un gel de contact entre la sonde et la peau peut être perçue comme une contrainte importante. Toutefois, plusieurs approches peuvent être envisagées pour s'en affranchir. On citera ici l'existence de matériaux nommés « gels d'alcool polyvinylique » (en anglais, *Polyvinyl alcohol gel* ou PVA), d'aspect caoutchouteux, utilisables pour le couplage acoustique d'une sonde ultrasonore (Hayakawa *et al.*, 1989).

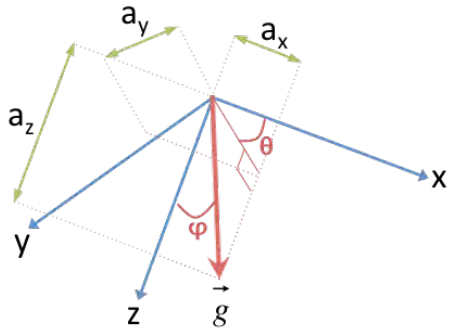
Figure 2.4 : Système de fixation « tête-sonde » HATS

Les systèmes de type « tête et sonde fixes » sont néanmoins assez contraignants pour le locuteur, pour qui une longue session d'acquisition peut s'avérer pénible. Pour rendre au sujet d'étude une certaine liberté de mouvement, il est également possible de fixer le transducteur à un casque, comme celui développé par Alan Wrench (Wrench *et al.*, 2007). Ce dernier maintient le référentiel « tête-sonde » fixe, tout en permettant un mouvement de la tête (et du corps).

La seconde approche consiste à laisser un des éléments du référentiel « tête-sonde » libre de bouger. La première possibilité consiste à fixer de manière rigide la sonde ultrasonore et à laisser la tête du sujet libre. Pendant la production, la mâchoire inférieure ne peut s'abaisser car elle est contrainte par la sonde : c'est alors la partie supérieure du crâne qui bascule vers l'arrière. La seconde possibilité consiste, à l'inverse, à maintenir la tête du sujet immobile, mais sans y attacher de manière rigide la sonde (cas d'une sonde tenue à la main, par exemple). Cette dernière se trouve alors libre d'accompagner les mouvements de la mâchoire. Dans ce cas, le mouvement observé est le mouvement relatif de la langue par rapport à la mâchoire inférieure. Dans ces deux cas, il peut s'avérer nécessaire d'estimer les mouvements relatifs de la sonde et de la tête (par exemple pour obtenir la véritable position de la langue par rapport au palais, ou pour recaler les images ultrasonores). Dans ce but, plusieurs approches ont été proposées. Dans le système *Palatron* (Mielke *et al.*, 2005), des marqueurs ponctuels, repérant les positions de la tête et du transducteur, sont superposés à l'image ultrasonore à l'aide de la technique d'incrustation dite par « fond bleu », couramment utilisée pour les effets spéciaux cinématographiques. Dans le système *Hocus* (Whalen *et al.*, 2005), des diodes infrarouges sont collées sur la tête et sur la sonde, et les informations fournies par un système de suivi optique (en anglais, *optotracking*) permettent la correction de l'image ultrasonore. Enfin, dans le système proposé par Aron (dans le cadre du projet européen ASPI[26]), la même tâche de suivi des mouvements « tête-sonde » est effectuée par articulographie électromagnétique (Aron *et al.*, 2007).

[26] ASPI : *Audiovisual-to-articulatory speech inversion* (http://aspi.loria.fr/)

Nous proposons ici une autre méthode pour mesurer les mouvements relatifs de la tête par rapport à la sonde[27]. Cette dernière est basée sur l'utilisation de capteurs inertiels du type « accéléromètres trois-axes ». Un accéléromètre est un capteur qui, fixé à un objet, permet de mesurer l'accélération de ce dernier. Cependant, en régime quasi-statique, un accéléromètre peut être utilisé comme inclinomètre. En effet, comme l'illustre la Figure 2.5, les accélérations mesurées par le capteur peuvent être interprétées comme les coordonnées du vecteur gravitationnel $\vec{g}$ dans le repère associé à l'accéléromètre (angles θ et φ).



Accélérations mesurées

$$\begin{cases} a_x = \|\vec{g}\| \sin\varphi \cos\theta \\ a_y = \|\vec{g}\| \sin\varphi \sin\theta \\ a_z = \|\vec{g}\| \cos\varphi \end{cases}$$

Figure 2.5 : Mesure de l'orientation d'un solide à l'aide d'un accéléromètre 3-axes.

Considérons à présent deux accéléromètres, attachés respectivement aux solides représentés par les repères (tridimensionnels) R_1 et R_2. Les accélérations mesurées par chacun des capteurs permettent d'exprimer les coordonnées du vecteur gravitationnel $\vec{g}$ dans chacun des repères. Or, la connaissance des coordonnées d'un même vecteur dans deux repères distincts, suffit à déterminer l'orientation relative de ces repères, à la rotation près autour de ce vecteur. En effet, le mouvement relatif d'un solide par rapport à un autre solide, dans l'espace, est entièrement déterminé par trois rotations[28]. La connaissance des coordonnées du vecteur $\vec{g}$ dans chacun des repères permet de déterminer deux de ces rotations. En effet, décrire la position relative de deux repères revient à décrire les rotations à effectuer pour les confondre. Ainsi, superposer un des vecteurs de base de R_1 avec un des vecteurs de base de R_2, peut s'effectuer en les alignant tous les deux sur la verticale $\vec{g}$. Dans le cas de la Figure 2.5, une rotation d'axe $\vec{z}$ d'angle θ puis une rotation d'angle φ et d'axe $\vec{y}$, permet effectivement de rendre le vecteur $\vec{x}$ colinéaire à $\vec{g}$. Néanmoins, pour déterminer complètement l'orientation relative des repères R_1 et R_2, la connaissance d'un troisième paramètre est nécessaire pour déterminer la dernière rotation à effectuer dans le plan orthogonal à $\vec{g}$ (Figure 2.6).

[27] La méthode proposée a fait l'objet d'un dépôt de brevet (Hueber *et al.*, 2009b).

[28] Le mouvement absolu d'un solide par rapport à un autre est lui entièrement déterminé par trois rotations et trois translations.

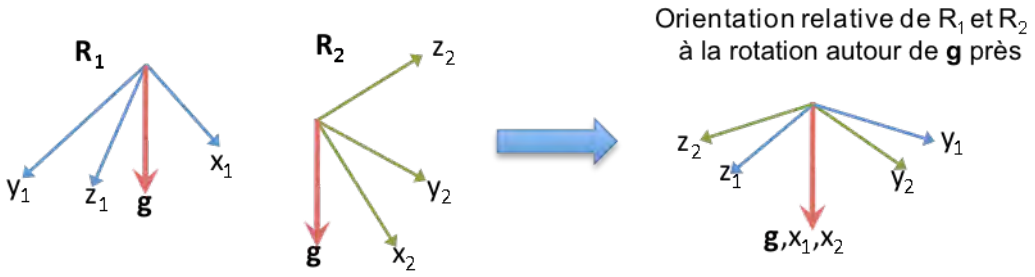


Figure 2.6 : Détermination de l'orientation relative de deux repères

Ainsi, il est possible, à l'aide de deux accéléromètres trois-axes, de déterminer deux des trois angles qui caractérisent l'orientation relative de deux solides. Ce principe est maintenant utilisé pour la mesure de l'orientation de la sonde ultrasonore par rapport à la tête, comme l'illustre la Figure 2.7.

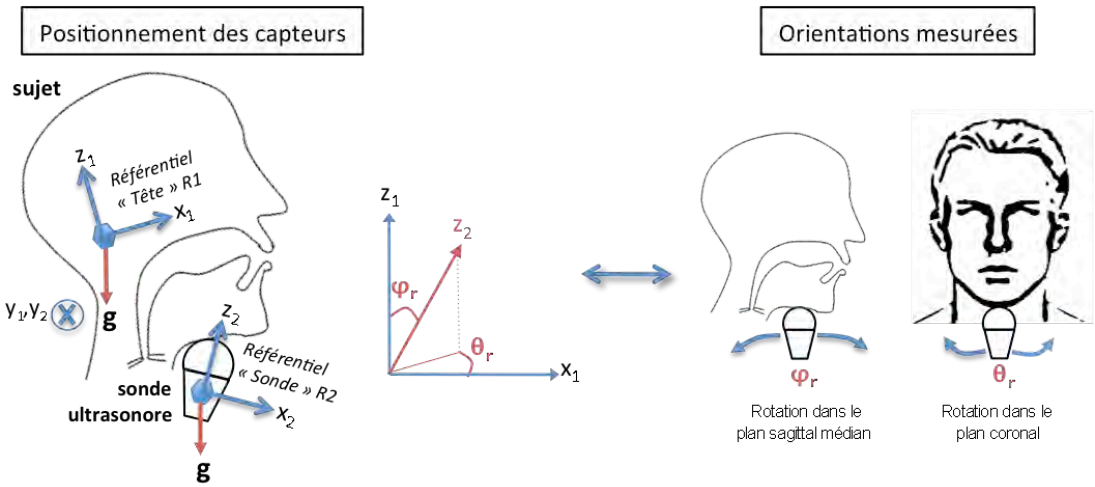


Figure 2.7 : Mesure de l'orientation relative de la sonde et de la tête à l'aide de deux accéléromètres « trois-axes »

En disposant les deux accéléromètres de telle sorte qu'un des axes du premier soit colinéaire à un des axes du second ($\vec{y_1}$ et $\vec{y_2}$ dans la Figure 2.7), il est possible de mesurer, de façon séparée, les mouvements de rotation dans le plan sagittal médian et ceux dans le plan coronal. Bien entendu, toute rotation de la sonde d'axe vertical (colinéaire à $\vec{g}$) ne peut être mesurée[29]. Un exemple de réalisation est présenté à la Figure 2.8 ; l'accéléromètre solidaire des mouvements du crâne est fixé sur une paire de lunettes. Enfin, cette méthode de mesure est autonome ; elle ne nécessite aucune « référence » externe. Aussi, sa mise en œuvre peut être envisagée dans le cadre d'un dispositif portable. La précision de la méthode proposée est discutée en Annexe A.

[29] L'angle de rotation autour de l'axe vertical pourrait être déterminé à l'aide d'un capteur gyroscopique.

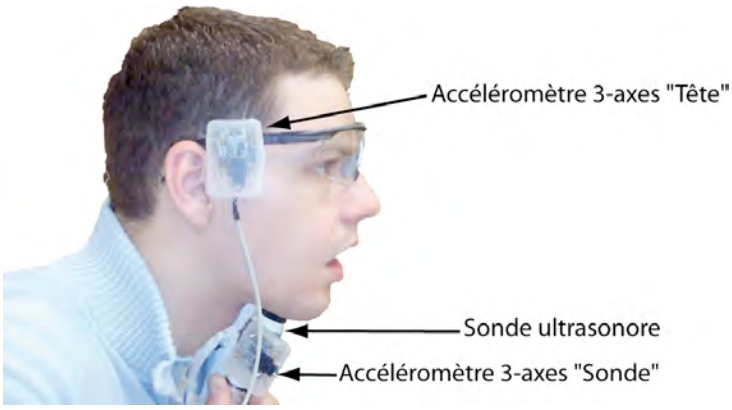


Figure 2.8 : Réalisation du système de mesure de l'orientation relative de la sonde ultrasonore et de la tête à l'aide de deux accéléromètres « trois-axes »

Configuration de l'échographe

Un des premiers paramètres à choisir sur un échographe est la profondeur maximale d'exploration. Comme nous le verrons dans la section 2.3.2, la structure visible la plus éloignée de la sonde est, sous certaines conditions, une partie de la voute palatine. Ainsi, pour un sujet adulte, une distance de 7cm semble appropriée (cette valeur peut être inférieure pour une étude chez l'enfant). A cette profondeur, pour une sonde à 128 éléments, et avec une unique zone de focalisation du faisceau, la cadence des images est d'environ 85 Hz. Pour éviter une trop forte atténuation du signal, la fréquence de l'onde ultrasonore émise est choisie dans la partie basse de la bande passante disponible (3 MHz par exemple). A cette fréquence, avec une largeur des éléments piézoélectriques de l'ordre de 0.5 mm, et une distance de focalisation de 5 cm (surface de la langue), les résolutions latérales et axiales sont respectivement de l'ordre de 2 mm et 0.7 mm.

Enfin, la plupart des systèmes existants effectuent, par défaut, un post-traitement des images dans le but d'en améliorer la qualité. Un de ces traitements consiste à moyenner l'image courante avec un certain nombre d'images adjacentes (en anglais, *image persistence*), ce qui permet de renforcer les contours dans l'image. Ce traitement est ici désactivé car il peut rendre plus difficile la localisation temporelle d'un événement dans la série d'images enregistrée.

2.3.2. Analyse qualitative de l'image ultrasonore

Lorsque la barrette d'éléments piézoélectriques est orientée dans le sens de la longueur de la langue, on obtient une coupe de la cavité buccale dans le plan sagittal médian, comme l'illustre la Figure 2.9.

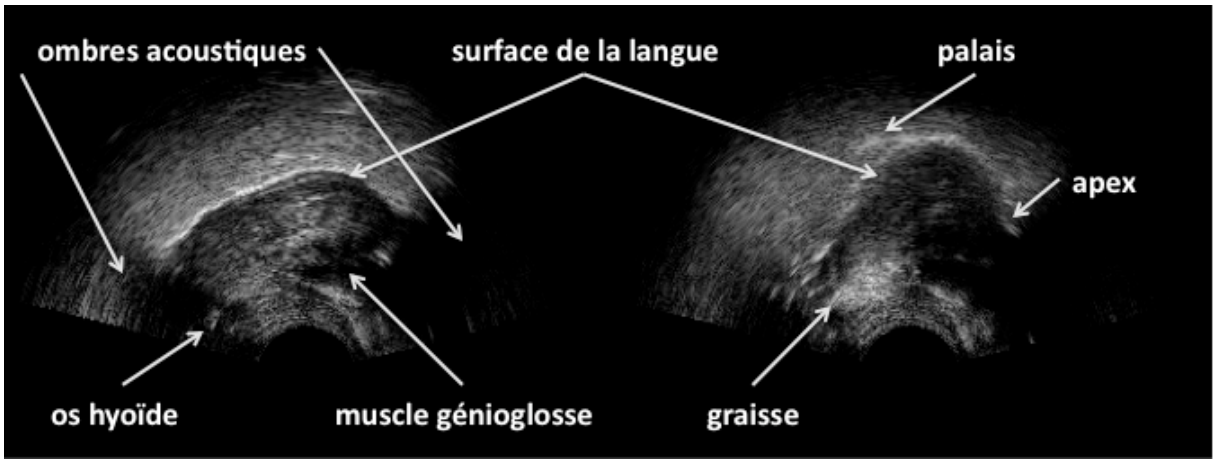


Figure 2.9 : Images ultrasonores de la langue dans le plan sagittal médian (position « relâchée » à gauche, lors d'un [k] à droite)

La structure la plus évidente dans ce type d'image est la surface supérieure de la langue qui est délimitée par la limite inférieure de la longue et fine bande très échogène. La surface de la langue est interrompue de part et d'autres par deux zones sombres, nommées « ombres acoustiques » dans la Figure 2.9. Ces ombres sont causées par la présence, sur la trajectoire du faisceau ultrasonore, de l'os hyoïde pour celle de gauche, et de l'os de la mâchoire pour celle de droite. Les tissus osseux ayant une impédance acoustique bien supérieure à celle des tissus mous, ils réfléchissent presque entièrement le faisceau ultrasonore et empêchent donc sa transmission. Ainsi, l'os hyoïde peut apparaître sous la forme d'une tâche lumineuse, à la base de l'ombre (voir Figure 2.9). De plus, l'ombre acoustique de la mâchoire peut parfois être à l'origine d'une occultation de l'apex. L'absence d'une information fiable sur la position du bout de la langue peut constituer une forte limitation pour une analyse précise des mouvements linguaux. Chez certains sujets, il est également possible d'observer le muscle génioglosse, identifiable par sa forme caractéristique en éventail (voir la Figure 2.9 et la Figure 1.3). Néanmoins, l'orientation oblique des fibres de ce muscle par rapport à l'axe du faisceau ultrasonore incident, limite fortement l'échogénicité de ces dernières. Déviée d'un angle identique à l'angle d'incidence (Figure 2.1), l'onde réfléchie ne retourne pas vers l'élément piézoélectrique émetteur : aucun « écho de réflexion » n'est donc observé. Ne sont alors perçus que les « échos de diffusion » dus à la réémission isotrope de l'onde incidente. L'amplitude de ces derniers étant néanmoins plus faible que celle des échos de diffusion, l'interface est mal imagée. Selon ce même principe, la surface de la langue peut parfois « disparaître » de l'image lorsque cette dernière se positionne de façon très oblique par rapport à l'axe du faisceau ultrasonore incident, lors de l'articulation de la voyelle [iy] par exemple (comme dans *beet*, en anglais).

Par ailleurs, le palais et le velum ne sont presque jamais visibles en imagerie ultrasonore. L'impédance acoustique de l'air étant bien inférieure à celle des tissus mous, le faisceau ultrasonore est presque entièrement réfléchi par la surface de la langue. Ainsi, le palais et le velum ne sont visibles que lors d'un contact avec cette dernière. Sans air, le faisceau ultrasonore

est en effet libre de traverser les tissus de la langue puis les muqueuses de la cavité buccale, jusqu'à atteindre le palais où il est réfléchi presque entièrement, rendant ce dernier visible sur l'image. Ceci se produit notamment lors de la déglutition, pendant laquelle la langue postérieure épouse la forme du palais, et lors de l'articulation de certaines consonnes palatales ou vélaires (ce qui est le cas à la Figure 2.9).

Enfin, une coupe de la cavité buccale dans le plan coronal est obtenue en plaçant la barrette d'éléments piézoélectriques dans le sens de la largeur de la langue. Ce mode de visualisation, qui laisse apparaître la surface de la langue et révèle les détails du septum lingual (fente médiane), reste assez peu utilisé car il ne permet pas la visualisation simultanée des parties postérieures et antérieures de la langue lors de la production. Aussi, l'utilisation d'images dans le plan coronal n'est pas envisagée dans le cadre de cette étude.

2.4. Construction des bases de données audiovisuelles

2.4.1. Dispositif expérimental

Pour la constitution des bases de données audiovisuelles, il est nécessaire d'acquérir de façon synchrone, le flux d'images ultrasonores, le flux d'images vidéo et le signal audio. Un schéma du dispositif expérimental nécessaire est proposé à la Figure 2.10.

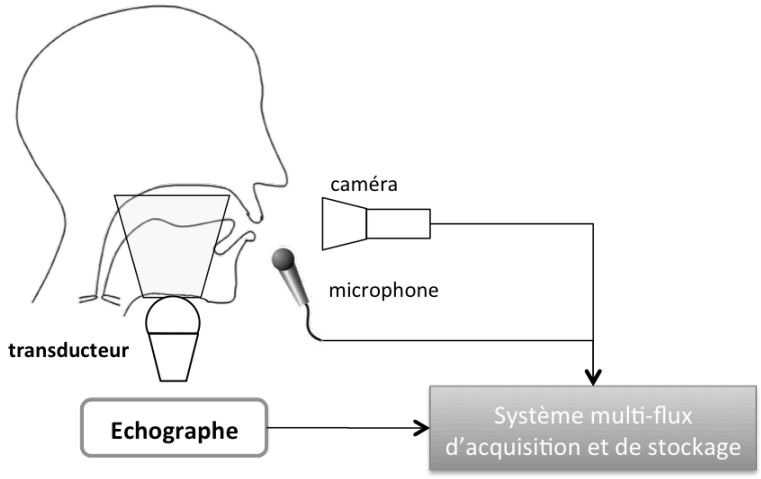


Figure 2.10 : Schéma du dispositif expérimental d'acquisition des données ultrasonores, vidéo et audio.

Dans le cadre de ce travail, deux dispositifs expérimentaux ont été mis en œuvre. Le premier, conçu et réalisé par Maureen Stone au *Vocal Tract Visualization Laboratory* à l'Université de Maryland (USA), s'appuie sur le système de fixation « tête-sonde » HATS (voir la Figure 2.4). Ce système, très robuste, présente néanmoins certaines contraintes, qui seront décrites dans les paragraphes suivants. C'est pourquoi un second dispositif expérimental a été développé dans le cadre du présent travail. Dans la suite de ce document, ces deux systèmes seront identifiés par les acronymes SA1 et SA2.

La mise en place d'un dispositif expérimental tel que celui présenté à la Figure 2.10, pose le délicat problème de la récupération des données visuelles en provenance de l'échographe et leur synchronisation avec les autres flux de données. En effet, un échographe standard à usage médical n'est pas conçu pour s'interfacer « facilement » avec d'autres systèmes d'acquisition. Deux types d'accès aux données ultrasonores sont généralement proposés.

Le premier s'appuie sur la mise à disposition, en temps réel, du flux d'images ultrasonores sur une sortie vidéo analogique de type *S-Video*. En utilisant un mélangeur vidéo et un ordinateur équipé d'une carte d'acquisition, il est alors possible d'enregistrer et de synchroniser le flux ultrasonore avec le flux vidéo et le flux audio. C'est cette technique qui est utilisée dans le système SA1. Cependant, le transport du flux sur ce type de connectique est effectué selon les protocoles NTSC ou PAL qui imposent respectivement une cadence d'images de 29.97 Hz et 25 Hz. Ce problème de la cadence des images ultrasonores sera discutée dans les sections suivantes. De plus, comme il a été décrit dans (Wrench et Scobbie, 2006), le processus de sous-échantillonnage en temps réel du flux d'images ultrasonores peut être à l'origine d'artéfacts parfois importants (comme un dédoublement du contour de la surface de la langue).

Le second procédé permettant la récupération des données ultrasonores, s'appuie sur l'accès à une zone de mémoire tampon de l'échographe, communément nommée *cineloop*. Cette mémoire de type FIFO (*first in, first out*), contient en permanence les *n* dernières images qui ont été acquises (sur les systèmes actuels, *n* est de l'ordre de 1000). La récupération des données de cette mémoire permet d'obtenir des séquences à pleine résolution temporelle et sans artéfact. Cependant, sur la plupart des échographes, l'accès à ces données ne peut (à ce jour) pas s'effectuer en temps réel, mais uniquement après avoir ordonné manuellement leur transfert sur un disque dur interne à l'échographe, dont le contenu est récupérable depuis un autre poste (généralement par protocole *Ethernet*, au format standard de l'imagerie médicale *DICOM*). De ce fait, la synchronisation avec d'autres flux d'informations est difficile, mais néanmoins possible, comme l'a montré Aron dans (Aron *et al.*, 2007) et Miller dans (Miller, 2008).

Récemment est apparu une nouvelle génération d'échographes, construits comme des périphériques informatiques, connectables à un ordinateur via des interfaces standard comme le *FireWire* ou l'*USB*. Ces appareils sont souvent de taille réduite car ils sont destinés à une utilisation « nomade ». Grâce notamment aux kits de développement logiciel (en anglais, *Software Development Kit* ou SDK) fournis par le constructeur, il est désormais possible de construire des applications *ad hoc*, permettant de piloter l'échographe et de récupérer les données ultrasonores, en temps réel, à leur résolution temporelle et spatiale maximale. C'est cette nouvelle possibilité qui a motivé le développement du système d'acquisition SA2. Ce dernier est basé sur la synchronisation des capteurs ultrasonores, vidéo, audio, et inertiels (accéléromètres), par voie logicielle (Hueber *et al.*, 2008c). Les éléments matériels (capteurs) mis en œuvre dans le système SA2 sont décrits à la Figure 2.11. Le logiciel chargé de leur

synchronisation, développé dans le cadre de ce travail, est nommé *Ultraspeech*[30]. Un aperçu de ce dernier est présenté à la Figure 2.12. *Ultraspeech* est une application développée en langage *C++* dont l'interface graphique s'appuie sur les bibliothèques *MFC*[31], et qui utilise les bibliothèques multimédia *Microsoft DirectX* et *OpenCV*[32]. La gestion « simultanée » des différents flux de données est rendue possible par l'utilisation des techniques de programmation parallèle (en anglais, *multithreading*). En interne, chacun des flux est accessible par l'intermédiaire d'une zone mémoire de type FIFO, en écriture pour le capteur, en lecture pour *Ultraspeech*. Lors de l'acquisition, chaque ajout d'une donnée par l'un des capteurs à l'une de ces zones mémoire, fait l'objet d'une interruption logicielle ; la donnée ajoutée est alors étiquetée par la valeur d'une horloge de synchronisation indépendante du capteur concerné. Les différents flux de données, ainsi étiquetés, sont ensuite post-synchronisés puis mis en forme pour être transférés sur le disque dur (séries d'images *Bitmap* pour les flux visuels, fichiers *WAV* pour le flux audio).

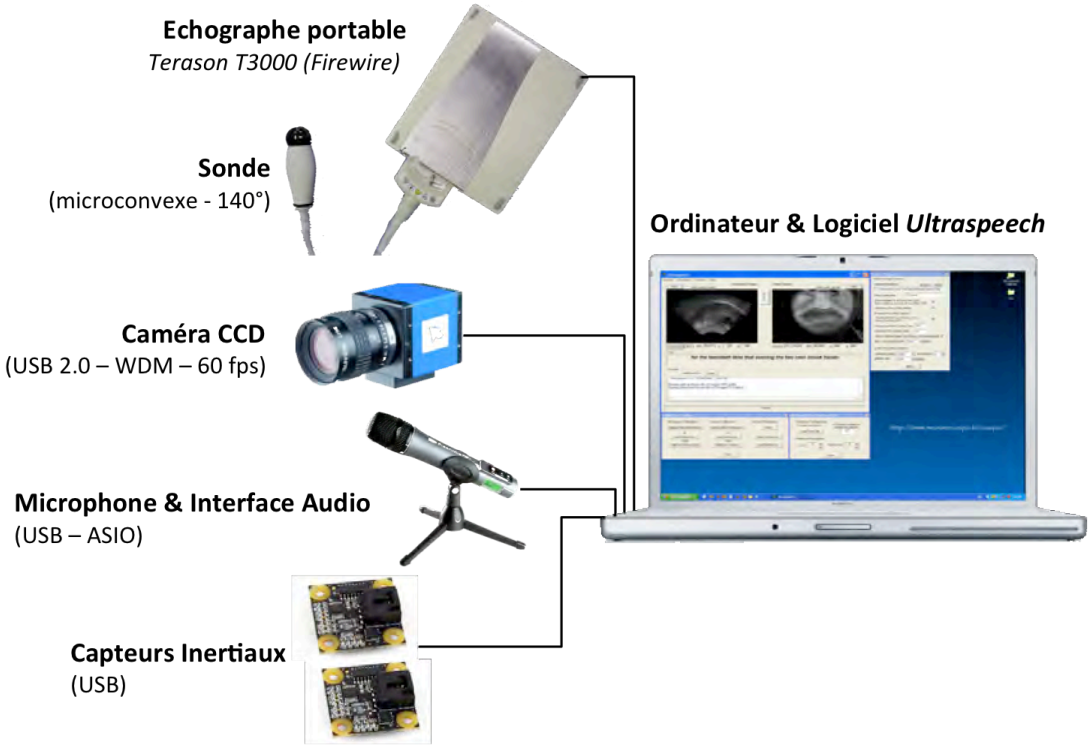


Figure 2.11 : Système d'acquisition multi-flux SA2 – Éléments matériels

[30] Le logiciel *Ultraspeech* a fait l'objet d'un dépôt à la Société des Gens de Lettres (http://www.sgdl.org)

[31] MFC : *Microsoft Foundation Classes*

[32] Librairie graphique *Open Source* - http://opencvlibrary.sourceforge.net

Figure 2.12 : Système d'acquisition multi-flux SA2 – Logiciel d'acquisition multi-flux *Ultraspeech*

Afin de valider cette approche de synchronisation par voie logicielle, un protocole expérimental simple est mis en œuvre. Ce dernier tente de recréer dans les différents flux, les conditions d'un « clap » de cinéma. Il s'agit de produire un événement le plus bref possible, identifiable dans chacun des flux. Dans le protocole adopté, une masse vient heurter une bouteille munie d'un bouchon-pompe et remplie de gel ultrasonore. Elle propulse ainsi une goutte de ce gel sur un transducteur ultrasonore situé à proximité de la bouteille (le temps de vol de la goutte de gel est négligeable). La caméra vidéo filme l'expérience. L'impact de la masse est facilement repérable sur le signal audio et sur la vidéo de l'expérience, et l'impact de la goutte apparaît clairement sur l'image ultrasonore[33], comme l'illustre la Figure 2.13.

[33] Un procédé similaire est utilisé dans (Aron *et al.*, 2007) ; une tige rigide vient heurter le transducteur ultrasonore immergé dans un bac d'eau. L'instant d'impact de la tige est identifiable à la fois dans le flux ultrasonore et dans le flux audio.

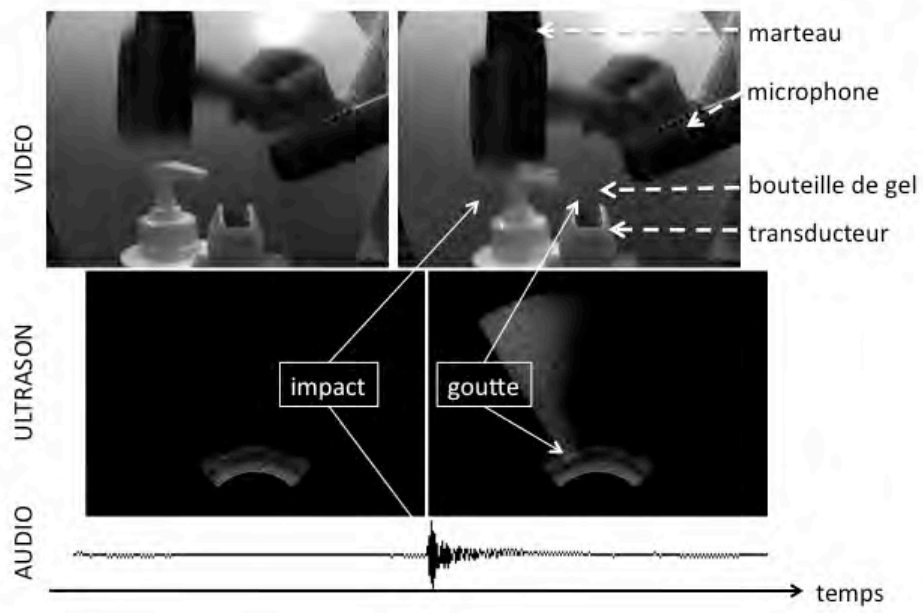


Figure 2.13 : Procédure expérimentale de vérification de la synchronie entre le flux d'images ultrasonores, le flux vidéo et le flux audio.

Le système développé a ainsi pu être validé pour l'enregistrement synchrone :

- du flux d'images ultrasonores à 60 ips (images par seconde).
- d'un flux d'images vidéo à 60 ips.
- du signal audio à 16 kHz (16 bits, mono).

Tout au long du processus d'enregistrement des données, il est nécessaire que les conditions d'observation des gestes articulatoires restent identiques. C'est pour atteindre cet objectif qu'est mis en œuvre le dispositif expérimental présenté à la Figure 2.14.

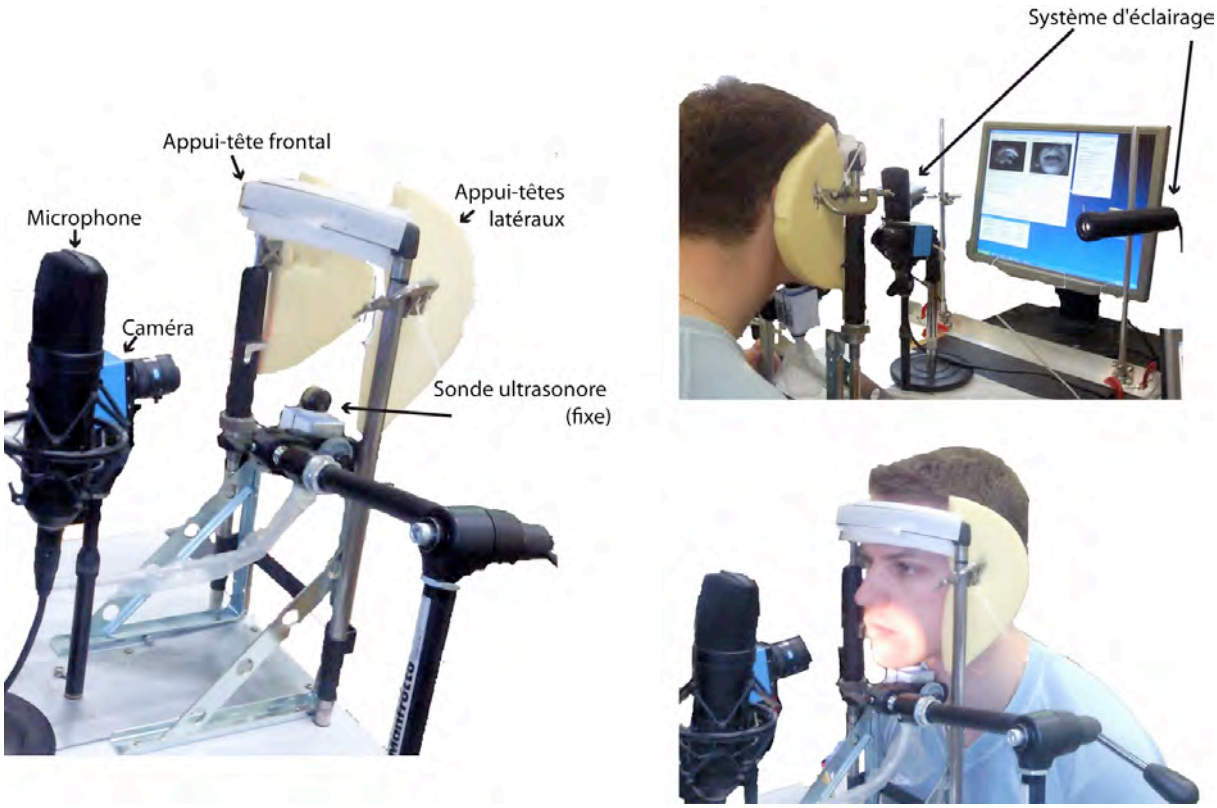


Figure 2.14 : Fixation « tête-sonde » et agencement des capteurs dans le système SA2

Les différents capteurs (sonde ultrasonore, caméra vidéo et microphone) sont ici fixes. La tête du locuteur est contrainte de rester à la même distance de la caméra (appuie-tête frontal) et dans le même axe (appuie-têtes latéraux). Néanmoins, contrairement au dispositif HATS (Figure 2.4), on notera que le crâne est ici libre de bouger par rapport à la sonde. L'utilisation d'une sonde microconvexe permet néanmoins un léger abaissement de la mâchoire.

Pour rendre possible la constitution de bases de données de taille importante (plusieurs centaines de phrases), il est apparu nécessaire d'envisager l'enregistrement des données en plusieurs sessions distinctes. Cela permet notamment au locuteur de quitter, un temps, la position relativement inconfortable imposée par ce type de dispositif expérimental. Pour maintenir une cohérence dans les données enregistrées, il est nécessaire, au début d'une nouvelle séance, de repositionner le locuteur à la place qu'il occupait à la session précédente. Deux mécanismes sont alors proposés. Le premier est basé sur l'utilisation des capteurs inertiels présentés à la section 2.3.1. A l'aide d'une interface graphique (voir Figure 2.12) permettant la visualisation en temps réel des positions angulaires relatives de la sonde et de la tête, il est demandé au locuteur d'ajuster sa tête de sorte qu'il s'aligne sur des positions de référence. Le second mécanisme proposé s'appuie sur la création, en temps réel, d'une image dite de « contrôle », obtenue en moyennant l'image vidéo correspondant à la position cible, avec l'image vidéo de la position actuelle. Ainsi, plus la position courante du locuteur est différente de la position cible, plus l'image de contrôle apparait « floue ». Pour se repositionner correctement, le locuteur ajuste la position de sa tête de sorte que l'image obtenue sous aussi nette que possible,

comme l'illustre la Figure 2.15. Ce procédé est également utilisé pour le contrôle de la position de la tête par rapport à la sonde ; l'objectif est alors, pour le locuteur, de superposer dans l'image, le contour de sa langue à un contour de référence, obtenu par exemple en vocalisant un phonème particulier.

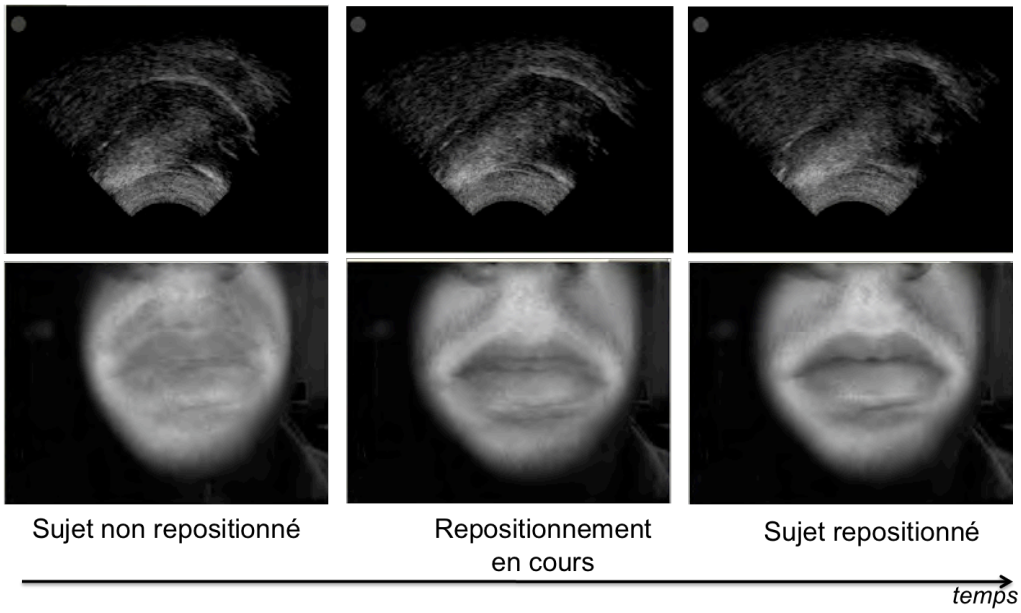


Figure 2.15 : Procédure interactive de repositionnement du locuteur dans le dispositif expérimental SA2, entre deux sessions d'acquisition

Ainsi, deux systèmes d'acquisition ont été mis en œuvre dans le cadre de ce travail. Le tableau ci-dessous résume les spécificités de chacun de ces systèmes.

	Système SA1	**Système SA2**
Echographe (marque)	Aloka SSD 1000	Terason T3000
Sonde (type, angle d'ouverture bande passante)	Convexe, 120° 2.5-6 MHz	Microconvexe, 140° 4-8 MHz
Camera (capteur, type, marque)	Camera CMOS caméscope analogique Sony	Camera CCD vision industrielle Imaging Source
Microphone (type)	Microphone à électret	Microphone à condensateur
Référentiel tête-sonde	Fixe (système HATS)	Mobile (légèrement)
Système de mixage des flux	Analogique	Numérique (logiciel *Ultraspeech*)
Résolution temporelle - Ultrasons	29.97 ips (NTSC)	60 ips
Résolution temporelle - Caméra	29.97 ips (NTSC)	60 ips

Tableau 2.1 : Comparaison des systèmes d'acquisition multi-flux SA1 et SA2

2.4.2. Bases de données enregistrées

Choix de la langue

La langue privilégiée dans le cadre de ce travail est l'anglais. Ce choix est avant tout « historique » ; les premières acquisitions de données ont en effet été réalisées en enregistrant des

étudiants américains de l'Université du Maryland, à l'aide du système SA1. En outre, pour une ICPS basée l'analyse du conduit vocal par imagerie ultrasonore, la langue anglaise peut apparaître comme une meilleure candidate que la langue française. Le velum n'étant visible qu'en cas de contact linguovélaire (voir section 2.3.2), l'imagerie ultrasonore ne renseigne donc que très peu sur la nasalisation. Or, si les deux langues donnent lieu à la réalisation de consonnes nasales, la langue française, à la différence de la langue anglaise, fait également l'objet de voyelles nasales (voir section 1.1). La langue française fera donc *a priori* l'objet d'un nombre plus important de configurations articulatoires difficilement interprétables.

Choix du corpus textuel

La constitution d'une base de données nécessite le choix du texte à faire prononcer par le locuteur. Afin d'être en mesure de modéliser une grande variété de gestes articulatoires, ce texte doit tout d'abord présenter une bonne couverture de l'espace phonétique. De plus, comme indiqué dans le dernier chapitre, une des techniques de synthèse du signal acoustique est basée sur la concaténation de diphones. Aussi, le texte choisi pour la constitution des bases de données, donc pour la construction du dictionnaire de synthèse, doit également présenter une bonne couverture de l'espace des diphones. Enfin, bien que des méthodes aient été proposées pour décomposer le processus d'acquisition des données en plusieurs sessions distinctes, permettant ainsi *a priori* l'enregistrement d'une très grande quantité de données, il reste en pratique préférable de se limiter à un corpus d'une taille « acceptable ». C'est pour ces différentes raisons qu'a été choisi le corpus C*MU Arctic*. Ce corpus, développé dans le cadre du projet de synthèse à partir du texte *Festival (Black et Lenzo, 2000)*, est constitué de 1132 phrases en langue anglaise, sélectionnées majoritairement parmi des ouvrages du projet Gutemberg[34], principalement ceux d'auteurs anglais du début du XX^{ème} siècle. Les propriétés de ce corpus sont rappelées dans le tableau ci-après.

Nombre d'unités	
Nombres de phrases	1132
Nombre de mots	10045
Nombre de mots différents	2974
Nombre de phones	39153
Couverture	
Phones	100 %
Diphones	79.6 %
Triphones	13.7 %

Tableau 2.2 : Propriétés du corpus CMU Arctic (l'anglais est ici décrit par 40 phonèmes)

[34] Projet visant à mettre à disposition gratuitement sous format électronique tous les livres, en langue anglaise, tombés dans le domaine publique (*out-of-copyright*) - http://www.gutenberg.org

Bases de données enregistrées

Dans le cadre de ce travail, deux bases de données ont été construites. La première est constituée avec le système SA1, à partir des enregistrements d'une jeune femme (native des Etats-Unis) à qui il est demandé de prononcer les phrases du corpus CMU Arctic, d'une façon la plus « neutre » possible (seule l'intonation interrogative est conservée). Cette base de données sera identifiée dans la suite du document par le terme B1. Une image caractéristique[35], obtenue à la sortie du mélangeur vidéo analogique utilisé pour combiner les différents flux, est présentée à la Figure 2.16. On notera la présence d'une seconde source vidéo, utilisée pour enregistrer une vue de profil des lèvres du locuteur. L'utilisation de cette vue sera discutée plus loin.

Figure 2.16 : Image extraite de la base de données B1, construite à l'aide du système SA1. Les images vidéo sont incrustées dans l'image ultrasonore à l'aide d'un mélangeur vidéo analogique.

Dans le cadre de cet enregistrement, aucune procédure de repositionnement du locuteur n'est utilisée. Ce dernier est contraint de rester immobilisé dans le système HATS pendant toute la durée de l'enregistrement, soit pendant plus de 2h30 dans notre cas. Pour des raisons liées principalement à la fatigue du locuteur, il n'a pas été possible d'enregistrer la totalité du corpus CMU Arctic ; 1020 phrases, sur les 1132 disponibles, ont néanmoins pu être acquises (soit environ 90 % du corpus). Pour constituer la base de données, une convention est adoptée pour segmenter précisément chaque phrase dans la séquence vidéo enregistrée. Entre deux phrases, il est demandé au locuteur de garder la bouche fermée. N'est ensuite gardé qu'un intervalle délimité par l'image précédant la première ouverture de la bouche et l'image qui suit la dernière fermeture de la bouche. Après cette mise en forme, la base de données B1 contient 61 minutes de parole continue, représentées par 109 553 images ultrasonores et vidéo (29.97 ips).

[35] Un extrait vidéo de la base B1 est disponible sur la page Internet associée à ce manuscrit (URL indiquée dans l'introduction générale).

Une seconde base de données a ensuite été constituée à l'aide du système SA2. Le même corpus CMU Arctic est utilisé ; il est prononcé par une autre locutrice, également native des Etats-Unis. Cependant, grâce aux procédures de repositionnement du sujet présentées à la section 2.4.1, l'acquisition a pu s'effectuer en plusieurs sessions distinctes. Ainsi, pour la constitution de cette seconde base de données, nommée B2, 10 sessions d'enregistrement ont été effectuées (soit environ une centaine de phrases enregistrées par session, deux sessions étant espacées de 24 heures au minimum et de 3 jours en moyenne). Dans ce protocole, la fatigue du locuteur n'étant plus un facteur limitant, la totalité du corpus CMU Arctic a ainsi pu être enregistrée. Après mise en forme des données (à l'aide de la même procédure que celle utilisée pour la base B1), la base de données B2 contient 66 minutes de parole continue, soit 237 764 images (flux visuels cadencés à 60 ips). Un couple d'images (ultrasonore et vidéo) caractéristique[36] de cette base est présenté à la Figure 2.17.

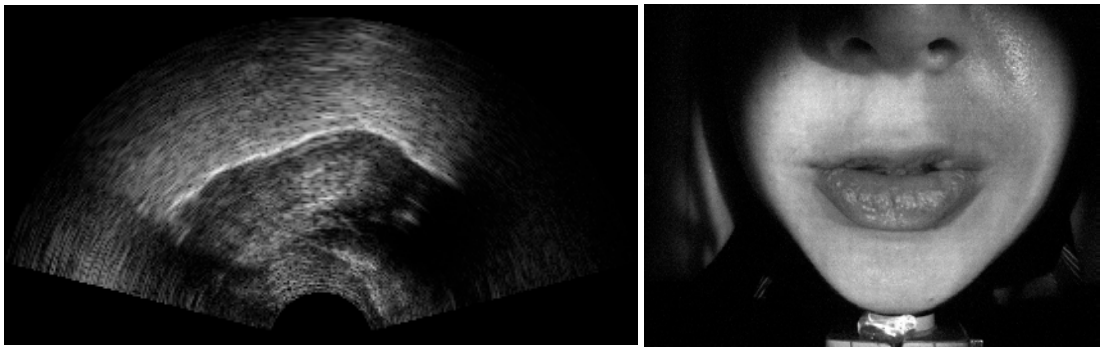


Figure 2.17 : Images extraites de la base de données B2, construite à l'aide du système SA2. A la différence du système SA1, les flux ultrasonores et vidéo sont ici dissociés, de l'acquisition jusqu'au stockage

Pour permettre la modélisation des liens entre le geste articulatoire et le signal de parole, ces différents enregistrements sont effectués en parole vocalisée et non en « articulation silencieuse ». Aussi, afin d'étudier les éventuelles spécificités de ce mode de production, il a donc enfin été demandé à la locutrice de la base B2, d'articuler les 60 premières phrases du corpus CMU Arctic, mais sans les vocaliser (ni même les murmurer).

[36] Un extrait vidéo de la base B2 est disponible sur la page Internet associée à ce manuscrit.

Chapitre 3. Traitement des données audio-visuelles, extraction des caractéristiques

3.1. Avant-propos

Ce chapitre présente les différentes techniques mises en œuvre pour le traitement et l'analyse des données ultrasonores, vidéo et audio, acquises à l'aide des systèmes d'acquisition SA1 et SA2 décrits au chapitre précédent. L'objectif poursuivi ici est la caractérisation de ces données, c'est-à-dire l'extraction automatique de caractéristiques discriminantes (visuelles et acoustiques), dans la perspective de la conversion visuo-acoustique.

3.2. Traitement des images ultrasonores

3.2.1. Pré-traitement – Réduction du bruit de speckle

Le phénomène de diffusion, introduit brièvement à la section 2.2.1, est à l'origine du *speckle*[37], qui donne à l'image échographique un aspect granuleux. Le *speckle* est la conséquence de l'interférence des ondes ultrasonores rétrodiffusées par les nombreuses inhomogénéités contenues dans le milieu biologique étudié (les diffuseurs). Le *speckle* est entièrement déterminé par la distribution des diffuseurs dans ce milieu ; il en est donc une mesure caractéristique. Néanmoins, sa présence peut fortement perturber l'interprétation de l'image, sa segmentation, ou toute autre approche d'extraction de l'information en vue de sa classification. Aussi, le *speckle* est généralement considéré comme un bruit qu'il est préférable de filtrer. Le problème de la réduction du bruit de *speckle* dans les images ultrasonores fait l'objet d'une abondante littérature. Pour une description des diverses approches disponibles, on pourra notamment se référer à (Tauber, 2005).

L'amplitude du bruit de *speckle* augmente avec l'intensité moyenne des pixels de la région concernée. Le *speckle* est donc un bruit multiplicatif. Pour le filtrer, il est préférable d'adapter une approche dite homomorphique, qui consiste à se ramener au traitement d'un bruit additif grâce à une compression logarithmique de l'image préalable[38].

Une des techniques couramment utilisée pour le traitement du *speckle* dans une image échographique est le filtrage par diffusion anisotrope, introduit par Perona et Malik dans (Perona et Malik, 1990). Un filtre de diffusion anisotrope tente d'effectuer un lissage de l'image isotrope dans les zones de réflectivité homogène, et anisotrope à proximité des contours. Ce filtre est un filtre itératif qui modifie l'intensité u d'un pixel en respectant la relation suivante (t est le paramètre d'échelle, c'est à dire l'itérateur) :

[37] *Speckle* peut se traduire en français par « tavelure ». Par la suite, on utilisera exclusivement le terme anglais.

[38] En pratique, cette compression logarithmique est généralement effectuée en interne par l'échographe, sur les échos ultrasonores reçus, avant la formation de l'image.

$$\frac{\partial u}{\partial t} = div(g(|\nabla u|)\nabla u) \text{ avec } g(s) = \frac{1}{1+(\lambda s)^2} \text{ et } \lambda \text{ une constante} \qquad \text{(Équation 3.1)}$$

Le paramètre λ est le facteur d'échelle pour l'amplitude du gradient ; il contrôle le déclenchement du lissage. En effet, en présence d'un contour (variation rapide d'intensité donc gradient fort), $|\nabla u| \gg 1/\lambda$, $g(|\nabla u|) \to 0$ et donc $\partial u/\partial t \to 0$. L'intensité u reste donc inchangée et le contour préservé. A l'inverse, dans une zone homogène (faible variation d'intensité, gradient faible), $|\nabla u| \ll 1/\lambda$, $g(|\nabla u|) \to 1$, et l'équation 3.1 devient :

$$\frac{\partial u}{\partial t} = div(\nabla u) \qquad \text{(Équation 3.2)}$$

Il s'agit là d'une équation classique de diffusion isotrope : les zones d'intensité homogène sont donc lissées. Ce filtre fournit de bons résultats pour le traitement des images optiques traditionnelles car le gradient d'intensité est, dans ce cas, un bon détecteur de contour. En imagerie ultrasonore, une région dont la réflectivité réelle est homogène peut être représentée par un ensemble de pixels dont l'intensité est dispersée (bruit de *speckle*). Le gradient d'intensité est donc soumis à de fortes variations locales même en l'absence de contour. Il ne constitue donc pas un bon détecteur de contour. Un détecteur de contour mieux adapté au *speckle* est le coefficient de variation local γ. Ce dernier est défini sur une fenêtre centrée sur le pixel s par :

$$\gamma^2(s) = \frac{1}{|\eta_s|} \sum_{p \in \eta_s} \frac{(I_p - \bar{I}_s)^2}{\bar{I}_s^2} \qquad \text{(Équation 3.3)}$$

où η_s est le voisinage du pixel s et $\bar{I}_s$ la valeur moyenne des intensités des pixels de η_s. La Figure 3.1 propose une comparaison du gradient d'intensité et du coefficient de variation local dans le cas de l'image ultrasonore du conduit vocal. Le gradient d'intensité, très bruité, apparaît effectivement comme un détecteur de contour peu robuste en présence de *speckle*. Tout en faisant apparaître les structures principales (ici la surface supérieure de la langue), le coefficient de variation local, lui, limite les variations locales d'intensité dans les zones à réflectivité réelle homogène (comme par exemple dans la zone située au dessus de la surface supérieure de la langue).

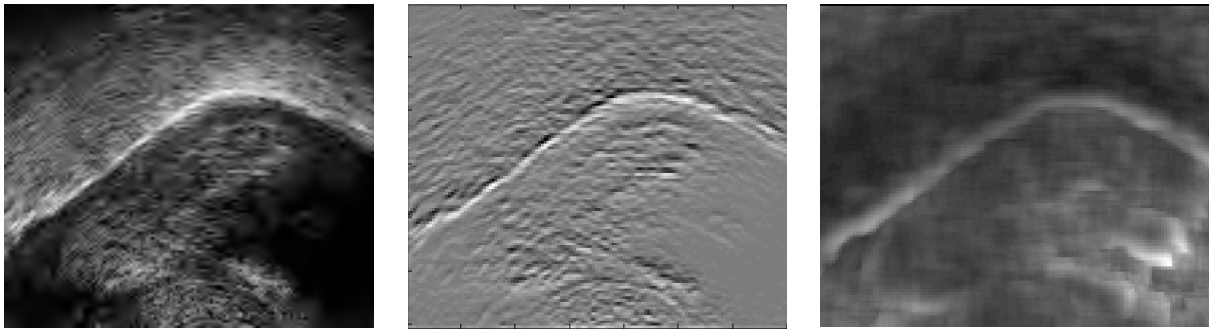


Figure 3.1 : Coefficient de variation local *versus* gradient d'intensité dans une image ultrasonore. De gauche à droite: l'image originale, le gradient d'intensité (calculé à l'aide du masque de Soebel), le coefficient de variation local (voisinage de 3x3 pixels)

Dans (Yu et Acton, 2002) est proposé un filtre de diffusion anisotrope basé sur l'utilisation de ce coefficient de variation local comme détecteur de contour. L'équation qui régit le comportement de ce filtre est donc :

$$\frac{\partial u}{\partial t} = div(g(|\gamma(u)|)\nabla u) \qquad \text{(Équation 3.4)}$$

L'action de ce filtre (implémenté sous *Matlab*) sur une image ultrasonore du conduit vocal est illustrée à la Figure 3.2.

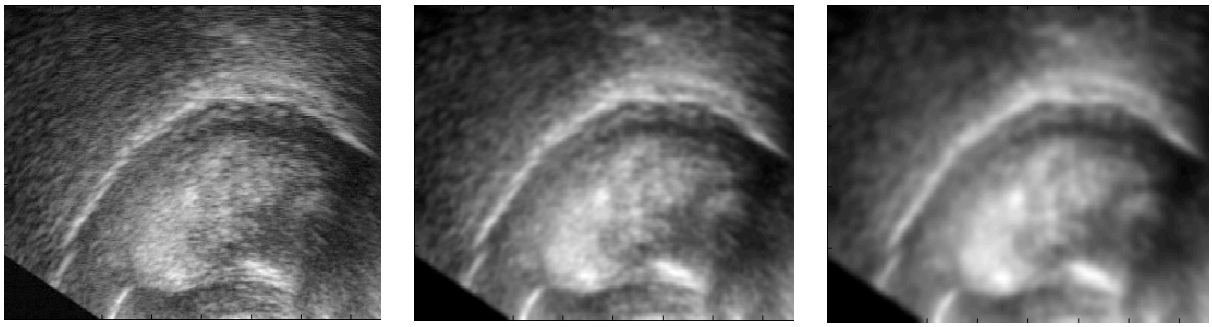


Figure 3.2 : Réduction du *speckle* à l'aide d'un filtre de diffusion anisotrope. De gauche à droite, image originale, image filtrée après 50 itérations, image filtrée après 100 itérations

3.2.2. Extraction des caractéristiques visuelles – Approches par segmentation

La principale structure visible dans l'image ultrasonore est la surface supérieure de la langue, le principal articulateur de la cavité buccale. L'approche la plus intuitive la caractérisation de ce type d'image est la segmentation et le codage de cette structure. Bien que facilitée par la réduction du *speckle*, l'extraction du contour de la surface de la langue reste une tâche relativement difficile. Dans le cadre de cette étude, deux approches par segmentation sont envisagées.

Segmentation par contours actifs

La première approche est basée sur la technique dite de « segmentation par contours actifs », également nommée « *snake* » (Kass, 1987). Cette technique est semi-automatique.

L'opérateur initialise le processus de segmentation en délimitant, de manière grossière, une ligne initiale de contour autour de l'objet à segmenter. La recherche du contour optimal s'effectue ensuite de façon itérative, en minimisant une certaine énergie qui caractérise le contour. Cette énergie est composée d'un terme d'attache aux données et d'un terme de régularisation interne. La minimisation du terme d'attache aux données attire le contour vers les zones à fort gradient. Le terme de régularisation interne vérifie que le contour actif respecte des contraintes, liées non pas à l'image, mais à la physique de l'objet à segmenter (courbure maximale, taille minimale, etc.). La méthode des contours actifs permet donc d'introduire des connaissances *a priori* sur la forme de l'objet à segmenter.

Dans (Li, 2003), Li et Stone ont proposé une formulation du terme d'énergie interne, adaptée à la segmentation de la surface de la langue. La méthode proposée est disponible dans le logiciel *EdgeTrak*. Afin de segmenter une séquence d'images, le contour actif à l'image *n* est initialisé par le contour optimal déterminé à l'image *n-1*. Cette méthode fournit généralement de bons résultats, comme le montre la Figure 3.3, où une série de 30 images est segmentée à l'aide de cet outil (ne sont représentées que trois images uniformément réparties dans la série). Néanmoins, lorsque l'image du contour est trop altérée (apparition de discontinuités, voir la section 2.3.2), la segmentation est imprécise, comme c'est le cas de la dernière image de la Figure 3.3 (image n° 30). Une réinitialisation manuelle du processus de segmentation est nécessaire dans ce cas.

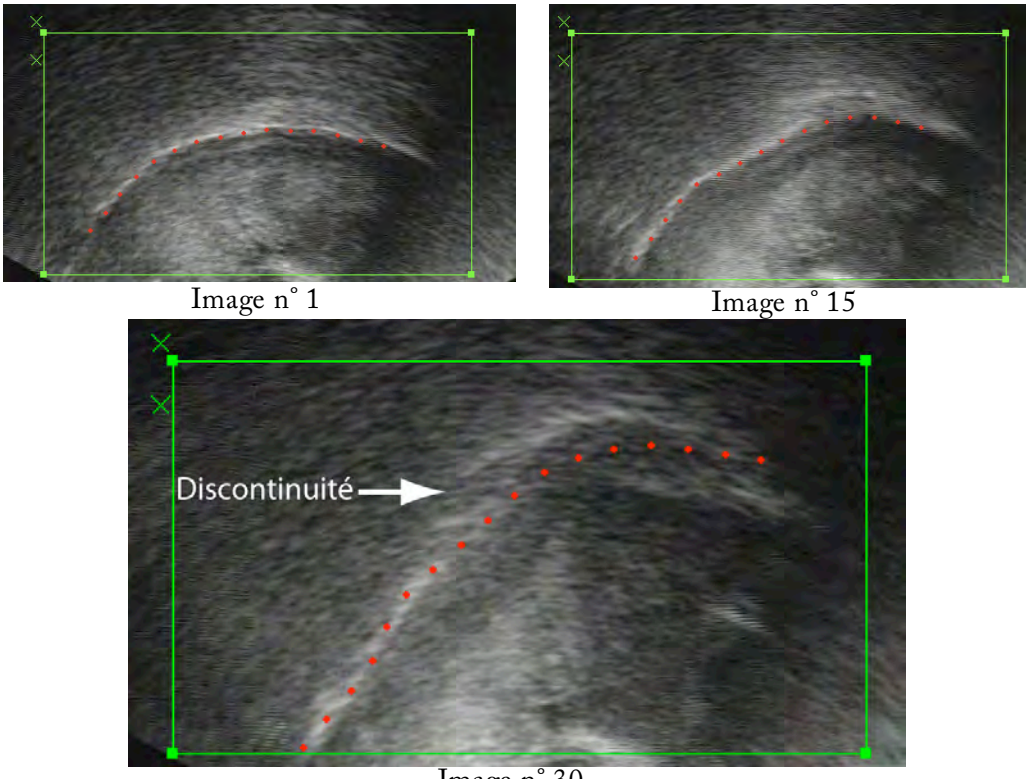


Figure 3.3 : Extraction du contour de la langue par la méthode des contours actifs (à l'aide du logiciel *EdgeTrak)*

Segmentation par recherche des maxima du gradient d'intensité

Pour extraire le contour de la surface supérieure de la langue de façon automatique (sans avoir à initialiser « manuellement » le processus de segmentation), une seconde approche a été proposée dans (Hueber, 2006). Cette dernière se décompose en deux étapes. Tout d'abord, des « points candidats » au contour sont extraits en recherchant les extrema du gradient d'intensité. Puis, la courbe formée par cette suite de points candidats est approchée, au sens des moindres carrés, par une fonction de type « *spline* cubique ». En raison de la présence (résiduelle) du *speckle* et de celle d'autres structures très échogènes, certains points, pourtant très éloignés du contour réel de la surface de la langue, sont déclarés à tort « candidats » et conduisent donc à des approximations aberrantes. Aussi, il a été proposé un algorithme qui vise à résoudre conjointement le problème de l'approximation des points candidats par une fonction *spline* et celui de l'élimination de ces points aberrants. Cet algorithme, qui ne sera pas entièrement redéveloppé ici, est basé sur la recherche, par une approche du type Monte Carlo, de sous-ensembles de points candidats pour lesquels la « *spline* interpolante » satisfait les contraintes géométriques illustrées à la Figure 3.4.

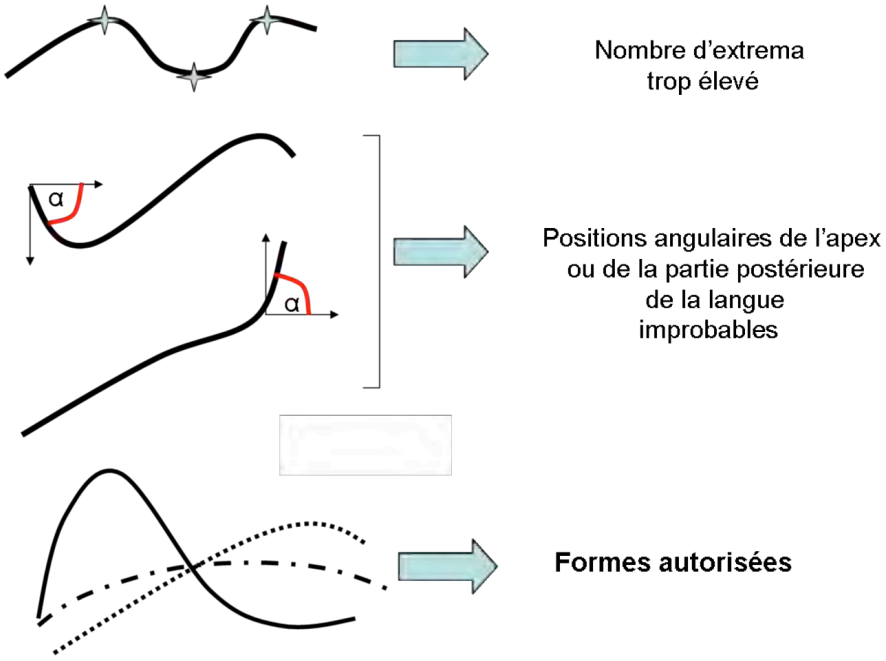
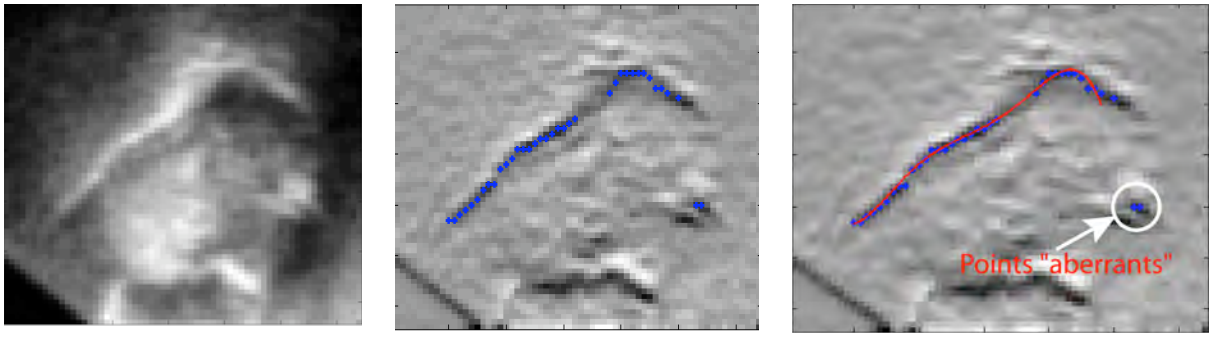


Figure 3.4 : Extraction du contour de la surface supérieure de la langue – Contraintes géométriques sur le contour recherché

Les différentes étapes de cette méthode de segmentation sont résumées à la Figure 3.5.



1) Réduction du *speckle* par filtrage de diffusion anisotrope (100 itérations)

2) Recherche de « points candidats » (en bleu)

3) Approximation des points candidats par une fonction *spline* cubique (en rouge) et rejet des points aberrants.

Figure 3.5 : Segmentation de l'image ultrasonore par recherche des maxima du gradient d'intensité

Dans la plupart des cas, cette seconde approche de segmentation automatique de la surface de la langue fournit des résultats satisfaisants. Néanmoins, tout comme pour la première technique basée sur les contours actifs, cette approche échoue presque systématiquement face à une image pour laquelle la surface de la langue apparaît sous la forme d'un contour très flou, discontinu, voir même complètement absent. Dans le cadre de l'imagerie ultrasonore de l'appareil vocal, il est ainsi apparu difficile de rendre un processus d'extraction des caractéristiques visuelles, basé exclusivement sur une approche par segmentation, à la fois robuste et automatique.

De plus, comme il a été décrit à la section 2.3.2, d'autres structures autres que la surface supérieure de la langue sont visibles dans l'image ultrasonore, certaines de façon permanente (structures graisseuses et musculaires de la langue, ombres acoustiques de l'os hyoïde et de la mâchoire), d'autres de façon intermittente (palais, velum, os hyoïde). La présence de ces structures et leurs positionnements dans l'image renseignent sur la géométrie de la cavité buccale au cours de la production. Ces structures fournissent donc des informations supplémentaires qui peuvent s'avérer utiles pour la conversion visuo-acoustique.

Aussi, afin de permettre un codage d'une image ultrasonore à la fois « robuste », c'est-à-dire capable de s'affranchir des altérations éventuelles du contour de la surface de la langue, et « global », car prenant en considération l'ensemble des structures visibles dans la région d'intérêt, un autre type d'approche est proposé. Ce dernier fait l'objet de la section suivante.

3.2.3. Extraction des caractéristiques visuelles – Approche globale

Un processus d'extraction des caractéristiques visuelles basé sur une approche globale, considère la région d'intérêt (*Region Of Interest* ou ROI en anglais) comme une source d'informations insécable, et fait l'hypothèse que l'intégralité de cette dernière est porteuse d'une information pertinente par rapport à la production de la parole. Dans le cadre des travaux sur la transcription de la parole audio-visuelle, de nombreuses techniques ont été proposées pour

décrire l'image des lèvres du locuteur, utilisée en complément du signal acoustique. Un bref état de l'art sur ces techniques sera présenté à la section 3.3.1 (consacrée au traitement des images vidéo). Deux techniques sont ici mises en œuvre dans le cadre de l'imagerie ultrasonore du conduit vocal. La première est basée sur la transformée en cosinus discrète, la seconde, sur l'analyse en composantes principales (approche dite des « *EigenTongues* »).

Sélection de la région d'intérêt et redimensionnement

Les deux approches mises en œuvre nécessitent la sélection, dans l'image ultrasonore, d'une région d'intérêt. Dans le cadre de la base B1 , cette dernière est un rectangle délimité verticalement par les ombres acoustiques latérales, et horizontalement par l'os hyoïde et le palais (voir Figure 2.16), ce dernier étant observé pendant une phase de déglutition (voir section 2.3.2). Dans le cadre de la base B2 , la totalité de l'image est utilisée (voir Figure 2.17). Afin de réduire la dimensionnalité de l'espace image, la région d'intérêt est redimensionnée (par interpolation cubique), pour les deux bases, à 32x32 pixels dans le cadre de l'approche par transformée en cosinus discrète, et à 64x64 pixels dans le cadre de l'approche par analyse en composantes principales.

Transformée en cosinus discrète

La transformation en cosinus discrète (TCD ou *Discrete Cosinus Transform*, DCT en anglais) est une technique très utilisée en compression d'image ; elle est notamment à la base du format JPEG. C'est une transformation linéaire inversible proche de la transformée de Fourier discrète. La TCD d'une matrice I de taille NxN est une matrice de même taille définie, au coefficient multiplicatif près, par :

$$TCD(u,v) = \sum_{i=1}^{N}\sum_{j=1}^{N} I(i,j)\cos\left[\frac{\pi}{N}(i-\frac{1}{2})(u-1)\right]\cos\left[\frac{\pi}{N}(j-\frac{1}{2})(v-1)\right] \qquad \text{(Équation 3.5)}$$

La TCD fournit une représentation du contenu fréquentiel de l'image. Les premiers coefficients (en haut à gauche de la matrice) correspondent aux basses fréquences spatiales, c'est à dire aux variations (spatiales) d'intensité lentes. Les derniers coefficients correspondent aux hautes fréquences spatiales, c'est-à-dire aux variations rapides. Un processus d'extraction des caractéristiques visuelles par TCD présuppose que l'information pertinente est celle qui est portée par les basses fréquences spatiales. Ces dernières codent en effet la forme générale des principales structures qui apparaissent dans l'image. Les hautes fréquences se spécialisent plutôt dans le codage des détails liés, certes aux contours, mais également au bruit. La Figure 3.6 résume le processus d'extraction des caractéristiques visuelles basé sur la TCD.

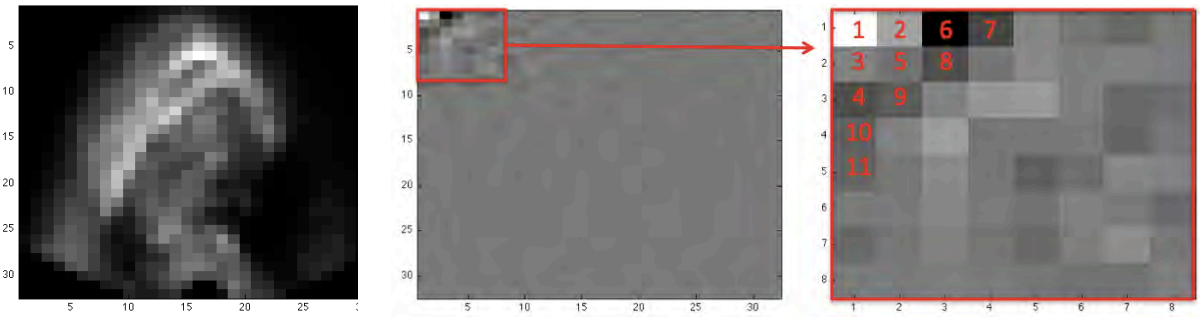


Image originale filtrée et redimensionnée (32x32 pixels) – Base B2

Transformée en cosinus discrète et extraction des coefficients correspondant aux basses fréquences spatiales (parcours « zigzag »)

Figure 3.6 : Extraction des caractéristiques visuelles – Approche globale par transformée en cosinus discrète

Dans la perspective de la conversion visuo-acoustique, il est nécessaire de se fixer un critère pour déterminer le nombre de coefficients utilisés comme « caractéristiques visuelles ». La procédure adoptée pour définir ce nombre est la suivante. Soit I l'image originale de taille NxN, $\{\alpha_k\}_{k=1..N^2}$ la suite ordonnée des coefficients de la TCD de I (parcours « zigzag ») et $\hat{I}_n$ l'image reconstruite à partir des n premiers coefficients. On définit E_n, l'erreur quadratique de reconstruction normalisée, par la relation :

$$E_n = \frac{\left\| I - \hat{I}_n \right\|^2}{\left\| I \right\|^2} = \frac{\sum_{k=n+1}^{N^2} \alpha_k^2}{\sum_{k=1}^{N^2} \alpha_k^2} \qquad \text{(Équation 3.6)}$$

Pour déterminer le nombre de coefficients à conserver, on se propose de déterminer n tel que :

$$0.1 < E_n < 0.2 \qquad \text{(Équation 3.7)}$$

Il s'agit, en d'autres termes, de conserver les coefficients qui, à eux seuls, représentent entre 80% et 90% de l'énergie totale de l'image originale. Ceci permet d'encoder une partie importante des informations basses fréquences tout en rejetant (en théorie) le bruit. Cette approche est illustrée par la Figure 3.7, où est représentée, pour chaque base de données B1 et B2 (voir section 2.4.2), l'erreur quadratique moyenne de reconstruction en fonction du nombre de coefficients de la TCD : cette moyenne est calculée sur un ensemble de 500 images.

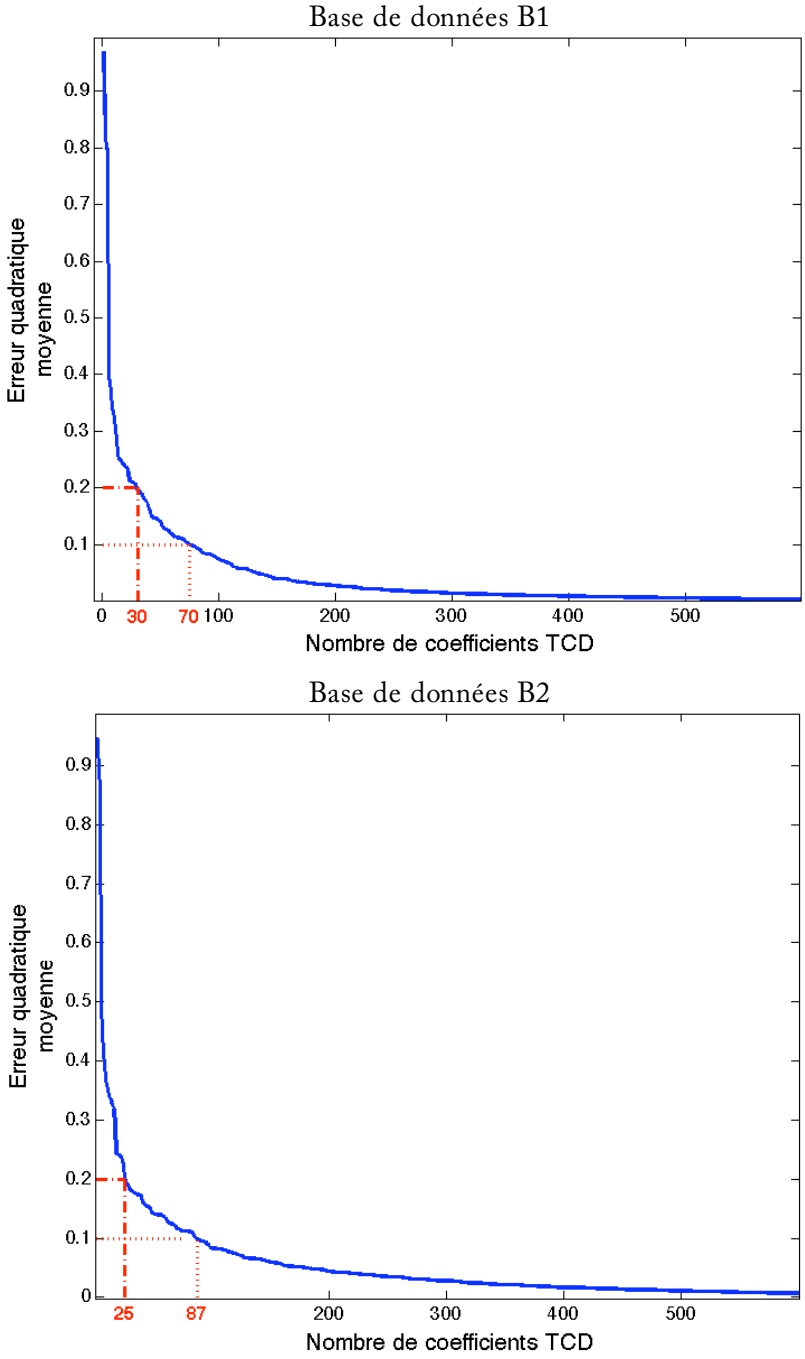


Figure 3.7 : Erreur quadratique moyenne de reconstruction de l'image ultrasonore en fonction du nombre de coefficients TCD utilisés

Il apparaît ainsi que 80% de l'énergie de l'image originale est portée par environ 3% des coefficients de la TCD seulement (1024 coefficients en totalité). Afin de limiter le nombre de descripteurs visuels, et d'uniformiser le traitement des deux bases de données, les 30 premiers coefficients de la TCD sont utilisés comme caractéristiques visuelles de l'image ultrasonore.

EigenTongues

La seconde approche mise en œuvre est une approche similaire à la technique dite des « *EigenFaces* », introduite par Turk et Pentland dans le cadre de la reconnaissance des visages (Turk et Pentland, 1991). Adaptée au contexte du traitement de la parole, et appliquée à des images du conduit vocal, l'approche proposée est nommée ici « *EigenTongues* » (Hueber *et al.*, 2007a). Le principe de cette approche est de construire un espace de représentation qui explique au mieux les variations d'intensité observées dans un ensemble d'images d'apprentissage, chaque nouvelle image étant ensuite codée par ses coordonnées dans cet espace. Ce dernier est obtenu comme suit. Soit E un ensemble d'apprentissage de M images de taille NxN, un parcours par ligne d'une image de taille NxN formant un vecteur de taille N^2 (une image de taille NxN est donc un point dans un espace de dimension N^2). L'ensemble d'apprentissage E peut donc être représenté par une matrice A de taille N^2xM (une image par colonne). L'espace de représentation recherché est obtenu par analyse en composantes principales (ACP) de E, c'est à dire après décomposition en valeurs propres de la matrice de covariance de A, notée C, tel que :

$$R^T C\ R = \Lambda \ \text{ avec } \ C = \frac{1}{M}(AA^T) \qquad \text{(Équation 3.8)}$$

où R et Λ sont respectivement la matrices des vecteurs et des valeurs propres. Les vecteurs propres ou « composantes principales » représentent les « directions de plus grande variance » observées dans l'ensemble E. L'espace des vecteurs propres est ordonné : le premier vecteur propre est orienté selon la direction de plus grande variance observée dans E, le second dans la direction de seconde plus grande variance, etc. Un vecteur propre étant de taille N^2, il peut être visualisé sous la forme d'une image de taille NxN que l'on nomme ici *EigenTongue* (voir Figure 3.8). L'interprétation de ces « images propres » n'est cependant pas intuitive. En effet, les zones de fortes variations d'intensité décrites par chacune des *EigenTongues* ne sont pas forcément localisées dans l'image. En d'autres termes, il n'est pas possible de dire qu'une *EigenTongue* se spécialise dans le codage exclusif d'une des structures visibles (apex, dos de la langue, os hyoïde, etc.).

Après détermination de la matrice des *EigenTongues* R, l'extraction des n caractéristiques visuelles $\{\alpha_k\}_{k=1..n}$ d'un nouveau « vecteur image » I de taille N^2 s'effectue simplement par détermination de ses coordonnées dans le sous-espace propre formé par les n premières *EigenTongues,* selon l'équation suivante :

$$\alpha_k = \sum_{i=1}^{N^2} I_i R_{ik} \qquad \text{(Équation 3.9)}$$

Le principe d'utilisation des premières composantes principales pour le codage d'une image repose sur l'hypothèse que les axes de grandes variances, c'est-à-dire les axes selon lesquels les données sont très dispersées, correspondent au signal « utile ». A l'inverse, les axes de faibles

variances sont considérés comme peu discriminants. Cette hypothèse reste ici raisonnable. Dans une image ultrasonore du conduit vocal, le signal « utile » (pour la conversion visuo-acoustique) est principalement la position de la langue et en effet, c'est bien sur les zones que cette dernière est susceptible d'occuper, que seront observées les plus fortes variations d'intensité. Ainsi, ce sont bien les « axes de grandes variances » qui portent l'information relative à la position de la langue dans l'image.

Par ailleurs, l'ensemble d'apprentissage *E* est, dans l'implémentation proposée, construit de façon « phonétiquement équilibrée », afin de couvrir au mieux l'espace articulatoire. Ceci nécessite une étape préalable de segmentation du flux vidéo au niveau phonétique, qui sera décrite au chapitre 5. Le processus d'extraction des caractéristiques visuelles basé sur l'approche *EigenTongues* est illustré par la Figure 3.8 (sur des données extraites de la base B1).

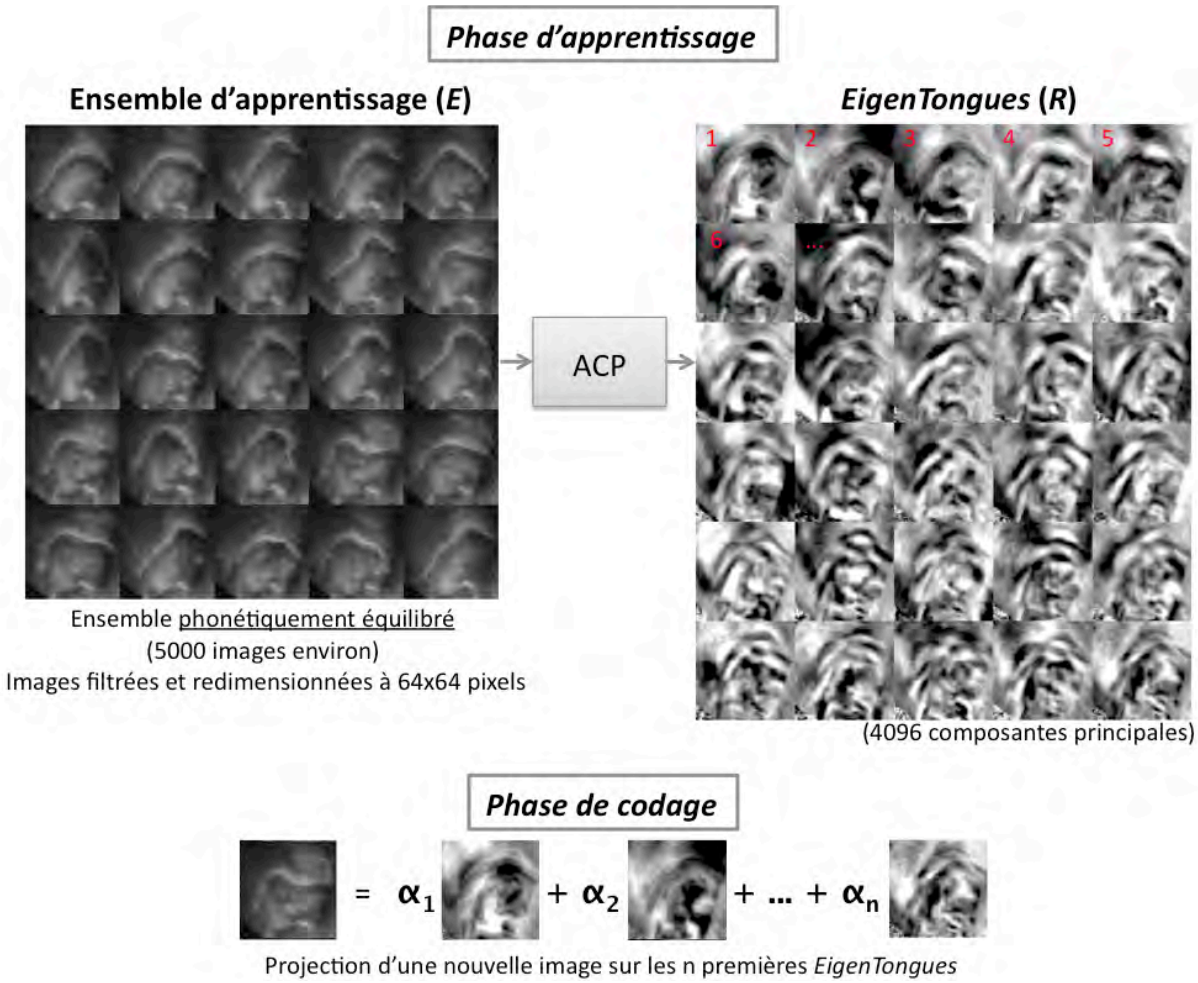


Figure 3.8 : Extraction des caractéristiques visuelles - Approche globale par décomposition d'une image ultrasonore dans l'espace des *EigenTongues*.

Pour déterminer le nombre de composantes principales à conserver pour l'extraction des caractéristiques visuelles, une procédure identique à celle utilisée dans le cadre de l'approche par TCD est mise en œuvre. L'erreur quadratique de reconstruction moyenne est calculée, sur le même ensemble de développement que celui utilisé pour l'approche par TCD (500 images), à

l'aide de l'équation 3.6 (les $\{\alpha_k\}_{k=1..N^2}$ étant cette fois, les coordonnées de l'image dans l'espace des *EigenTongues*).

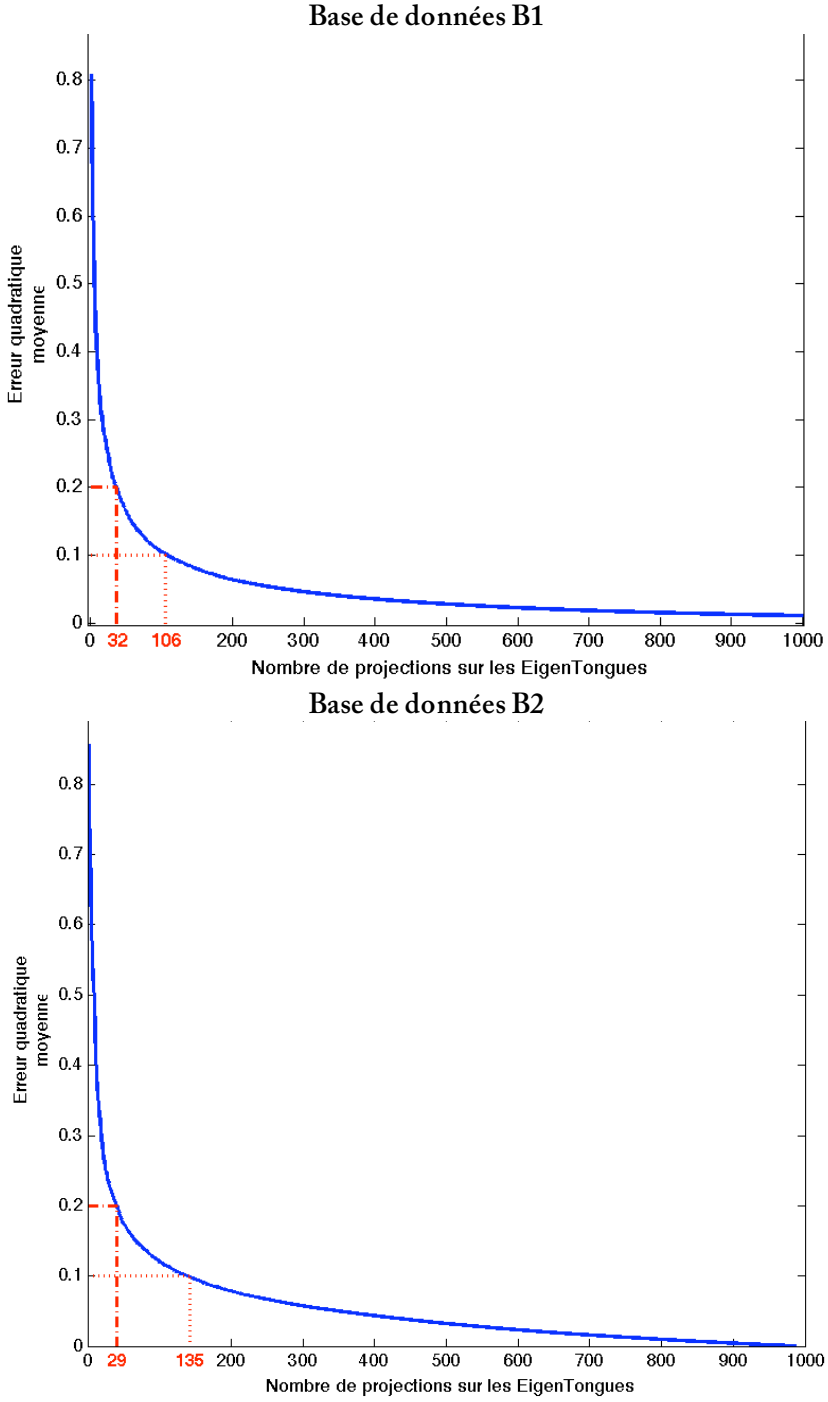


Figure 3.9 : Erreur quadratique moyenne de reconstruction de l'image ultrasonore en fonction du nombre de *EigenTongues* utilisées

En observant les courbes représentées à la Figure 3.9, il apparaît que 0.8 % des coefficients suffisent à expliquer 80 % de l'énergie de l'image originale (4096 coefficients en totalité). L'analyse en composante principale est en effet la technique de compression dite « optimale », en ce sens qu'elle fournit la base de décomposition qui permet l'approximation linéaire d'un

signal avec une erreur quadratique moyenne minimale (Mallat, 2001). Notons cependant que, contrairement à la TCD, cette performance est atteinte au prix d'une phase d'apprentissage préalable qui rend cette approche dépendante des données.

Afin d'uniformiser ici encore le traitement des deux bases de données, les caractéristiques visuelles d'une image ultrasonore sont définies, dans cette approche, comme étant ses 30 premières coordonnées dans l'espace des *EigenTongues*.

Normalisation des caractéristiques visuelles

Dans les travaux sur la transcription audio-visuelle de la parole, une normalisation des caractéristiques visuelles est souvent effectuée afin, notamment, de compenser les variations d'illuminations observables entre les différentes séquences d'images des lèvres (Vanegas *et al.*, 1998), (Potamianos et Neti, 2001). Bien qu'il puisse paraitre délicat de parler « d'illumination » dans le cas d'une image ultrasonore, l'intensité moyenne des pixels de l'image peut néanmoins varier d'une séquence à l'autre. Ceci peut notamment s'expliquer par la présence plus ou moins importante de gel de contact et par le caractère humide ou sec des tissus de la cavité buccale (une bouche « sèche » est peu échogène, (Stone, 2005)). Aussi, afin de vérifier si une normalisation des caractéristiques visuelles est justifiée dans notre cas, nous représentons à la Figure 3.10, dans le cadre de l'approche *EigenTongues*, l'évolution des premières caractéristiques visuelles pour trois occurrences des mots « Alpha » et « Juliet » (enregistrés à l'aide du système SA2, voir section 2.4.2). Pour un même mot (donc pour un même contenu linguistique), les évolutions absolues des caractéristiques visuelles sont, aux déformations temporelles près (dues aux vitesses de prononciations différentes), relativement similaires d'une occurrence à l'autre. Néanmoins, en comparant les évolutions relatives, par exemple dans le cas de « Alpha – *EigenTongues* 2 » ou dans celui de « Juliet – *EigenTongues* 1 », on observe que la valeur moyenne de la caractéristique visuelle varie d'une occurrence à l'autre. Bien qu'en toute rigueur cette analyse devrait être effectuée sur plus d'exemples, la normalisation des caractéristiques visuelles apparaît ici justifiée. Cette dernière s'effectue selon la procédure suivante.

On note s, le segment de parole correspondant au stimulus textuel présenté au locuteur lors de l'enregistrement des bases de données (typiquement, un mot ou une phrase), M le nombre d'images ultrasonores contenues dans s, $\mu_{s,k}$ et $\sigma_{s,k}$ respectivement la moyenne et la variance de la $k^{ème}$ caractéristique visuelle notée $y_k = \left[y_k(1),...,y_k(M) \right]$, calculée sur s. La $k^{ème}$ caractéristique visuelle normalisée, notée $\hat{y}_k$, est alors définie par :

$$\hat{y}_k = \frac{y_k - \mu_{s,k}}{\sigma_{s,k}} \qquad \text{(Équation 3.10)}$$

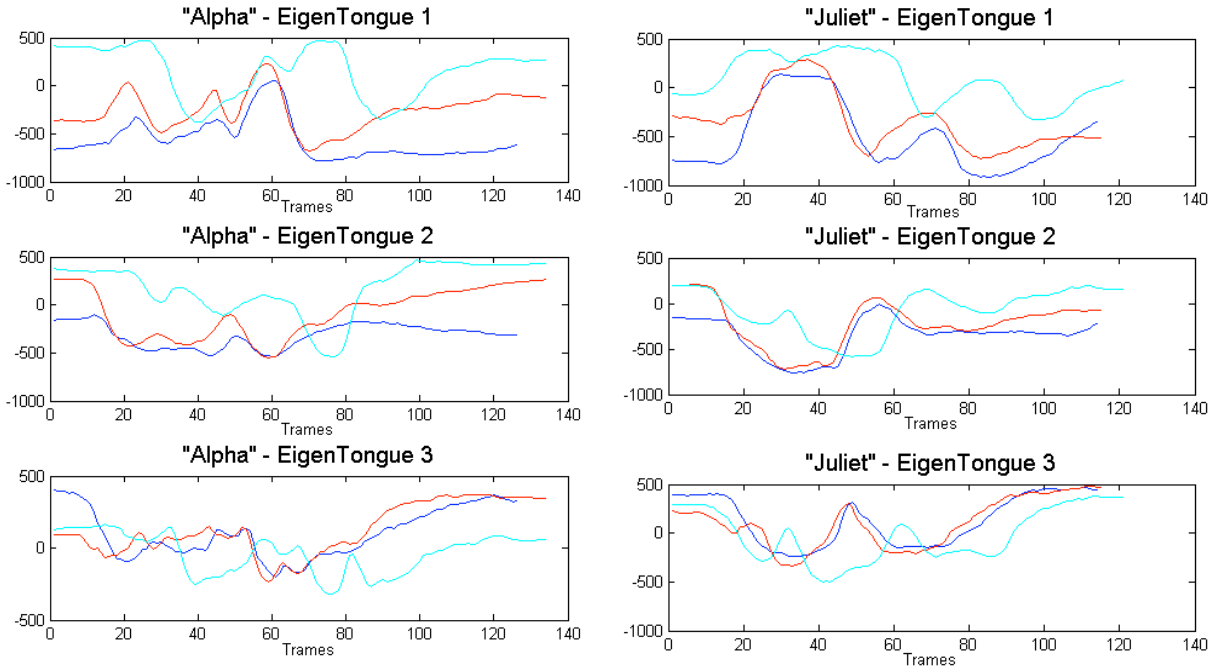


Figure 3.10 : Evolution des trois premières caractéristiques visuelles, dans le cas de l'approche *EigenTongues*, pour trois occurrences (bleu, rouge, cyan) des mots « Alpha » et « Juliet », base B2.

Cette équation fait apparaître, en plus de la moyenne, une normalisation par la variance. Il est difficile de motiver cette dernière par la seule analyse des courbes de la Figure 3.10. Bien que non indispensable, cette (classique) normalisation par la variance s'est néanmoins avérée profitable lors des diverses expériences de conversion visuo-acoustique décrites ultérieurement.

Ajout de caractéristiques visuelles dynamiques

L'appareil vocal est un système dynamique dont le comportement à un instant donné dépend de ses états antérieurs. En théorie phonétique, ce principe est notamment un des fondements de la phonologie articulatoire (Browman et Goldstein, 1990), qui décrit la production de la parole comme une réalisation planifiée de « gestes articulatoires », primitives caractérisant non seulement le lieu d'articulation mais également la dynamique du mouvement des articulateurs[39]. En transcription de la parole audio-visuelle, l'aspect dynamique de la production est généralement intégré dès l'étape d'extraction des caractéristiques visuelles, en complétant ces dernières par leurs dérivées premières et secondes[40].

[39] Certains des résultats issus de la phonologie articulatoire seront utilisés pour la conversion visuo-acoustique, dans le cadre de l'approche dite « indirecte » introduite au chapitre 5.

[40] Une stratégie identique est traditionnellement adoptée pour la transcription automatique de la parole acoustique (Rabiner et Juang, 1993). Dans (Potamianos et Neti, 2001), une stratégie alternative est proposée. Cette dernière consiste à compléter les caractéristiques statiques d'une image par celles des images qui lui sont

C'est cette approche qui est ici utilisée pour la description des données ultrasonores ; les dérivées premières et secondes de la séquence de caractéristiques visuelles *y*, notées respectivement Δ_y et Δ_y^2, sont approximées par la méthode des différences finies, à l'aide des relations suivantes :

$$\Delta_y(n) = \frac{y(n+h) - y(n-h)}{2h} \text{ et } \Delta_y^2(n) = \frac{y(n+h) - 2y(n) + y(n-h)}{h^2} \qquad \text{(Équation 3.11)}$$

avec *n* l'indice de la trame et *h*, la taille de la fenêtre de calcul, fixée ici à 1. Après ajout de ces composantes dynamiques, chaque image ultrasonore est à présent décrite par 90 caractéristiques visuelles.

3.3. Traitement des images vidéo

3.3.1. Etat de l'art

De l'ensemble des techniques utilisées en transcription de la parole audio-visuelle pour décrire l'image des lèvres, se dégagent trois grandes approches (Potamianos *et al.*, 2004):

- les approches par segmentation, basées sur l'extraction et le paramétrage du contour (intérieur et/ou extérieur) des lèvres (André-Obrecht *et al.*, 1997), (Rogozan et Deléglise, 1998).
- les approches globales, parmi lesquelles figurent notamment l'approche par analyse en composantes principales, dite des « *EigenLips* » (Bregler et Konig, 1994), la transformée en cosinus discrète (Duchnowski *et al.*, 1994) et la transformée en ondelettes discrètes (Potamianos *et al.*, 1998).
- les approches que nous qualifierons ici de « mixtes », qui combinent les informations sur la forme des lèvres obtenues par segmentation du contour, et celles fournies par une analyse globale de la région d'intérêt. Cette combinaison s'effectue de façon plus au moins directe, par concaténation, comme dans (Dupont et Luettin, 2000), ou via l'apprentissage d'un « modèle actif d'apparence » (Cootes *et al.*, 1998), comme dans (Matthews *et al.*, 2001).

Bien que les travaux les plus récents semblent privilégier les approches mixtes, basées notamment sur l'utilisation de « modèles actifs d'apparence », il n'y a pas, à notre connaissance, d'étude démontrant clairement la supériorité d'une technique en particulier.

3.3.2. Approches mises en œuvre

Dans (Hueber, 2006), une approche par segmentation a été adoptée pour décrire la vue de profil des lèvres du locuteur (sur les données acquises avec le système SA1). Dans le cadre de cette étude, la caractérisation de l'image des lèvres, en vue de face comme en vue de profil (cas de la base B1, voir section 2.4.2) est exclusivement effectuée à l'aide d'approches globales. Ce choix se justifie, ici encore, par la capacité de ces dernières à intégrer plus d'informations que le

adjacentes. N'ayant pas donné de résultats probants dans le cadre de notre étude, cette approche ne sera pas développée ici.

simple mouvement des lèvres, comme notamment la présence ou l'absence des dents dans l'image (pour la vue de face), voire celle de la langue. Ce dernier point est particulièrement intéressant dans le cadre de notre étude, notamment pour l'observation des consonnes interdentales (par exemple pour [th], comme dans *author* en anglais). Pour ces dernières, l'apex vient s'appuyer sur les incisives. Caché par l'ombre acoustique de la mâchoire, il n'est pas visible dans l'image ultrasonore mais peut apparaître distinctement dans l'image vidéo, comme l'illustre la Figure 3.11 (notamment sur les données de la base B2).

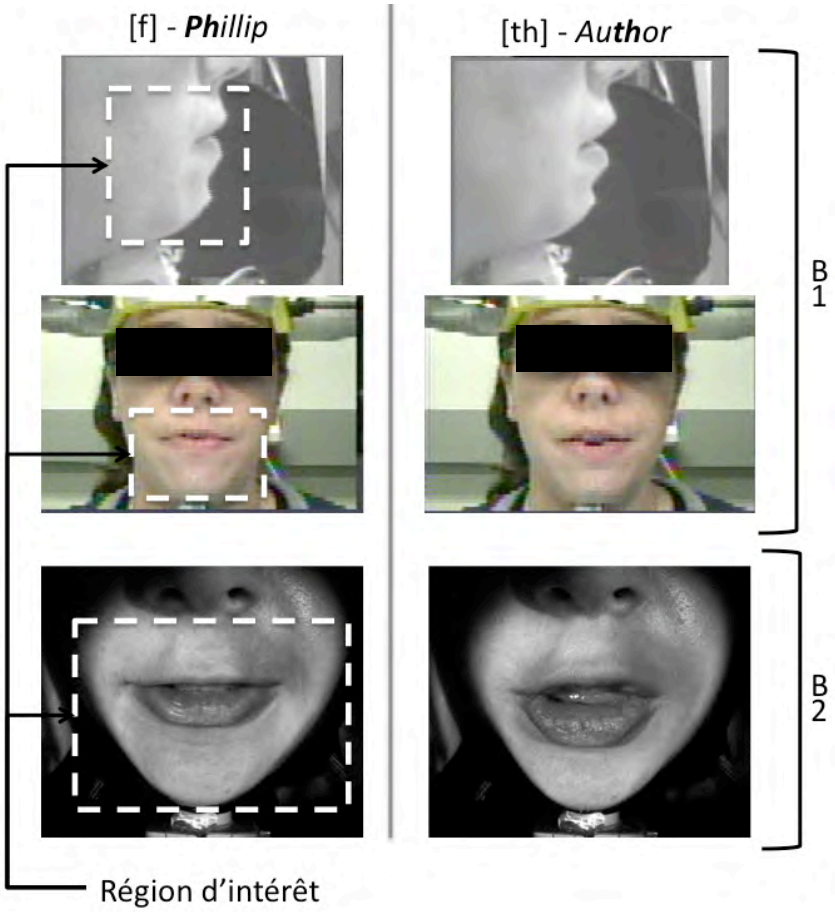


Figure 3.11 : Traitement des images vidéo - Consonnes labiodentales (cas du [f] à gauche) et interdentales ([th] à droite) - Régions d'intérêt pour les bases B1 et B2.

Pour traiter les images vidéo des bases B1 et B2 (vues de face comme de profil), un processus d'extraction des caractéristiques strictement identique à celui utilisé pour le traitement des images ultrasonores (voir section 3.2.3) est mis en œuvre. Après sélection de la région d'intérêt, chaque trame vidéo, est convertie (si nécessaire) en niveau de gris, puis est redimensionnée en une vignette de 64x64 pixels. Les caractéristiques visuelles associées à cette trame sont ensuite calculées soit par transformée en cosinus discrète, soit par analyse en composantes principales (qui construit, cette fois, l'espace des « *EigenLips* »). Une analyse similaire à celle menée dans le cas de l'image ultrasonore (voir page 71) est effectuée pour déterminer le nombre de coefficients à conserver. Pour les deux approches, il apparaît alors qu'une erreur de

reconstruction normalisée (équation 3.6) de 0.2 est atteinte à l'aide des 25 premiers coefficients. Afin d'uniformiser les différents processus d'extraction des caractéristiques visuelles, nous décidons néanmoins, pour les deux approches, d'utiliser les 30 premiers coefficients comme caractéristiques visuelles d'une image vidéo. A l'aide des mêmes procédures que celles décrites à la section 3.2.3, le vecteur résultant est ensuite normalisé, puis complété par des caractéristiques « dynamiques » (obtenues après calcul des dérivées premières et secondes). Chaque trame vidéo, représentant soit une vue de face des lèvres (bases B1 et B2), soit une vue de profil (base B1 uniquement) se voit alors représentée par un vecteur de 90 caractéristiques visuelles.

3.4. Techniques d'analyse-synthèse du signal acoustique

Dans le cadre des approches proposées pour la conversion visuo-acoustique, deux techniques d'analyse-synthèse du signal de parole sont mises en œuvre. La première, utilisée dans le cadre de l'approche de conversion dite « directe », est basée sur l'utilisation de coefficients mel-cepstraux, comme descripteurs acoustiques pour la phase de modélisation, puis comme paramètres du filtre de synthèse pour la phase de conversion. La seconde, mise en œuvre dans le cadre de l'approche dite « indirecte », est la technique d'analyse-synthèse « Harmonique plus Bruit ». Ces deux approches sont brièvement décrites dans les sections suivantes.

3.4.1. Analyse cepstrale

La production de la parole peut être décrite comme la mise en forme du flux d'air laryngé (pulsé ou continu) par les articulateurs des cavités orales (voir section 1.1). Ainsi, le signal de parole peut être décrit comme le résultat de l'excitation, par un signal source, d'un filtre résonnant dont les caractéristiques varient au cours du temps. Cependant, les images ultrasonores et vidéo ne renseignent que sur l'état du conduit vocal (le filtre), et non sur celui de l'appareil excitateur (la source). Aussi, il semble préférable, pour la modélisation visuo-acoustique, d'opter pour une technique de description du signal acoustique qui isole les contributions du filtre de celles de la source. Cette déconvolution « source-filtre » peut être effectuée par analyse cepstrale.

Le cepstre réel est défini comme la transformée de Fourier inverse, du logarithme du spectre (déconvolution homomorphique). Soit $e(n)$ le signal d'excitation (discret) caractérisant la source glottique, $h(n)$ le filtre résonnant caractérisant le conduit vocal ; le signal de parole résultant $x(n)$ s'écrit :

$$x(n) = e(n) * h(n) \quad \text{avec } (*) \text{ le produit de convolution} \qquad \text{(Équation 3.12)}$$

En notant F, la transformée de Fourier discrète, le cepstre C du signal $x(n)$ est défini par :

$$C\{x(n)\} = F^{-1}\left\{\log\left|F\{x(n)\}\right|\right\} \qquad \text{(Équation 3.13)}$$

En combinant ces deux dernières équations, on obtient :

$$\begin{aligned} C\{x(n)\} &= F^{-1}\left\{\log\left|F\{e(n)\}F\{h(n)\}\right|\right\} \\ &= F^{-1}\left\{\log\left|F\{e(n)\}\right| + \log\left|F\{h(n)\}\right|\right\} \\ &= F^{-1}\left\{\log\left|F\{e(n)\}\right|\right\} + F^{-1}\left\{\log\left|F\{h(n)\}\right|\right\} \\ &= \hat{e}(n) \; + \; \hat{h}(n) \end{aligned} \qquad \text{(Équation 3.14)}$$

Les contributions de la source, $\hat{e}(n)$, font l'objet de variations rapides dans le spectre ; elles se situent donc dans les hautes « quéfrences » (la partie haute du cepstre). En revanche, les contributions du filtre, $\hat{h}(n)$, c'est à dire l'enveloppe spectrale, correspondent aux variations lentes dans le spectre ; elles se situent donc dans les basses quéfrences (partie basse du cepstre). Ainsi, par « liftrage », c'est-à-dire par filtrage dans l'espace des quéfrences, il est possible d'isoler les deux contributions.

En reconnaissance (acoustique) de la parole, l'analyse cepstrale est utilisée pour l'extraction des coefficients dits « mel-cepstraux » ou MFCC (pour *Mel Frequency Cepstral Coefficient*). Il s'agit alors d'évaluer le contenu spectral du signal[41] sur une échelle fréquentielle non-linéaire, dite échelle de Mel, qui rend compte des caractéristiques propres à la perception des sons par l'oreille humaine[42]. Ce traitement est généralement effectué dans le domaine fréquentiel, en multipliant le module de la TFD par le gabarit d'un banc de filtres triangulaires répartis sur l'échelle de Mel, puis en intégrant les coefficients résultants sur chacun des intervalles définis par ce banc de filtre. Un schéma récapitulatif de la procédure couramment utilisée pour le calcul des coefficients mel-cepstraux est proposé à la Figure 3.12.

Ce schéma de calcul, très utilisé dans le cadre de la reconnaissance de la parole, n'est pas applicable dans un contexte de synthèse. En effet, il est impossible de retrouver l'enveloppe spectrale originale à partir des coefficients mel-cepstraux calculés selon ce schéma pour deux raisons. D'une part le filtrage par banc de filtres effectue, sur chacun de ces filtres, une moyenne du spectre : il s'agit d'une opération destructrice, donc irréversible. D'autre part ce schéma de calcul est basé sur l'utilisation du cepstre réel. Ce dernier ne considère que le spectre d'amplitude sans tenir compte des informations fournies par le spectre de phase.

[41] De façon classique, le contenu spectral est évalué sur une fenêtre d'analyse, c'est-à-dire sur un segment sur lequel le signal est supposé stationnaire, pondéré, par exemple, par une fenêtre du type Hamming.

[42] Etablie à l'aide de tests perceptifs, l'échelle de Mel montre que la capacité du système auditif humain à évaluer la différence entre deux fréquences, diminue quand la fréquence augmente. En d'autres termes, l'oreille humaine est plus sensible aux variations dans les basses que dans les hautes fréquences. C'est pourquoi l'échelle de Mel est construite de façon linéaire dans les basses fréquences et logarithmique dans les hautes fréquences.

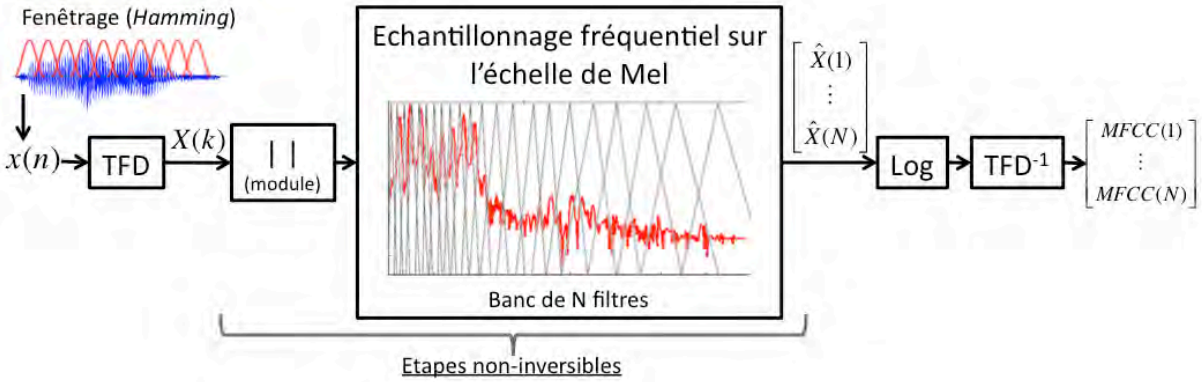


Figure 3.12 : Extraction des coefficients mel-cepstraux ou MFCC. Présentant de meilleures propriétés de regroupement de l'information basse fréquence, la transformée en cosinus discrète (TCD) est souvent utilisée à la place de la transformée de Fourier inverse (TFD^{-1}). Le spectre d'amplitude est parfois remplacé par le spectre de puissance. Typiquement, les *N/2* premiers coefficients MFCC sont utilisés comme descripteurs acoustiques.

Pour permettre une transformation réversible utilisable dans un contexte de synthèse, un autre schéma de calcul, basé sur le cepstre complexe, a été proposé dans (Imai, 1983). Dans cette approche, le spectre complexe d'un signal *x(n)*, noté $X(z)$, est modélisé par *M* coefficients mel-cepstraux[43] $c = [c_\alpha(0),...,c_\alpha(M)]$, tel que :

$$X(z) = \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m}$$

avec (Équation 3.15)

$$\tilde{z}^{-1} = \Psi_\alpha(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad |\alpha| < 1$$

Le filtre définie par la fonction de transfert (complexe) $\Psi_\alpha(z)$ est un filtre déphaseur, tel que :

$$\arg(\Psi_\alpha(e^{j\omega})) = \beta_\alpha(\omega) = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \qquad \text{(Équation 3.16)}$$

Avec un choix approprié du paramètre α, la fonction β_α réalise une bonne approximation de l'échelle de Mel[44]. A la différence de l'approche par banc de filtres, l'échantillonnage du spectre sur cette échelle ne se fait pas ici de façon discrète, mais continue. Une méthode de calcul des coefficients mel-cepstraux, basée sur cette approche, a été proposée par Imai dans (Imai et Furuichi, 1988)[45]. Cette méthode, qui ne sera pas détaillée ici, ramène le problème de l'estimation des coefficients à un problème d'optimisation.

[43] De façon classique, le nombre *M* de coefficients mel-cepstraux est ici fixé à 24.

[44] Typiquement, pour signal échantillonné à 16 kHz, α = 0.42.

[45] Cette méthode est celle utilisée pour le calcul des coefficients mel-cepstraux dans la boite à outils SPTK (http://sp-tk.sourceforge.net), utilisée dans le cadre de cette étude.

Enfin, à partir de l'équation 3.15, il est possible de définir la structure d'un filtre numérique, nommé filtre MLSA (pour *Mel Log Spectral Approximation*), utilisable, dans une approche « source-filtre », pour la synthèse du signal à partir de coefficients mel-cepstraux. Introduite et décrite dans (Imai *et al.*, 1983), cette technique est notamment utilisée dans le cadre de la « synthèse par MMC » (Tokuda *et al.*, 2000).

3.4.2. Modélisation « Harmonique plus Bruit »

La modélisation « Harmonique plus Bruit », introduite par Stylianou (Stylianou, 1990) et plus connue sous l'acronyme HNM (pour *Harmonic plus Noise Model*), suppose que le signal de parole *x(n)* peut se décomposer en une partie dite harmonique, *h(n)*, modélisant les structures quasi-périodiques du signal, et en une partie dite « bruitée », *b(n)*, qui décrit les composantes apériodiques comme les bruits de friction et les variations inter-périodes de l'excitation glottale. Le signal *x(n)* peut ainsi s'écrire :

$$x(n) = h(n) + b(n) \qquad \text{(Équation 3.17)}$$

Le codeur HNM agit différemment selon la caractéristique voisée ou non-voisée du signal ; la première étape du codage est donc une analyse de la fréquence fondamentale[46]. Dans le cas d'un segment non-voisé, la composante harmonique est considérée comme nulle et le signal est décrit à l'aide d'un modèle autorégressif obtenu par prédiction linéaire (analyse LPC). Une estimation à intervalles réguliers de la variance du signal permet de définir un gain (variable) pour le filtre « tout pôle ». Dans le cas d'un segment voisé, le spectre du signal est décomposé en deux sous-bandes de fréquence, délimitées par la « fréquence maximale de voisement ». Cette dernière, variable dans le temps, est définie comme la dernière harmonique « visible » de la fréquence fondamentale f_0 (Figure 3.13). Elle s'obtient à l'aide d'une analyse de la structure fine du spectre basée sur la détection d'irrégularités dans la répartition des harmoniques (les partiels).

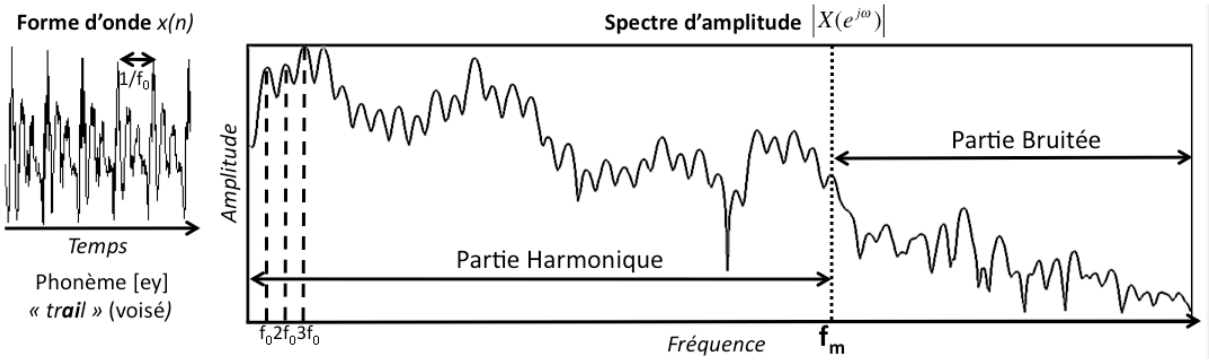


Figure 3.13 : Codage HNM – Décomposition du spectre en bandes « harmonique » et « bruit » délimitées par la fréquence maximale de voisement f_m (f_0 est la fréquence fondamentale)

[46] De multiples méthodes pour l'estimation de la fréquence fondamentale ont été proposées dans la littérature. Pour une synthèse des principales approches, le lecteur pourra par exemple consulter (Doval, 1994).

La composante harmonique *h(n)* d'un segment voisé du signal *x(n)*, est alors décrite à l'aide d'un modèle sinusoïdal tel que[47] :

$$h(n) = \sum_{k=-L(n_a^i)}^{L(n_a^i)} A_k(n_a^i) e^{j2\pi k f_0(n_a^i)(n-n_a^i)}$$
$$\text{avec } n_a^{i+1} = n_a^i + f_0^{-1}(n_a^i) \qquad \text{(Équation 3.18)}$$
$$L(n_a^i) = E\left[f_m(n_a^i) / f_0(n_a^i) \right]$$
$$A_k \in \mathbb{C} \text{ et } A_{-k} = A_k^*$$

Les n_a^i sont les instants d'analyse, définis de façon « pitch-synchrone », c'est-à-dire espacés d'une période fondamentale. f_0 et f_m sont respectivement la fréquence fondamentale et la fréquence maximale de voisement (normalisées), et L, le nombre d'harmoniques résultant (E est la fonction partie entière). Les amplitudes (complexes) A_k de ces harmoniques sont les paramètres du modèle à estimer à partir du signal. En considérant ces derniers constants sur une fenêtre d'analyse de taille $2N$ échantillons centrée sur l'instant d'analyse n_a^i (avec $N \approx 1 / f_0(n_a^i)$), une approche par minimisation au sens des moindres carrés (pondérées) fournit l'expression analytique suivante :

$$A_k = \frac{\sum_{n=n_a^i-N}^{n_a^i+N} \omega^2(n) x(n) e^{-j2\pi k f_0 n}}{\sum_{n=n_a^i-N}^{n_a^i+N} \omega^2(n)} \qquad \text{(Équation 3.19)}$$

avec $\omega(n)$ une fenêtre de pondération qui vise à donner plus d'importance aux échantillons proches de l'instant d'analyse (fenêtre de *Hamming* par exemple).

Une fois la partie harmonique *h(n)* estimée, la partie bruitée *b(n)* peut se définir, dans le domaine temporel, comme le signal résiduel suivant :

$$b(n) = x(n) - h(n) \qquad \text{(Équation 3.20)}$$

Comme dans le cas d'une trame non-voisée, ce signal est alors décrit par un modèle autorégressif obtenu par prédiction linéaire[48]. Une estimation de la variance du signal aux instants d'analyse n_a^i fournit le gain du filtre tout-pôle.

L'implémentation de la technique d'analyse-synthèse « Harmonique plus Bruit » utilisée dans le cadre de cette étude, présente certaines spécificités par rapport à celle décrite dans

[47] La formulation utilisée ici est dérivée du modèle « HNM1 » décrit dans (Stylianou, 1990).

[48] Dans (Stylianou, 1990), l'extraction de la partie bruitée s'effectue par modélisation autorégressive du signal original et non du signal résiduel. La partie bruitée n'est donc pas obtenue en phase d'analyse mais en phase de synthèse, par filtrage passe-haut du signal synthétisé à partir du modèle AR du signal « pleine bande ».

(Stylianou, 1990). Dans l'équation 3.18, $L(n_a^i)$ (nombre de paramètres utilisés pour décrire la partie harmonique) varie d'un instant d'analyse à l'autre. Afin d'obtenir un système de codage présentant un nombre constant de paramètres, une modélisation autorégressive de la partie harmonique estimée est également effectuée. Un schéma général de l'approche mise en œuvre est proposé à la Figure 3.14.

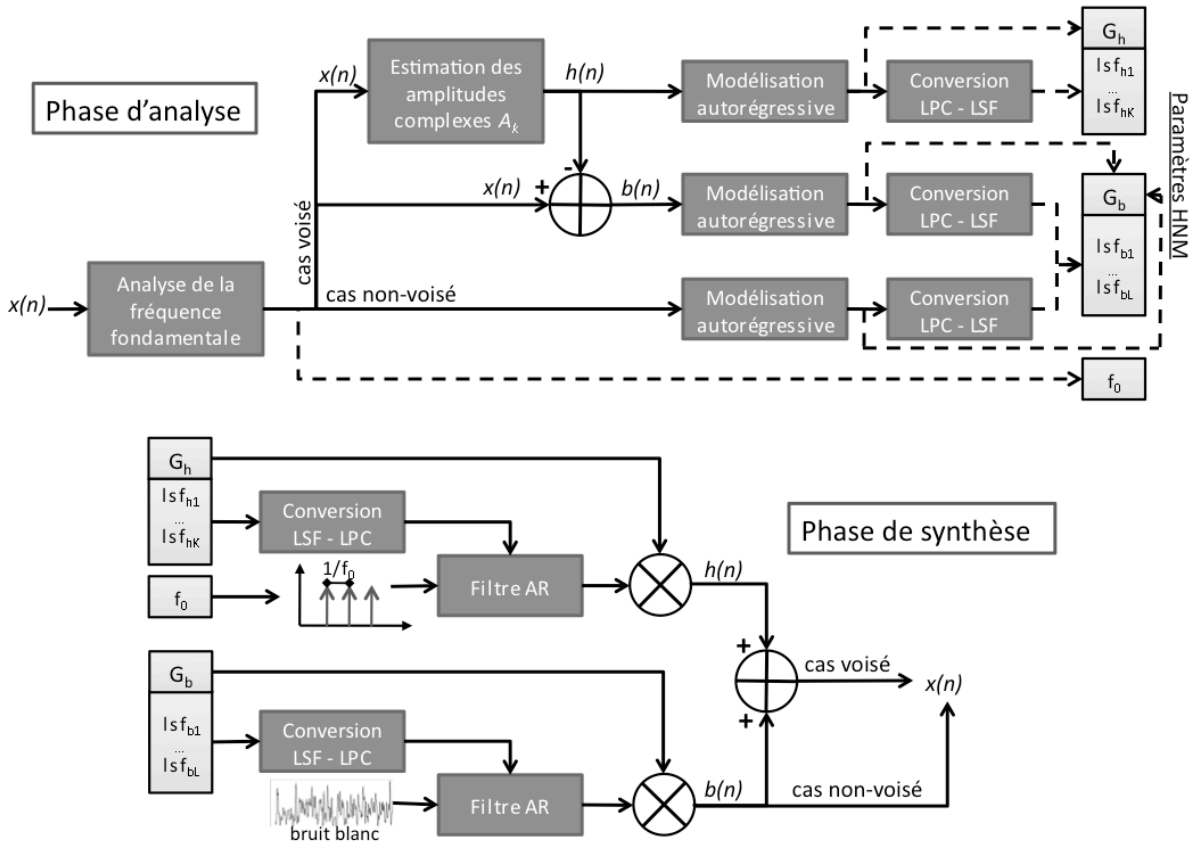


Figure 3.14 : Schéma général de fonctionnement du système d'analyse-synthèse « Harmonique plus Bruit » mis en œuvre dans le cadre de cette étude. G_h et G_b sont respectivement les gains des modèles AR (autorégressifs) des parties harmonique et bruit. K et L sont l'ordre de ces modèles et sont fixés respectivement à 12 et 16 pour un signal échantillonné à 16 kHz.

Ce schéma fait également apparaître une représentation des coefficients des modèles AR des parties harmonique et bruit, par des coefficients LSF (*Line Spectrum Frequencies*) (Itakura, 1975). Soit *F(z)* un modèle AR d'ordre p défini tel que :

$$F(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad \text{(Équation 3.21)}$$

Les coefficients LSF sont définis comme la suite des racines (entrelacées) des polynômes P et Q, définis par :

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \end{aligned} \qquad \text{(Équation 3.22)}$$

La représentation d'un modèle AR à l'aide de coefficients LSF est connue comme étant plus robuste que la simple utilisation des coefficients du filtre tout-pôle. En effet, une faible variation d'un seul de ces coefficients peut suffire à rendre le modèle AR instable. En revanche, une propriété simple sur les coefficients LSF garantit la stabilité du modèle. En effet, un modèle AR décrit par des coefficients LSF est stable si et seulement si la suite formée par ces coefficients est strictement croissante. Cette propriété rend la représentation LSF particulièrement bien adaptée à une utilisation dans un contexte d'apprentissage artificiel. En effet, tant que l'ordre des coefficients LSF n'est pas affecté, une erreur de prédiction effectuée sur un ou plusieurs de ces coefficients n'empêche pas la synthèse du signal. Aussi, afin de garantir cette relation d'ordre lors de l'inférence, nous remplaçons le vecteur des coefficients LSF par un vecteur construit en calculant les différences (positives) entre deux coefficients consécutifs. Ainsi, les vecteurs $[lsf_{h1},...,lsf_{hK}]$et $[lsf_{b1},...,lsf_{bL}]$ (voir Figure 3.14) sont respectivement remplacées, pour la conversion visuo-acoustique, par les vecteurs $[lsf_{h1}, lsf_{h2}-lsf_{h1},...,lsf_{hK}-lsf_{h(K-1)}]$ et $[lsf_{b1}, lsf_{b2}-lsf_{b1},...,lsf_{bL}-lsf_{b(L-1)}]$.

Enfin, à l'aide de son schéma d'analyse-synthèse « *pitch-synchrone* », la technique HNM permet d'effectuer, de façon assez simple et directe, des transformations du signal du type « correction de la hauteur » (en anglais *pitch-shifting*) et « déformation temporelle » (en anglais *time-stretching*)[49]. Cette propriété a notamment motivé l'utilisation de HNM dans le cadre de l'approche indirecte de la conversion visuo-acoustique, qui sera introduite au dernier chapitre.

[49] Ces transformations peuvent s'effectuer par exemple par simple interpolation des paramètres HNM.

Chapitre 4. Conversion visuo-acoustique, approche directe

4.1. Avant-propos

Le geste articulatoire, capturé par imagerie ultrasonore et vidéo, et sa réalisation acoustique associée, accessible *via* le signal audio, peuvent être considérés comme deux représentations d'une même « information linguistique ». Une première approche (peut être la plus « naturelle ») pour aborder le problème de la conversion visuo-acoustique consiste à le considérer comme un problème d'association de deux espaces de données. Il s'agit alors de déterminer une ou plusieurs « fonctions de transformation » qui établissent un lien entre caractéristiques visuelles et caractéristiques acoustiques. La conversion visuo-acoustique est alors effectuée au niveau du « signal » et n'utilise aucunes des caractéristiques « haut niveau » du signal de parole, comme celles fournies par exemple, par une description de ce dernier à des niveaux linguistiques supérieurs. Par opposition à la méthode décrite au chapitre suivant, cette approche est donc qualifiée de « directe ».

Une approche « source-filtre », basée sur l'utilisation du filtre MLSA (voir section 3.4.1), est ici adoptée pour la synthèse du signal. Deux ensembles de paramètres doivent alors être inférés à partir des données visuelles : les coefficients du filtre de synthèse d'une part, les caractéristiques de la fonction par laquelle ce dernier doit être excité, d'autre part. Les coefficients du filtre MLSA sont dérivés de coefficients mel-cepstraux. L'inférence de ces derniers à partir des caractéristiques visuelles est un problème de régression. Deux approches sont proposées et comparées. La première met en œuvre un réseau de neurones, la seconde s'appuie sur la modélisation de l'espace conjoint « visuo-acoustique » par un mélange de gaussiennes. Dans le synthétiseur « source-filtre » mis en œuvre, la fonction d'excitation du filtre est entièrement déterminée à partir de la caractéristique « voisée/non-voisée » de la trame à synthétiser, et dans le cas d'une trame voisée, par sa fréquence fondamentale. L'inférence de la caractéristique « voisée/non-voisée » est un problème de classification, que nous tentons ici de résoudre à l'aide d'un réseau de neurones. La prédiction de la fréquence fondamentale, dans le cas des trames voisées, est un problème de régression, abordé ici à l'aide d'une approche par mélange de gaussiennes. Un schéma récapitulatif de cette approche directe pour la conversion visuo-acoustique est proposé à la Figure 4.1.

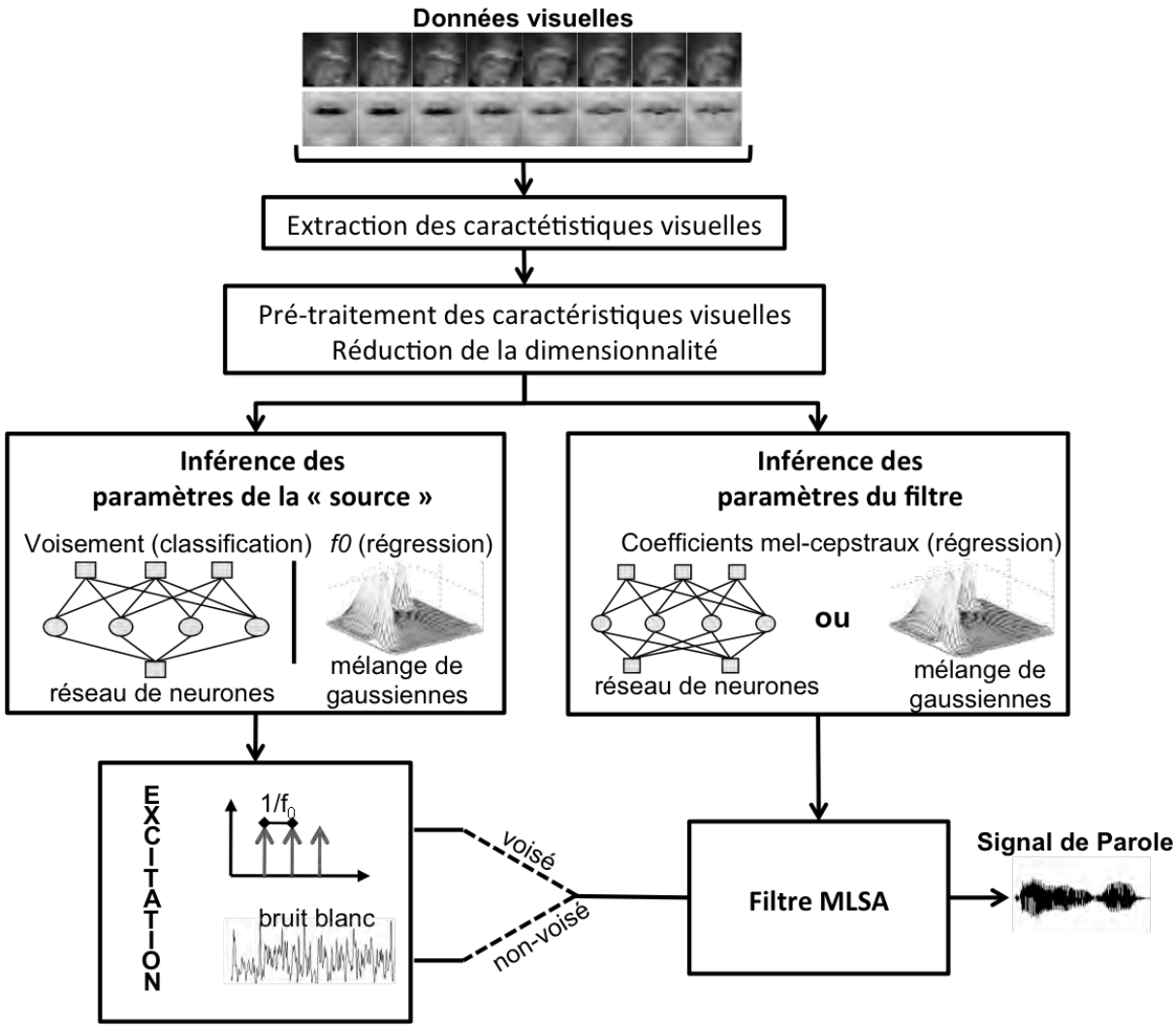


Figure 4.1 : Conversion visuo-acoustique, approche directe

Les résultats présentés dans ce chapitre sont basés sur l'utilisation de caractéristiques visuelles obtenues exclusivement par des approches globales (section 3.2.3). Une comparaison des approches par segmentation (extraction du contour de la langue, section 3.2.2) et des approches globales pour la conversion visuo-acoustique directe est effectuée dans (Hueber *et al.*, 2007a)[50]. Cette étude montre une supériorité de l'approche globale. Moins performantes et, comme précédemment évoqué, peu robustes face à certaines configurations articulatoires, les approches par segmentation sont, à ce stade de l'étude, abandonnées.

[50] Dans cette étude, les paramètres d'un filtre de synthèse autorégressif (représentés par des coefficients LSF) sont inférés à l'aide d'un réseau de neurones. Les caractéristiques visuelles ultrasonores sont obtenues à l'aide de l'approche *EigenTongues*. Une base différente des bases B1 et B2 (section 2.4.2) est utilisée. Cette dernière est construite à l'aide du système d'acquisition SA1 (section 2.4.1), contient 45 minutes de parole environ, et est basée sur le corpus de texte « IEEE Harvard » (Anon, 1969). Une vue de profil des lèvres, décrite à l'aide d'une approche par segmentation, est utilisée en complément de l'image ultrasonore.

4.2. Prétraitement des caractéristiques

4.2.1. Sur-échantillonnage des caractéristiques visuelles

L'approche directe de la conversion visuo-acoustique est basée sur une description « trame-synchrone » des flux visuels et du flux audio[51]. Pour les bases de données B1 et B2, les flux d'images sont cadencés respectivement à 29.97 ips et 60 ips, soit une image (donc un vecteur de caractéristiques visuelles) toutes les 33 ms et 16 ms. Or, pour les deux bases, l'analyse par « fenêtre glissante » des signaux audio, réalisée pour le calcul des coefficients mel-cepstraux et pour l'estimation de la fréquence fondamentale, fournit un vecteur de caractéristiques acoustiques toutes les 10 ms (fenêtre de 20 ms et décalage de 10 ms). Pour permettre la mise en correspondance des deux modalités, les flux de caractéristiques visuelles sont sur-échantillonnés à 100 Hz, par interpolation linéaire.

4.2.2. Fusion des caractéristiques visuelles des modalités ultrasonore et vidéo

Chacune des modalités visuelles est décrite par un flux de vecteurs de caractéristiques de dimension 90, chaque vecteur étant composé de 30 coefficients statiques (pour les approches par TCD et *EigenTongues*/*EigenLips*) auxquels s'ajoutent 60 coefficients dynamiques (dérivées premières et secondes). Afin de combiner ces deux modalités en vue de la modélisation visuo-acoustique, une stratégie dite de « fusion au niveau des caractéristiques »[52] est ici adoptée; un vecteur de caractéristiques « global » est obtenu par concaténation des vecteurs associés aux modalités ultrasonore et vidéo.

Cette stratégie facilite la mise en œuvre de la modélisation visuo-acoustique. Néanmoins, elle ne permet pas de prendre en compte le phénomène d'anticipation gestuelle. Une illustration de ce phénomène est donnée dans (Livescu, 2005), pour le mot anglais « *several* » ([s eh v r ah l]). Lors de l'articulation de ce mot, la langue se « désynchronise » des lèvres, elle anticipe en débutant la rétroflexion nécessaire à la production du [r], avant même que les lèvres ne se soient rapprochées pour articuler la fricative [v]. La prise en considération de ce phénomène de désynchronisation des gestes articulatoires entre eux, sera discutée ultérieurement dans le cadre de l'approche indirecte de la conversion visuo-acoustique.

4.2.3. Choix des caractéristiques acoustiques

L'analyse mel-cepstrale du signal audio décrit chaque trame de ce signal par 25 coefficients. Pour l'inférence des paramètres du filtre, seuls les 13 premiers coefficients sont

[51] L'asynchronie entre le geste articulatoire et la réalisation acoustique n'est donc pas pris en compte. Ce point sera discuté en fin de chapitre.

[52] Ce terme, traduit de l'anglais *Feature Fusion*, est emprunté au domaine de la reconnaissance audiovisuelle de la parole.

utilisés. De façon classique, nous considérons ici qu'ils décrivent la forme générale de l'enveloppe spectrale, les derniers coefficients ne concernant que les structures fines du spectre.

4.2.4. Réduction de la dimensionnalité de l'espace des caractéristiques visuelles

Chaque couple d'images ultrasonore et vidéo est donc décrit par un vecteur de 180 caractéristiques visuelles. Afin de faciliter l'apprentissage de la fonction de conversion visuo-acoustique en limitant le nombre de paramètres à estimer, différentes techniques de réduction de la dimensionnalité ont été envisagées. Les deux principales méthodes mises en œuvre dans le cas de l'approche directe sont l'analyse en composante principale et l'analyse de corrélation canonique. Cette dernière méthode, qui s'est avérée être la plus performante, est brièvement détaillée ci-après[53].

On note *n* et *m*, les dimensions respectives des vecteurs de caractéristiques visuelles et acoustiques[54] (dans notre cas, *n=180* et *m=13*), $X \in \mathfrak{M}_{n,k}(\mathbb{R})$ et $Y \in \mathfrak{M}_{m,k}(\mathbb{R})$, deux ensembles (matrices) de *k* vecteurs de caractéristiques, respectivement visuelles et acoustiques. L'objectif de l'analyse de corrélation canonique est de déterminer les directions $\mathbf{a}_1 \in \mathbf{U}^n$ et $\mathbf{b}_1 \in \mathbf{U}^m$ (avec $\mathbf{U}^d = \{\mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\| = 1\}$) telles que les projections de *X* sur $\mathbf{a}_1$ et de *Y* sur $\mathbf{b}_1$ soient les plus corrélées possibles. Ceci s'écrit :

$$(\mathbf{a}_1, \mathbf{b}_1) = \underset{(\mathbf{a}_1, \mathbf{b}_1) \in \mathbf{U}^n \mathrm{x} \mathbf{U}^m}{\operatorname{argmax}} \; corr(\mathbf{a}_1^t X, \mathbf{b}_1^t Y) \qquad \text{(Équation 4.1)}$$

En notant C_{AB} la matrice de covariance de deux vecteurs aléatoires *A* et *B*, on montre que $\mathbf{a}_1$ est le premier vecteur propre de la matrice *M* définie par $M = C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{YX}$, le vecteur $\mathbf{b}_1$ se déduisant ensuite de la relation $\mathbf{b}_1 = C_{YY}^{-1} C_{YX} \mathbf{a}_1$. En prenant en considération les *d* vecteurs propres de *M*, on obtient alors deux bases orthonormées, notées $\{\mathbf{a}_i\}_{i \in [1..d]}$ et $\{\mathbf{b}_i\}_{i \in [1..d]}$, qui décrivent deux espaces dans lesquels les corrélations des projections de *X* et *Y* sont maximisées, dimension à dimension. En d'autres termes, l'analyse de corrélation canonique fournit les combinaisons linéaires des caractéristiques visuelles et acoustiques qui sont, entre elles, les plus corrélées possibles. Par ailleurs, le nombre *d* de vecteurs propres de *M* étant égal au rang de la matrice C_{XY}, on obtient $d = \min(n, m)$. C'est cette propriété qui permet d'utiliser ici l'analyse par corrélation canonique comme technique de réduction de la dimensionnalité.

Dans notre cas, les vecteurs de caractéristiques visuelles de dimension 180 (section 4.2.2) sont projetés dans un espace de dimension 13 (section 4.2.3). La dimension des vecteurs de caractéristiques acoustiques reste, elle, inchangée. L'impact de l'analyse de corrélation canonique sur les données visuelles et acoustiques est illustré à la Figure 4.2. Sont représentées les

[53] Une description plus complète de l'analyse de corrélation canonique est disponible dans (Bredin, 2007).

[54] La fréquence fondamentale n'ayant de sens que sur les trames voisées, seuls les coefficients mel-cepstraux sont utilisés comme « caractéristiques acoustiques » pour l'analyse de corrélation canonique.

évolutions, sur une phrase (400 trames), des deux premières projections des vecteurs de caractéristiques visuelles et acoustiques. Ces projections sont notées respectivement $\{X_i\}$ et $\{Y_i\}$ avec $i = \{1,2\}$ et définies par $X_i = \sum_{k=1}^{n} a_{ik} X$ et $Y_i = \sum_{k=1}^{m} b_{ik} Y$. L'analyse de corrélation canonique est ici effectuée sur un ensemble d'apprentissage de 48559 trames, construit à partir des 150 premières phrases de la base de données ; les caractéristiques visuelles sont, dans cet exemple, obtenues à l'aide de l'approche *EigenTongue/EigenLips*.

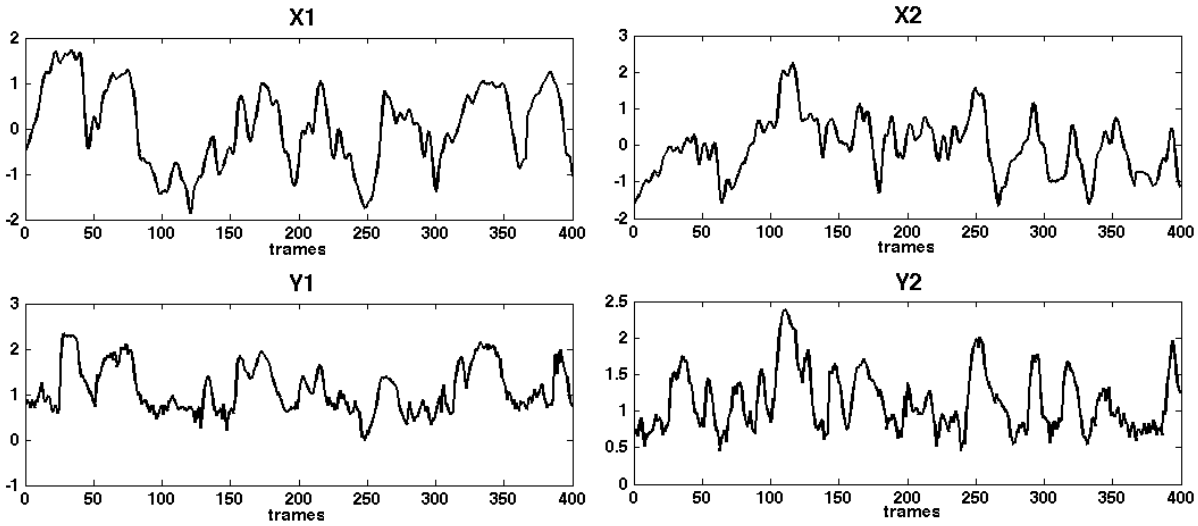


Figure 4.2 : Analyse de corrélation canonique. Evolution des deux premières projections des caractéristiques visuelles (X_1, X_2) et acoustiques (Y_1,Y_2), base B1, caractéristiques visuelles (initiales) du type *EigenTongue/EigenLips*.

4.3. Inférence des paramètres du filtre : approche par réseaux de neurones

4.3.1. Principe

La brève introduction aux réseaux de neurones proposée dans cette section s'appuie sur (Dreyfus *et al.*, 2008), ouvrage que le lecteur pourra consulter pour une présentation complète de cet outil.

Les Réseaux de Neurones Artificiels (RNA) sont des combinaisons de fonctions non linéaires élémentaires appelées « neurones formels » ou simplement « neurones ». Un neurone réalise une fonction non linéaire bornée de ses entrées :

$$y = F(\sum_{i=1}^{n} \omega_i x_i + \beta) \qquad \text{(Équation 4.2)}$$

où $\{x_i\}_{i\in[1..n]}$ sont les variables (les entrées), $\{\omega_i\}_{i\in[1..n]}$ les paramètres (poids synaptiques) (β est le « biais ») et F, une fonction non-linéaire nommée « fonction d'activation ». Il existe de multiples façons d'agencer ces neurones élémentaires et construire ainsi des réseaux complexes. Pour l'inférence des coefficients mel-cepstraux, nous utilisons un perceptron multicouche

(PMC) dont l'architecture est illustrée à la Figure 4.3. Le PMC est un réseau de neurones à couches, non bouclé, dont les neurones de la couche cachée ont une fonction d'activation du type sigmoïde (tangente hyperbolique). Soit M le nombre de neurones de la couche cachée, la $i^{ème}$ sortie du réseau notée y_i est définie par (cas d'un neurone de sortie linéaire) :

$$y_i = \sum_{j=1}^{M} w_{ij} F(\sum_{k=1}^{N} \omega_{jk} x_k + \beta_j) + b_i \qquad \text{(Équation 4.3)}$$

L'estimation des paramètres ω et w s'effectue par apprentissage supervisé, c'est-à-dire par minimisation d'une fonction de coût qui reflète l'écart entre les observations qui figurent dans une base d'apprentissage et les prédictions du modèle. Ce problème d'optimisation peut se résoudre à l'aide de différents algorithmes qui ne seront pas détaillés ici. On citera néanmoins la méthode du « gradient conjugué », utilisée dans les expériences décrites ci-après.

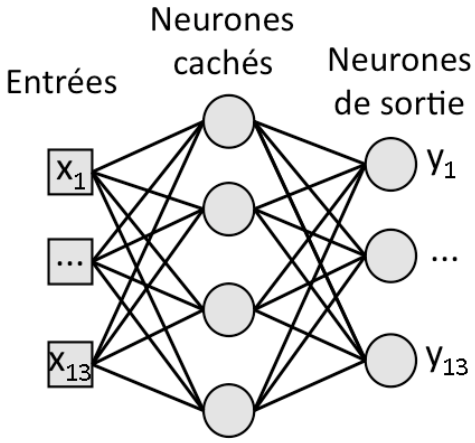


Figure 4.3 : Perceptron multicouches mis en œuvre : 13 entrées, 13 sorties, 4 neurones cachés (le biais n'est pas représenté)

4.3.2. Mise en œuvre

Bases d'apprentissage, de validation et de test

Les bases de données B1 et B2 sont respectivement divisées en 34 et 37 listes de 30 phrases chacune. Pour les deux bases, les 10 premières listes sont utilisées comme base d'apprentissage, les 5 listes suivantes comme base de validation et les listes restantes comme base de test (soit 19 listes de test pour la base B1, et 22 pour la base B2).

Le choix d'un corpus d'apprentissage de taille relativement restreinte par rapport à la quantité de données disponibles, est motivé par l'expérience. L'utilisation de corpus de taille plus importante n'a en effet pas permis d'obtenir des modèles présentant une meilleure capacité de généralisation.

Architecture du réseau

Le PMC mis en œuvre dispose de 13 entrées, d'une seule couche de neurones cachés dont la fonction d'activation est la fonction sigmoïde, et d'un étage de sortie composé de 13 neurones dont la fonction d'activation est linéaire. La sélection de modèle, c'est-à-dire le choix du nombre de neurones sur la couche cachée, s'effectue par validation croisée. La performance du modèle est alors évaluée au sens des moindres carrés (voir section 4.6). Dans les diverses expériences réalisées, les meilleurs résultats sont obtenus avec des architectures à 50 neurones cachés, l'utilisation d'architectures plus complexes n'apportant pas d'amélioration significative.

Régularisation

Deux méthodes de régularisation sont mises en œuvre afin d'éviter de créer des modèles surajustés : « l'arrêt prématuré » *(early stopping* en anglais) et la « modération des poids » (*weight decay*). La première technique consiste à suivre, pendant l'apprentissage, l'évolution de la performance du modèle sur l'ensemble de validation et d'arrêter l'apprentissage lorsque cette dernière commence à diminuer. La seconde technique consiste à ajouter un terme supplémentaire à la fonction de coût qui pénalise les poids trop élevés[55] (paramètres du modèle). La fonction de coût régularisée J_α peut s'écrire sous la forme :

$$\begin{aligned} J_\alpha(\omega) &= \alpha J(\omega) + (1-\alpha)\Omega(\omega) \\ &= \alpha \sum_{k=1}^{N_a} \left\| \mathbf{y}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}_k, \omega) \right\|^2 + (1-\alpha) \sum_{i=1}^{q} \left\| \omega_i \right\|^2 \end{aligned} \qquad \text{(Équation 4.4)}$$

où $\mathbf{x}_k$ désigne le vecteur des entrées pour l'exemple *k*, $\mathbf{y}(\mathbf{x}_k)$ la valeur de la mesure correspondante, $\mathbf{g}(\mathbf{x}_k, \omega)$ la valeur calculée par le réseau de neurones muni des poids ω pour le vecteur de variables $\mathbf{x}_k$, N_a le nombre d'exemples dans l'ensemble d'apprentissage, q le nombre de paramètres du modèle. $J(\omega)$ désigne la fonction de coût des moindres carrés et $\Omega(\omega)$ le terme de régularisation. Le choix de l'hyperparamètre α fait l'objet d'un compromis entre faible erreur de modélisation d'une part, et risque de surajustement et donc faible capacité de généralisation d'autre part. Le choix du paramètre α est effectué par validation croisée, il est fixé à 0.6.

[55] Avec une valeur élevée du poids associé à un neurone caché, une faible variation de la variable en entrée du neurone peut entraîner une variation brutale de sa fonction d'activation sigmoïde et donc de sa sortie. Le réseau de neurones dans son ensemble peut alors devenir instable.

4.4. Inférence des paramètres du filtre : approche par mélange de gaussiennes

4.4.1. Principe

La seconde méthode proposée pour l'inférence des paramètres du filtre s'appuie sur la modélisation de l'espace visuo-acoustique par un « mélange de gaussiennes », plus connu sous les acronymes anglo-saxons GMM (pour *Gaussian Mixture Model*). Cette approche est inspirée des techniques mises en œuvre dans un système de conversion de la voix. Ce type de système effectue alors une conversion que l'on peut qualifier « d'acoustico-acoustique » puisqu'ayant lieu entre les caractéristiques acoustiques d'un locuteur « source », et celles d'un locuteur « cible ». Parmi les contributions majeures à la conversion de la voix, sur lesquelles s'appuie l'approche décrite ci-après, figurent (Stylianou, 1990), (Kain, 2001), (Toda *et al.*, 2001), et (Chen *et al.*, 2003).

Dans une modélisation par mélange de gaussiennes, chaque observation $\mathbf{x}$ est considérée comme une réalisation d'une variable aléatoire. La densité de probabilité[56] de cette variable aléatoire, notée ici $p(\mathbf{x} \mid \Theta)$ (Θ est l'ensemble des paramètres du modèle), est décrite comme une somme de m composantes, telle que :

$$p(\mathbf{x} \mid \Theta) = \sum_{i=1}^{m} \alpha_i p_i(\mathbf{x} \mid \theta_i) \qquad \text{(Équation 4.5)}$$

avec $\mathbf{x} = [x_1,...,x_d]$ un vecteur aléatoire de dimension d (égale à la dimension du vecteur de caractéristiques), $\theta_i = \{\mu_i, \Sigma_i\}$ les paramètres d'une distribution normale (ou gaussienne) de moyennes $\mu_i = [\mu_{i1},...,\mu_{id}]$, et de matrice de covariance Σ_i (avec $\Sigma_i \in \mathfrak{M}_{d,d}$), tel que :

$$p_i(\mathbf{x} \mid \theta_i) = N(\mathbf{x}, \mu_i, \Sigma_i)$$
avec (Équation 4.6)
$$N(\mathbf{x}, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{-1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right]$$

Les poids $\{\alpha_i\}_{i \in [1..m]}$ associés à chacune des m composantes vérifient les propriétés $\sum_{i=1}^{m} \alpha_i = 1$ et $\forall i \in [1..m],\ \alpha_i \geq 0$. Un modèle du type « mélange de gaussiennes » est alors entièrement déterminé par les m vecteurs de moyennes de dimension *1xd*, les m matrices de covariances de dimension *dxd* et les m poids $\{\alpha_i\}_{i \in [1..m]}$, tel que : $\Theta = \{\alpha_i, \mu_i, \Sigma_i\}_{i \in [1..m]}$.

L'estimation de ces paramètres à partir d'un ensemble d'observations (ensemble d'apprentissage) s'effectue traditionnellement à l'aide de l'algorithme « EM » (pour *Expectation-Maximization* en anglais) (Dempster *et al.*, 1977). Soit $[\mathbf{x}_1,...,\mathbf{x}_N]$ un ensemble de N

[56] Dans ce mémoire, nous utiliserons la notation *P(x)* pour les probabilités (lorsque la variable x est discrète), et *p(x)* pour les densités de probabilité continues et les fonctions de vraisemblance (lorsque la variable x est continue).

observations (chaque observation étant de dimension *d* telle que $\mathbf{x}_t = [x_{t1},...,x_{td}]^T$), l'algorithme EM trouve l'ensemble des paramètres qui maximise, en fonction de ces observations, le logarithme de la fonction de vraisemblance du modèle (ou « log-vraisemblance »), notée $L(\Theta)$ et défini par :

$$L(\Theta) = \sum_{t=1}^{N} \log p(\mathbf{x}_t \mid \Theta) \qquad \text{(Équation 4.7)}$$

L'algorithme EM, dont une description très complète figure notamment dans (Stylianou, 1990)[57] ne sera pas entièrement détaillé ici. On retiendra cependant qu'il s'agit d'un algorithme itératif, chaque itération se décomposant en deux étapes : l'étape d'estimation et l'étape de maximisation. Dans le cas d'un modèle du type « mélange de gaussiennes », l'étape d'estimation consiste, à partir du modèle Θ^n estimé à l'itération *n*, à évaluer la loi de probabilité de variables discrètes dites « cachées » (ou manquantes), notée c_i. Pour une observation donnée $\mathbf{x}_t$, cette dernière représente la probabilité conditionnelle *a posteriori* que cette observation soit « générée » par la $i^{ème}$ gaussienne, notée c_i . Elle est définie par :

$$P(c_i \mid \mathbf{x}_t, \Theta^n) = \frac{\alpha_i N(\mathbf{x}_t, \mu_i^n, \Sigma_i^n)}{\sum_{p=1}^{m} \alpha_p N(\mathbf{x}_t, \mu_p^n, \Sigma_p^n)} \qquad \text{(Équation 4.8)}$$

A partir des quantités estimées, les paramètres d'un nouveau modèle Θ^{n+1}, sont alors déterminés lors de l'étape de maximisation à partir des relations suivantes :

$$\begin{aligned}
\alpha_i^{n+1} &= \frac{1}{N} \sum_{t=1}^{N} P(c_i \mid \mathbf{x}_t, \Theta^n) \\
\mu_i^{n+1} &= \frac{\sum_{t=1}^{N} P(c_i \mid \mathbf{x}_t, \Theta^n) \mathbf{x}_t}{\sum_{t=1}^{N} P(c_i \mid \mathbf{x}_t, \Theta^n)} \\
\Sigma_i^{n+1} &= \frac{\sum_{t=1}^{N} P(c_i \mid \mathbf{x}_t, \Theta^n)(\mathbf{x}_t - \mu_i^n)(\mathbf{x}_t - \mu_i^n)^T}{\sum_{t=1}^{N} P(c_i \mid \mathbf{x}_t, \Theta^n)}
\end{aligned} \qquad \text{(Équation 4.9)}$$

Pour l'élaboration de la fonction de conversion visuo-acoustique, nous adoptons l'approche mise en œuvre dans (Kain, 2001). Cette dernière est basée sur la modélisation, par un mélange de gaussiennes Θ, de la densité conjointe $p(X,Y)$, où X et Y représentent les variables aléatoires associées respectivement aux espaces source et cible. Les paramètres du modèle sont estimés à

[57] On consultera notamment les pages 118 à 124.

l'aide de l'algorithme EM (critère du maximum de vraisemblance), à partir d'une matrice d'observations « conjointes », notée $\mathbf{Z} \in \mathfrak{M}_{2d,N}$, définie par:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{d1} & \cdots & x_{dN} \\ \\ y_{11} & \cdots & y_{1N} \\ \vdots & \ddots & \vdots \\ y_{d1} & \cdots & y_{dN} \end{pmatrix} \qquad \text{(Équation 4.10)}$$

Dans notre cas, $\mathbf{X} \in \mathfrak{M}_{d,N}$ et $\mathbf{Y} \in \mathfrak{M}_{d,N}$ sont deux ensembles constitués respectivement de N observations visuelles et acoustiques. La fonction de transformation F qui fournit l'estimation acoustique $\hat{\mathbf{y}}$ à partir de l'observation visuelle $\mathbf{x}$ est alors définie comme une somme pondérée de régressions linéaires, tel que :

$$\hat{\mathbf{y}} = F(\mathbf{x}) = \sum_{i=1}^{m} (W_i \mathbf{x} + b_i) \cdot P(c_i \mid \mathbf{x}) \qquad \text{(Équation 4.11)}$$

avec m le nombre de gaussiennes, $P(c_i \mid \mathbf{x})$ la probabilité conditionnelle *a posteriori* que l'observation $\mathbf{x}$ soit générée par la composante gaussienne c_i, W_i et b_i respectivement la matrice de transformation et le vecteur de biais associés à c_i. A l'aide des notations utilisées aux équations 4.5 et 4.6, et de façon similaire à l'équation 4.8, Kain montre que :

$$W_i = \Sigma_i^{YX} (\Sigma_i^{XX})^{-1}$$
$$b_i = \mu_i^Y - \Sigma_i^{YX} (\Sigma_i^{XX})^{-1} \mu_i^X$$
$$P(c_i \mid \mathbf{x}) = \frac{\alpha_i N(\mathbf{x}, \mu_i^X, \Sigma_i^{XX})}{\sum_{p=1}^{m} \alpha_p N(\mathbf{x}, \mu_p^X, \Sigma_p^{XX})} \qquad \text{(Équation 4.12)}$$

avec

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{XX} & \Sigma_i^{XY} \\ \Sigma_i^{YX} & \Sigma_i^{YY} \end{bmatrix} \text{ et } \mu_i = \begin{bmatrix} \mu_i^X \\ \mu_i^Y \end{bmatrix}$$

4.4.2. Mise en œuvre

Initialisation de l'algorithme EM

Un point délicat dans la mise en œuvre de cette approche de conversion concerne l'initialisation de l'algorithme EM. Dans l'implémentation effectuée, le modèle est initialisé par quantification vectorielle, à l'aide de l'algorithme « k-moyennes » (pour lequel le paramètre k est égal au nombre de gaussiennes du mélange). Cet algorithme partitionne un ensemble de N observations en k ensembles. Les moyennes des k gaussiennes du mélange sont alors initialisées

par la valeur des barycentres de ces ensembles. Le choix des N observations utilisées pour le partitionnement initial s'effectue par tirage aléatoire. Le nombre N est ici fixé à 5000 trames « visuo-acoustiques ». Le choix du nombre k est discuté plus loin.

Critère d'arrêt de l'algorithme EM

Un autre point important dans l'apprentissage des paramètres du modèle concerne le choix d'un critère d'arrêt pour l'algorithme EM. L'approche adoptée consiste à arrêter la ré-estimation des paramètres du modèle lorsque la croissance de la log-vraisemblance est inférieure à un certain seuil, fixé ici à 0.001. Ce seuil est généralement atteint après une dizaine d'itérations, comme l'illustre la Figure 4.4 (en pratique, 15 itérations de l'algorithme EM sont réalisées).

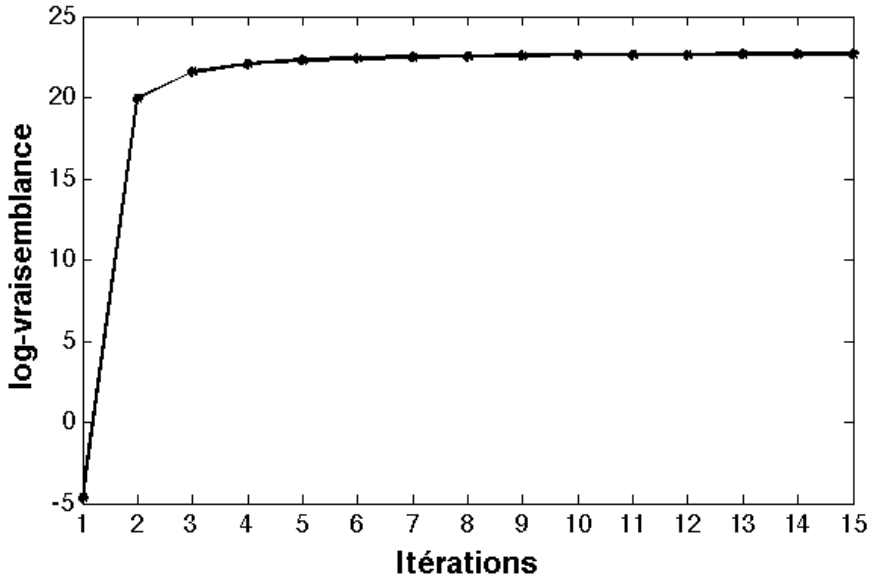


Figure 4.4 : Convergence de l'algorithme EM.

Choix du nombre de gaussiennes

Le choix du nombre de gaussiennes s'effectue par validation croisée. Les approches mises en œuvre pour l'évaluation des performances du modèle sur les ensembles de validation et de test sont décrites à la section 4.6.1. Les diverses expériences ont montré que le nombre de 25 gaussiennes était un nombre optimal, au dessus duquel aucune amélioration significative ne pouvait être observée.

4.5. Inférence des paramètres « de source »

4.5.1. Caractéristique « voisée/non-voisée »

L'inférence de la caractéristique « voisée/non-voisée » à partir des descripteurs visuels s'effectue ici à l'aide d'un classifieur binaire du type « réseau de neurones ». Pour sa mise en œuvre, une approche identique à celle utilisée pour l'inférence des coefficients mel-cepstraux est adoptée. Des procédures identiques à celles décrites à la section 4.3.2 sont donc utilisées pour le

partitionnement des bases de données, l'apprentissage, la régularisation et le choix des hyperparamètres. Le réseau est constitué de 13 entrées (projections des caractéristiques visuelles dans l'espace fourni par l'analyse de corrélation canonique), d'une couche de neurones cachés, et d'un unique neurone de sortie (classification binaire). A la différence du PMC mis en œuvre précédemment, la fonction d'activation du neurone de sortie est de type « sigmoïde », comme celle des neurones cachés. Dans les diverses expériences réalisées, les meilleurs résultats sont obtenus avec des architectures à 30 neurones cachés.

Pour l'apprentissage, les valeurs numériques associées aux caractères « non-voisés » et « voisés », sont respectivement 0 et 1. En phase de test, les prédictions du réseau se situent donc dans l'intervalle défini par ces deux valeurs. Le seuil de classification optimal, utilisé pour transformer les prédictions en valeurs binaires, est déterminé sur l'ensemble de validation, à l'aide d'une courbe ROC (en anglais *Receiver Operating Characteristic*). Cette dernière représente le taux de vrais positifs en fonction du taux de faux positifs, pour différents seuils (en pratique, 10 valeurs uniformément réparties sur l'intervalle [0 1]). Le seuil optimal est celui qui réalise le meilleur compromis entre fort taux de vrais positifs (prédiction correcte du caractère voisée) et faible taux de faux positifs (prédiction d'une trame voisée, à tort). La courbe ROC obtenue dans le cadre de la modélisation visuo-acoustique basée sur des caractéristiques visuelles du type *EigenTongue/EigenLips* est présentée à la Figure 4.5. Un seuil fixé à 0.6 apparaît comme un bon compromis.

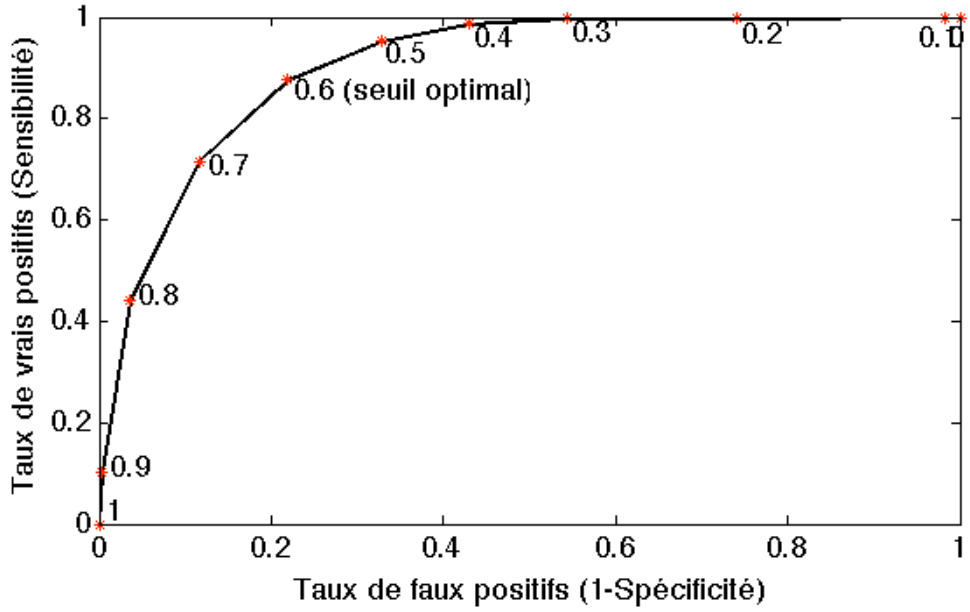


Figure 4.5 : Prédiction de la caractéristique « voisée/non-voisée ». Choix du seuil de classification optimal à l'aide d'une courbe ROC (ensemble de validation, base B2)

4.5.1. Inférence de la fréquence fondamentale

De façon similaire à la méthode décrite à la section 4.4, l'approche adoptée pour l'inférence de la fréquence fondamentale des trames voisées est basée sur une modélisation de l'espace

conjoint « visuo-acoustique » par un mélange de gaussiennes. L'unique caractéristique acoustique utilisée ici est la fréquence fondamentale. Les paramètres du modèle sont estimés à l'aide de l'algorithme EM, (section 4.4.1), à partir d'une matrice d'observation similaire à celle décrite à l'équation 4.10. $\mathbf{Y} \in \mathfrak{M}_{1,N}$ est ici un vecteur contenant les N fréquences fondamentales observées (trames voisées uniquement). Des procédures identiques à celles décrites à la section 4.4.2 sont utilisées pour l'initialisation de l'algorithme EM, son critère d'arrêt (en pratique, 15 itérations sont réalisées) et le choix du nombre de gaussiennes (les meilleurs résultats sont obtenus avec 6 gaussiennes).

4.6. Résultats et interprétations

4.6.1. Inférence des paramètres du filtre

Critères utilisés pour l'évaluation des performances

Plusieurs mesures sont utilisées pour évaluer les performances du modèle sur l'ensemble de test. La première mesure est l'erreur quadratique moyenne. Afin de pouvoir estimer l'intervalle de confiance de cette mesure, une stratégie dite de ré-échantillonnage (*bootstrap*) est adoptée. B sous-ensembles de N_{T_B} échantillons sont constitués par tirage aléatoire (avec remise) dans l'ensemble de test. Une erreur quadratique moyenne $\mathrm{EQMT_b}$ (avec $b \in [1..B]$) est calculée sur chacun de ces sous-ensembles. A l'aide des notations utilisées à l'équation 4.4, cette erreur s'écrit :

$$EQMT_b = \sqrt{\frac{1}{N_{T_B}} \sum_{k=1}^{N_{T_B}} \left\| \mathbf{y}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}_k, \omega) \right\|^2} \qquad \text{(Équation 4.13)}$$

En notant respectivement μ_{EQMT} et σ_{EQMT} la moyenne et la variance (empirique) des B $EQMT_b$, l'intervalle de confiance à 95 % est alors défini par $\mu_{EQMT} \pm 2\sigma_{EQMT}$ (approximation normale). L'erreur quadratique moyenne est par ailleurs le critère utilisé, en phase d'apprentissage, pour l'estimation des poids du réseau de neurones et pour l'ajustement par validation croisée, des différents hyperparamètres (architecture et régularisation du réseau de neurones, nombre de gaussiennes dans le mélange, etc.).

La seconde mesure mise en œuvre pour l'évaluation des performances du modèle est le coefficient de régression linéaire, noté r_i ($i \in [1..d]$, avec ici $d = 13$), défini pour la $i^{ème}$ variable de sortie par :

$$r_i = \frac{\sum_{k=1}^{N_T} (y_i(\mathbf{x}_k) - \overline{y}_i)(g_i(\mathbf{x}_k, \omega) - \overline{g}_i)}{\sqrt{\sum_{k=1}^{N_T} (y_i(\mathbf{x}_k) - \overline{y}_i)^2} \sqrt{\sum_{k=1}^{N_T} (g_i(\mathbf{x}_k, \omega) - \overline{g}_i)^2}} \qquad \text{(Équation 4.14)}$$

avec $\mathbf{y} = [y_1,...,y_d]$, $\mathbf{g} = [g_1,...,g_d]$, $\overline{y}_i$ et $\overline{g}_i$ respectivement les valeurs moyennes de y_i et g_i sur l'ensemble de test, et N_T, le nombre d'échantillons dans cet ensemble. Ce coefficient est compris entre [-1 et 1]. La performance du modèle est jugée d'autant meilleure que ce coefficient est loin de 0 (proche de -1 ou 1). Au vu des résultats expérimentaux exposés ci-après, il est également apparu intéressant d'évaluer les performances du modèle uniquement sur les d premiers coefficients mel-cepstraux. Ceci permet d'évaluer la capacité du modèle à prédire la forme très générale de l'enveloppe spectrale. Un coefficient de régression linéaire dit « moyen », noté R_d est alors introduit. Ce dernier est défini par :

$$R_d = \frac{1}{d}\sum_{1}^{d} r_i \qquad \text{(Équation 4.15)}$$

On s'intéressera plus spécifiquement aux valeurs de R_{13} et R_8, coefficients de régression linéaire moyens calculés respectivement sur l'ensemble des 13 coefficients, et sur les 8 premiers coefficients seulement.

Résultats expérimentaux

Le Tableau 4.1 et le Tableau 4.2 présentent les résultats obtenus à l'aide des approches par réseau de neurones et par mélange de gaussiennes, pour les deux bases de données et pour les deux approches d'extraction des caractéristiques visuelles (TCD et *EigenTongues/EigenLips*, voir les sections 3.2.3 et 3.3.2).

La Figure 4.6 et la Figure 4.7 présentent les diagrammes de dispersion obtenus pour chacune des deux approches, dans le cas des « meilleures » conversions visuo-acoustiques. Il s'agit des conversions effectuées sur la base B2, à partir de caractéristiques visuelles du type *EigenTongues/EigenLips*.

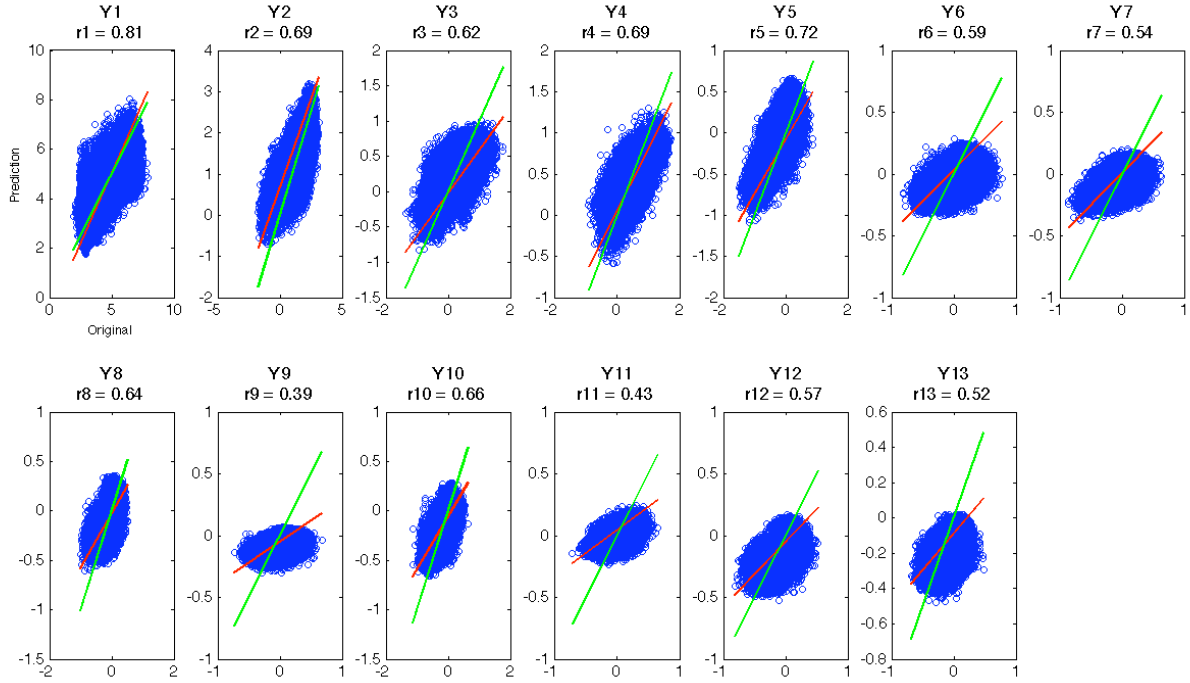


Figure 4.6 : Diagrammes de dispersion pour la conversion visuo-acoustique par réseau de neurones (base B2, approche *EigenTongues/EigenLips*, r_i est le coefficient de régression linéaire calculé à partir des prédictions du coefficient mel-cepstral Y_i, en rouge, la droite de régression linéaire, en vert la droite d'équation $y=x$).

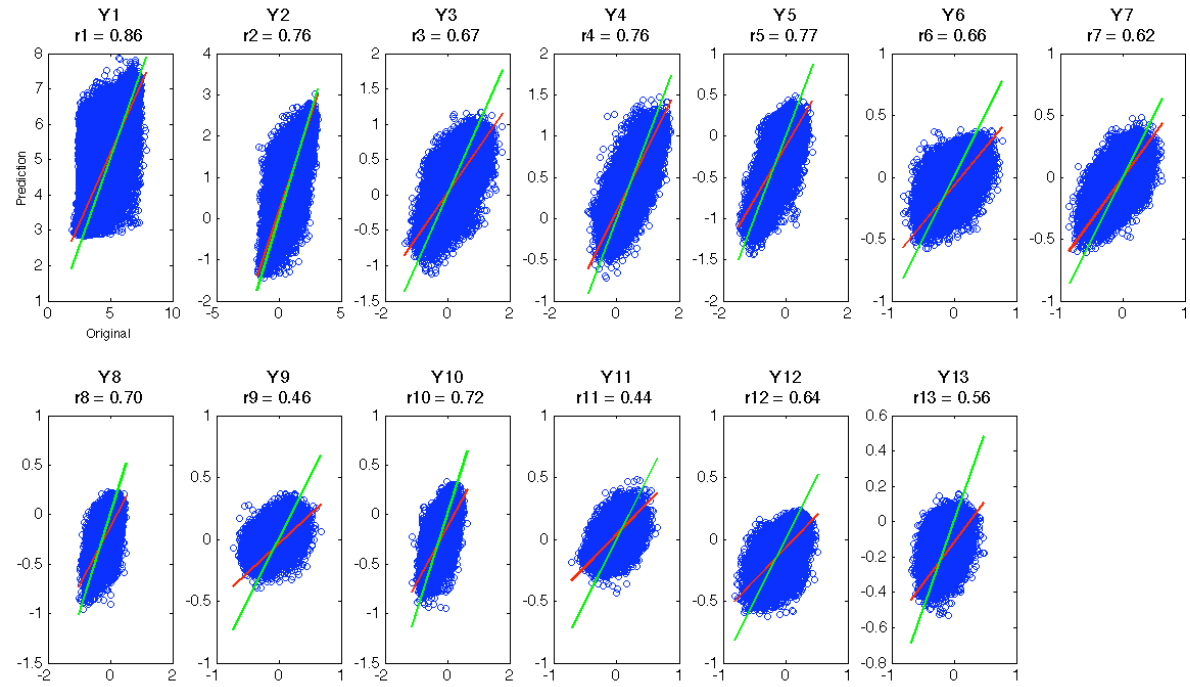


Figure 4.7 : Diagrammes de dispersion pour la conversion visuo-acoustique par mélange de gaussiennes (base B2, approche *EigenTongues/EigenLips*).

	Base B1		**Base B2**	
	TCD	*EigenTongues/EigenLips*	TCD	*EigenTongues/EigenLips*
$\mu_{EQMT} \pm \Delta_{95\%}$	1.46 ± 0.02	1.47 ± 0.02	1.49 ± 0.02	1.20 ± 0.02
(R_8,R_{13})	(0.61, 0.52)	(0.60, 0.52)	(0.61, 0.50)	(0.66, 0.61)

Tableau 4.1 : Evaluation globale de la conversion visuo-acoustique par réseau de neurones (prédiction des coefficients mel-cepstraux, $\Delta_{95\%}$ est l'intervalle de confiance à 95%).

	Base B1		**Base B2**	
	TCD	*EigenTongues/EigenLips*	TCD	*EigenTongues/EigenLips*
$\mu_{EQMT} \pm \Delta_{95\%}$	1.60 ± 0.02	1.40 ± 0.02	1.90 ± 0.02	1.10 ± 0.02
(R_8,R_{13})	(0.59, 0.51)	(0.65, 0.56)	(0.48, 0.44)	(0.73, 0.67)

Tableau 4.2 : Evaluation globale de la conversion visuo-acoustique par mélange de gaussiennes (prédiction des coefficients mel-cepstraux).

Interprétations

L'étude de ces résultats permet de comparer les différentes approches mises en œuvre pour chacune des étapes de la chaîne de traitement, de l'acquisition des données à la conversion visuo-acoustique, en passant par l'extraction des caractéristiques visuelles.

Ces résultats peuvent tout d'abord être interprétés en fonction de la base de données utilisée pour la modélisation. Rappelons que ces bases sont construites à partir du même corpus de texte, mais à l'aide de deux systèmes d'acquisition et deux locutrices différents (systèmes SA1 et SA2, voir section 2.4.1). Dans l'approche par réseau de neurones comme dans celle par mélange de gaussiennes, les résultats basés sur des caractéristiques visuelles du type *EigenTongue/EigenLips* sont nettement meilleurs sur la base B2 que sur la base B1. Dans le cas de caractéristiques visuelles du type TCD, une tendance opposée peut être observée, mais les différences entre les deux bases sont cependant moins importantes que dans le cas de l'approche *EigenTongue/EigenLips*. Ces bonnes performances obtenues sur la base B2 par rapport à la base B1, dans le cas de caractéristiques visuelles du type *EigenTongue/EigenLips*, peuvent en partie s'expliquer par la qualité des données visuelles de cette base. Les séquences ultrasonores et vidéo de la base de données B2 possèdent en effet de meilleures résolutions spatiale et temporelle que celles de la base B1.

Ces résultats permettent ensuite de comparer les deux approches mises en œuvre pour l'extraction des caractéristiques visuelles. Dans la quasi totalité des expériences, l'approche *EigenTongues/EigenLips* fournit des résultats qui sont, de façon statistiquement significative, meilleurs que ceux obtenus à partir de l'approche TCD. Bien qu'il paraisse difficile de fournir une véritable explication à ce résultat, rappelons néanmoins que les deux approches sont basées sur des hypothèses relativement différentes. L'approche par analyse en composantes principales considère comme pertinente l'information présentant une variabilité importante. L'estimation de cette variabilité nécessite une phase d'apprentissage qui rend la décomposition dépendante des données. Dans l'approche basée sur la TCD, seule l'information basse fréquence est

conservée. Certains détails portés par les hautes fréquences spatiales (notamment dans l'image ultrasonore), peuvent alors être exclus du codage, à tort.

Enfin, ces résultats permettent de comparer les deux techniques de modélisation mises en œuvre pour la conversion visuo-acoustique. Pour réaliser cette comparaison, nous nous plaçons dans le cadre des « meilleures » conversions visuo-acoustiques observées pour les deux approches. Il s'agit des conversions effectuées à partir des données de la base B2, traitées à l'aide de l'approche *EigenTongues/EigenLips*. Les deux mesures de performance que sont l'erreur quadratique moyenne et le coefficient de corrélation linéaire, montrent une assez nette supériorité de l'approche par mélange de gaussiennes. Ceci peut en partie s'expliquer par l'utilisation, dans cette approche, d'une information *a priori* sur la structure de l'espace des données à modéliser. Un modèle par mélange de gaussiennes fait référence de façon explicite à une organisation des observations en classes (*cluster*), chaque classe étant modélisée par une gaussienne. Le nombre de classes *a priori* est une information qui est « donnée » dans le cas d'une modélisation par mélange de gaussiennes (il s'agit du nombre de gaussiennes), mais qui est « inconnue » dans le cas d'une modélisation par réseau de neurones. L'apprentissage de la fonction de conversion peut donc, dans ce dernier cas, apparaître plus complexe.

L'analyse des diagrammes de dispersion (Figure 4.6 et Figure 4.7) permet une analyse des performances coefficient par coefficient. De façon attendue, les erreurs les moins importantes sont commises sur les premiers coefficients mel-cepstraux, c'est-à-dire sur la forme très générale de l'enveloppe spectrale. Dans le cas de la meilleure conversion, le coefficient de régression linéaire moyen sur les 8 premiers coefficients est égal à 0.73, contre 0.67 sur l'ensemble des coefficients mel-cepstraux. On notera par ailleurs que le meilleur coefficient de corrélation linéaire est obtenu, dans les deux approches, pour le premier coefficient (coefficient de régression linéaire supérieure à 0.8). Ce résultat peut paraitre surprenant, car ce coefficient renseigne principalement sur l'énergie moyenne du signal sur la trame d'analyse. Il ne semblait *a priori* pas évident que cette quantité puisse être bien prédite, de façon assez précise, uniquement à partir des observations articulatoires. Une illustration de la qualité de la prédiction obtenue sur ce coefficient est proposée à la Figure 4.8.

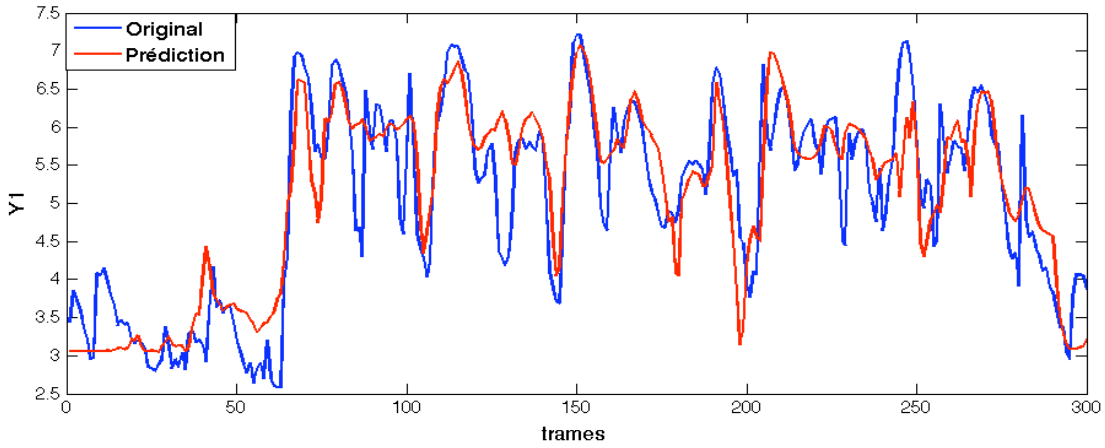


Figure 4.8 : Evolution, sur les 300 premières trames de l'ensemble de test, du premier coefficient mel-cepstral Y_1, inféré par conversion visuo-acoustique directe (approche par mélange de gaussiennes, base B2, caractéristiques visuelles du type *EigenTongues/EigenLips*).

4.6.2. Inférence des paramètres de « source »

Critères utilisés pour l'évaluation des performances

Pour évaluer la performance sur la prédiction de la caractéristique « voisée/non-voisée », trois mesures sont utilisées : la précision, la sensibilité et la spécificité. En notant VP le nombre de vrais positifs (trames réellement voisées prédites comme telles), VN le nombre de vrais négatifs (trames réellement non-voisées, prédites comme telles), FP le nombre de faux positifs (trames réellement non-voisées, prédites comme voisées) et FN le nombre de faux négatifs (trame réellement voisées, prédites comme non-voisées), ces trois mesures, notées respectivement *PRE*, *SEN* et *SPE*, sont définies par les relations suivantes :

$$PRE = \frac{VP + VN}{P + N}, \quad SEN = \frac{VP}{VP + FN}, \quad SPE = \frac{VN}{FP + VN} \qquad \text{(Équation 4.16)}$$

où P et N sont respectivement le nombre de trames voisées et non-voisées contenues dans l'ensemble de test.

L'évaluation de la performance sur la tâche de prédiction de la fréquence fondamentale est basée sur le calcul de l'erreur quadratique moyenne, à l'aide de la procédure de ré-échantillonnage décrite à la section 4.6.1.

Résultats expérimentaux

Les résultats obtenus sur les deux bases de données et pour les deux types de caractéristiques visuelles sont présentés au Tableau 4.3.

	Base B1		Base B2	
	TCD	*EigenTongues/EigenLips*	TCD	*EigenTongues/EigenLips*
PRE	0.70	0.74	0.74	0.82
SEN	0.67	0.84	0.46	0.80
SPE	0.74	0.58	0.90	0.84
$\mu_{EQMT} \pm \Delta_{95\%}(Hz)$	48 ± 5	46 ± 5	49 ± 5	45 ± 5

Tableau 4.3 : Evaluation globale de la conversion visuo-acoustique pour la prédiction de la caractéristique voisée/non-voisée (PRE, SEN, SPE) et de la fréquence fondamentale (μ_{EQMT}).

Un exemple d'évolution de la fréquence fondamentale prédite à partir des données visuelles est présenté à la Figure 4.9.

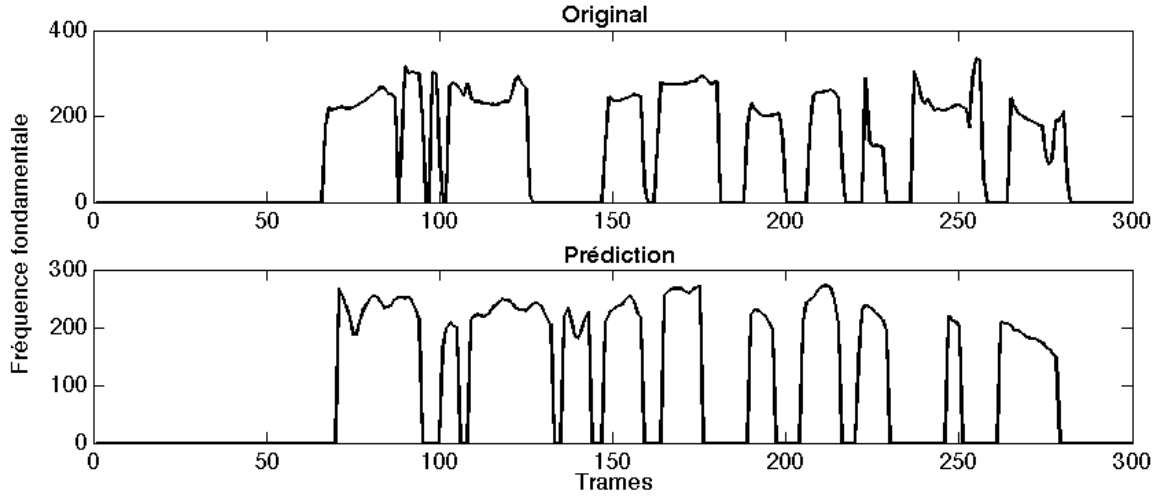


Figure 4.9 : Prédiction de la caractéristique voisée/non-voisée et de la fréquence fondamentale. En haut la référence, en bas la prédiction. Evolution sur les 300 premières trames de l'ensemble de test, base B2, conversion visuo-acoustique effectuée à partir de caractéristiques visuelles du type *EigenTongues/EigenLips*.

Interprétations

Pour la tâche de prédiction de la caractéristique « voisée/non-voisée », la meilleure précision observée est de l'ordre de 0.8, avec un couple « sensibilité/spécificité » du même ordre de grandeur. Ce résultat est obtenu sur la base B2, à partir de caractéristiques visuelles du type *EigenTongues/EigenLips*. De façon similaire à ce qui a été constaté sur la tâche d'inférence des paramètres du filtre (coefficients mel-cepstraux), et probablement pour les mêmes raisons que celles précédemment évoquées (section 4.6.1), les résultats obtenus sur cette base sont meilleurs que ceux obtenus sur la base B1. De même, l'approche *EigenTongues/EigenLips* semble mieux encoder l'information présente dans les données visuelles que l'approche par TCD. Cette bonne

performance observée peut néanmoins surprendre, la caractéristique « voisée/non-voisée » étant reliée à l'activité de l'appareil excitateur et non à celle du résonateur. Nous pouvons alors supposer que le lien entre la configuration articulatoire et le voisement est ici modélisé de façon « indirecte ». Les images ultrasonores et vidéo ne renseignent évidemment pas directement sur l'activité de l'appareil excitateur. Néanmoins, on peut raisonnablement supposer qu'il existe des corrélations assez fortes entre la configuration articulatoire et la caractéristique « voisée/non-voisée », corrélations sur lesquelles la modélisation visuo-acoustique peut s'appuyer. En mettant pour l'instant de côté les phénomènes d'anticipation gestuelle, de coarticulation et d'asynchronie entre le geste articulatoire et la réalisation acoustique (geste préphonatoire), on peut supposer qu'une image vidéo montrant une bouche fermée ne peut être associée qu'à une trame de signal non-voisée. A l'inverse, à une configuration articulatoire « stable » (identifiée par la valeur des descripteurs visuels dynamiques), typique de la production de certaines voyelles, correspond vraisemblement à une trame « voisée ». Pour mettre en défaut cet élément de réponse, on peut cependant considérer le cas des phonèmes qui ne se diffèrent que par leur caractéristique voisée/non-voisée, comme ([v],[f]) et ([z],[s]) par exemple. Dans ce cas, on peut alors supposer que la modélisation bénéficie de ce que nous appellerons ici un « effet de corpus », causé par la répartition non-uniforme des phonèmes sur les ensembles d'apprentissage et de test. Si, dans ces deux ensembles, le nombre d'occurrences du phonème [s] est, par exemple, dix fois supérieur au nombre d'occurrences du phonème [z], le modèle aura tendance presque systématiquement à associer à la configuration articulatoire du couple ([z],[s]), la caractéristique « non-voisée ».

Si les performances obtenues sur cette tâche de classification « voisée/non-voisée » sont relativement bonnes, celles relatives à la prédiction de la fréquence fondamentale sont, en revanche, assez faibles. Sur chacune des deux bases et pour les deux types de caractéristiques visuelles utilisées, une erreur de 50 Hz en moyenne est commise (voir la dernière ligne du Tableau 4.3). De plus, l'évolution de la fréquence fondamentale inférée n'est pas du tout réaliste (voir Figure 4.9). Ce résultat était cependant attendu, la corrélation entre la fréquence fondamentale et le geste articulatoire étant *a priori* quasi-inexistante. Cette approche « directe » de la conversion visuo-acoustique ne permet donc pas la prédiction d'une intonation acceptable.

4.6.3. Evaluation du signal synthétisé

Le signal de parole est ici synthétisé à l'aide du filtre MLSA, à partir des coefficients mel-cepstraux et des paramètres « de source » prédits (caractéristique voisée/non-voisée et fréquence fondamentale). La Figure 4.10 présente un exemple typique d'un signal de synthèse obtenu par cette méthode, accompagné d'un signal dit « de référence ». Ce dernier est obtenu par « analyse-synthèse » du signal de parole original, c'est-à-dire après caractérisation de son spectre par analyse mel-cepstrale, extraction de sa fréquence fondamentale, et synthèse à l'aide du filtre MLSA à partir des paramètres extraits.

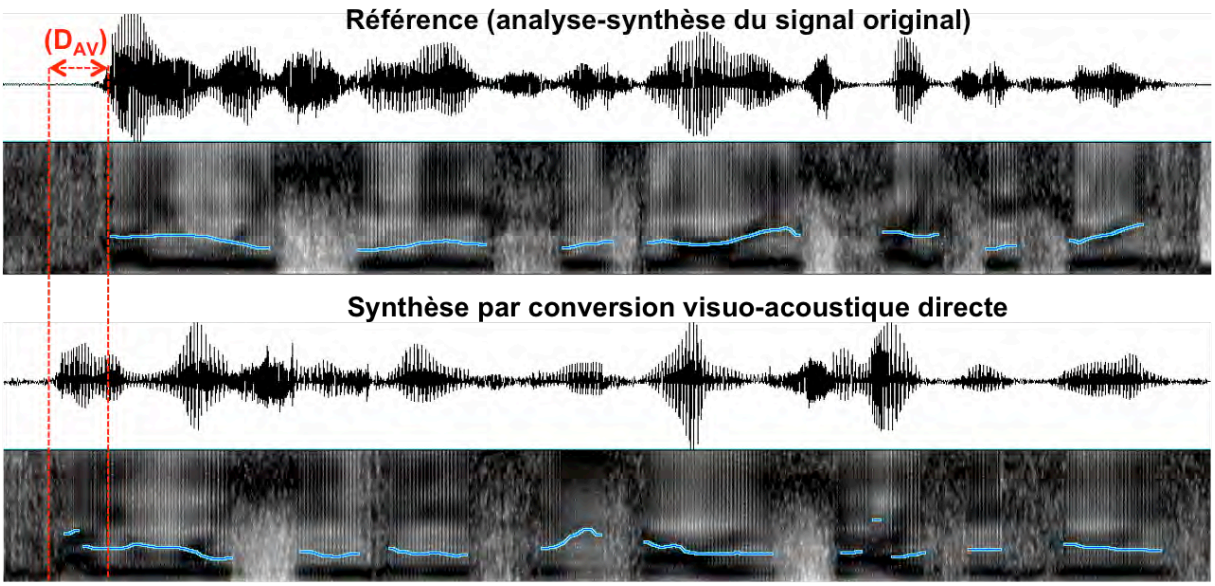


Figure 4.10 : Exemple d'un signal de synthèse dans le cas de l'approche directe de la conversion visuo-acoustique. En haut, forme d'onde et spectrogramme du signal de référence, obtenu par analyse-synthèse du signal original. En bas, forme d'onde et spectrogramme du signal reconstitué par conversion visuo-acoustique directe (approche par mélange de gaussiennes, base B2, caractéristiques visuelles du type *EigenTongues/EigenLips*). En bleu, la fréquence fondamentale[58], en rouge, une mise en évidence du phénomène d'asynchronie entre le geste articulatoire et la réalisation acoustique (discuté à la section 4.7).

D'autres exemples sonores sont accessibles depuis la page Internet associée à ce manuscrit[59]. En comparant « qualitativement » les spectrogrammes des signaux de synthèse avec ceux des signaux de référence, on remarque sur certains segments, de nombreuses similarités dans les trajectoires formantiques. A l'écoute, ces segments sont assez fidèlement reproduits. Néanmoins, de façon générale, la qualité[60] du signal de synthèse reste très insuffisante. Rappelons que le coefficient de corrélation moyen obtenu dans le cadre de la meilleure modélisation n'est que de 0.61. L'approche directe de la conversion visuo-acoustique ne semble donc pas être en mesure de fournir un signal de parole intelligible. Avec ce niveau de performance, il ne nous a pas paru utile de mettre en place un protocole d'évaluation des résultats plus approfondi (à l'aide par exemple d'un test d'intelligibilité).

4.7. Conclusions

L'approche « directe » de la conversion visuo-acoustique est une approche simple pour réaliser une transformation de l'espace des données visuelles vers l'espace des données

[58] Le calcul du spectrogramme et l'estimation de la fréquence fondamentale sont effectués à l'aide du logiciel *Praat* (Boersma et Weenink, 2009).

[59] L'URL de cette page est indiquée en introduction générale. L'exemple de la Figure 4.10 correspond à l'exemple 1 de la page Internet.

[60] Les notions de « qualité » et « d'intelligibilité » seront définies au chapitre suivant, dans le cadre de l'évaluation des résultats obtenus par l'approche indirecte de la conversion visuo-acoustique.

acoustiques. Les diverses expériences ont montré qu'il est possible, à partir de l'observation des mouvements articulatoires, de prédire de façon relativement précise la structure (très) générale de l'enveloppe spectrale (c'est-à-dire les premiers coefficients mel-cepstraux), ainsi que la caractéristique voisée/non-voisée. Cependant, l'intelligibilité du signal de synthèse obtenu par cette approche n'est pas acceptable. Bien que prometteurs, les résultats obtenus, notamment sur la tâche de prédiction des coefficients mel-cepstraux, restent insuffisants. Plusieurs raisons peuvent être évoquées pour expliquer cette performance.

Par son schéma de description « trame-synchrone » des flux visuels et du flux acoustique, cette approche ne prend pas en considération le phénomène de désynchronisation entre le geste articulatoire et la réalisation acoustique. Ce phénomène est notamment visible sur les signaux présentés à la Figure 4.10. Le geste articulatoire anticipant la réalisation acoustique pour la réalisation du premier phonème (qui dans le cas de cet exemple est le phonème [ay], comme dans le mot « *hide* » en anglais), le signal obtenu par conversion visuo-acoustique précède bien le signal de référence (d'un temps noté D_{av}). Par ailleurs, dans cette approche, la totalité de l'espace des configurations articulatoires n'est représenté que par un seul modèle (par exemple, un mélange de gaussiennes). Ce dernier n'est vraisemblablement pas assez complexe pour être robuste aux phénomènes de coarticulation, de réduction et d'assimilation. Mais la principale raison de l'échec (relatif) de cette approche tient sans doute au caractère « mal-posé » du problème que constitue la conversion visuo-acoustique. Comme nous l'avons précédemment souligné, les observations visuelles ne renseignent que partiellement sur la configuration articulatoire (position du vélum accessible seulement en cas de contact avec la langue), et évidemment pas sur l'activité laryngée. Ainsi, deux observations articulatoires identiques peuvent correspondre à deux réalisations acoustiques très différentes. Pour lever certaines de ces ambigüités, il apparaît nécessaire d'intégrer des informations supplémentaires dans la modélisation. Dans le cadre de l'approche indirecte de la conversion visuo-acoustique, qui fait l'objet du chapitre suivant, ces informations sont fournies par les niveaux de description linguistique supérieurs.

Chapitre 5. Conversion visuo-acoustique, approche indirecte

5.1. Avant-propos

La parole peut être décrite, à différents niveaux, comme un enchainement d'unités élémentaires, observables ou symboliques. On distingue généralement les niveaux : acoustique, phonétique et phonologique, morphologique et lexical, syntaxique, et sémantique (Boite *et al.*, 2000)[61]. A chacun de ces niveaux, on peut associer une unité de description caractéristique. Dans une description au niveau acoustique, l'unité caractéristique est la « trame », c'est-à-dire une portion du signal sur laquelle ce dernier est supposé stationnaire. Dans le cas de données visuelles, l'unité élémentaire de description est alors l'image. Aux niveaux phonétique[62] et phonologique, l'unité de description est le phonème. Il s'agit d'une unité symbolique, le signal observé n'est qu'une des réalisations possibles du phonème[63]. C'est de plus une unité dite « de contraste » (ou distinctive) : un phonème n'existe que parce qu'il permet de décrire un son élémentaire différent de tous les autres. Dans une description morphologique, l'unité caractéristique est le mot (forme lexicale), l'objectif étant de mettre en évidence les structures internes qui le constituent (les morphèmes). Enfin, les descriptions aux niveaux syntaxique et sémantique analysant respectivement la structure grammaticale et le sens d'un groupe de mots, on peut alors considérer que l'unité caractéristique est la phrase (et parfois même une partie plus importante du discours).

Dans l'approche directe présentée au chapitre précédent, la conversion visuo-acoustique s'effectue en reliant les plus bas niveaux de description des deux modalités, l'image pour la modalité visuelle et la trame pour la modalité acoustique. En raison du caractère « mal-posé » du problème de la parole silencieuse d'une part (absence d'activité laryngée), et de l'observation « incomplète » qui est effectuée de l'activité articulatoire d'autre part (position du vélum inconnue sauf en cas de contact avec la langue), il est apparu nécessaire d'intégrer des informations supplémentaires dans la modélisation.

Comme nous l'avons déjà souligné, les observations articulatoires et le signal audio peuvent être perçus comme deux représentations de la même information linguistique. D'un point de vue « théorique », ces dernières partagent donc *a priori* tous les niveaux de description linguistique supérieurs, du niveau phonétique jusqu'au niveau sémantique. Cette considération est le point de départ d'une nouvelle approche pour la conversion visuo-acoustique. Dans cette approche, la mise en correspondance des modalités visuelles avec la modalité acoustique ne s'effectue plus directement au niveau des observations (comme dans le cas de l'approche directe), mais à un niveau de description linguistique supérieur : le niveau phonétique. Cette

[61] Le niveau pragmatique, mentionné dans (Boite *et al.*, 2000), n'est pas considéré ici.

[62] Le niveau phonétique a notamment été abordé à la section 1.1 sous l'angle de la phonétique articulatoire.

[63] Une réalisation acoustique d'un phonème est généralement nommée « phone ».

approche est qualifiée « d'indirecte » car elle introduit dans le processus de conversion visuo-acoustique une étape intermédiaire dite de « décodage visuo-phonétique ». Cette étape, qui consiste à identifier dans le flux visuel la séquence de phonèmes la plus probable, permet l'introduction de connaissances linguistiques *a priori* sur la séquence observée. Le niveau phonétique « donnant accès » au niveau morphologique et lexical, ces informations peuvent notamment prendre la forme d'une limitation sur le vocabulaire autorisé. De plus, la connaissance éventuelle du contenu lexical permet d'envisager, par l'intermédiaire d'une analyse syntaxique, l'inférence d'une intonation cible « acceptable » pour le signal de synthèse. Ces différentes considérations sont illustrées de façon schématique à la Figure 5.1.

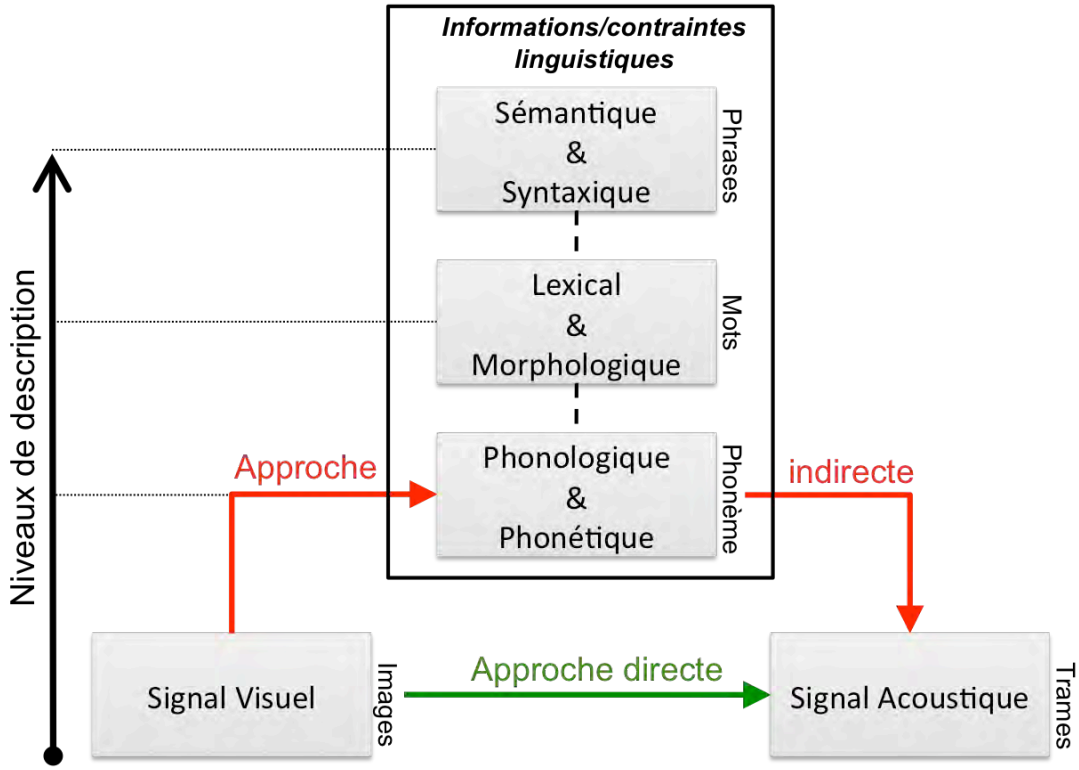


Figure 5.1 : Conversion visuo-acoustique, approche indirecte

L'apprentissage du décodeur « visuo-phonétique » s'appuie sur la segmentation automatique des flux visuels et audio au niveau phonétique et sur la modélisation des classes de segments ainsi trouvées par des modèles de Markov cachés (MMC). En phase de test, deux méthodes sont proposées pour l'inférence des paramètres acoustiques à partir de la séquence phonétique identifiée. La première utilise une approche par concaténation d'unité ; la seconde se base sur les techniques dites de « synthèse stochastique » ou « synthèse par MMC ».

5.2. Décodage visuo-phonétique

5.2.1. Principe

L'étape de décodage « visuo-phonétique »[64] fait référence aux mêmes concepts que ceux impliqués dans le problème de la reconnaissance de la parole dite « classique », c'est-à-dire celle basée sur l'analyse d'un signal audio. Ce domaine de recherche faisant l'objet d'une littérature abondante, les descriptions effectuées dans les sections suivantes resteront relativement succinctes. Elles s'appuient principalement sur (Rabiner et Juang, 1993; Barras, 1996; Boite *et al.*, 2000), références que le lecteur pourra notamment consulter pour une description plus approfondie des différents algorithmes cités.

Formulation générale du problème

En notant $\mathbf{x} = [x_1,...,x_N]$, une séquence des N observations visuelles, $\mathbf{m} \in M$ une suite de phonèmes et M l'ensemble des suites possibles, le problème du décodage visuo-phonétique consiste à déterminer la suite $\hat{\mathbf{m}} \in M$ qui vérifie la relation suivante :

$$\hat{\mathbf{m}} = \arg\max_{\mathbf{m} \in M} P(\mathbf{m} \mid \mathbf{x}) \qquad \text{(Équation 5.1)}$$

Deux approches peuvent être mises en œuvre pour résoudre ce problème. La première, dite « discriminative », consiste à postuler un modèle paramétré de la probabilité *a posteriori* $P(\mathbf{m} \mid \mathbf{x})$, et à déterminer les paramètres qui maximisent le taux de reconnaissance phonétique sur l'ensemble d'apprentissage[65]. La seconde, dite « générative » (et mise en œuvre ici pour le décodage visuo-phonétique), s'appuie sur la formule de Bayes selon laquelle :

$$P(\mathbf{m} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{m}) P(\mathbf{m})}{p(\mathbf{x})} \qquad \text{(Équation 5.2)}$$

Il s'agit de postuler, pour chaque suite de phonèmes $\mathbf{m}$ de l'ensemble d'apprentissage, un modèle paramétré de $p(\mathbf{x} \mid \mathbf{m})$, et de déterminer les paramètres tels que la séquence observée $\mathbf{x}$ correspondante ait la probabilité la plus grande possible. Une fois l'apprentissage effectué, lorsque l'on présente une séquence $\mathbf{x}$ qui n'appartient pas à l'ensemble d'apprentissage, on cherche, parmi les modèles qui ont été appris, celui qui, avec la plus grande probabilité, serait susceptible d'avoir engendré cette séquence (modèles génératifs). Dans le cadre du traitement de la parole, où les données sont naturellement des séquences, le modèle le plus couramment utilisé est le « modèle de Markov caché ». Ce dernier est décrit à la section suivante.

[64] Cette terminologie est choisie par analogie avec « décodage acoustico-phonétique ».

[65] Une approche discriminative a notamment été mise en œuvre, au chapitre précédent, pour la prédiction de la caractéristique « voisée/non-voisée » à partir des observations visuelles (voir section 4.5.1).

Modèle de Markov caché

Un modèle de Markov caché (MMC) est un modèle génératif dont le comportement peut être assimilé à celui d'un automate. A chaque unité de temps *t*, cet automate change d'état. Il émet alors une « observation » $\mathbf{o}_t$ qui est une réalisation d'une variable aléatoire qui suit la loi de probabilité dite « d'émission », b_j (avec j l'état courant tel que $j \in [1..N_q]$, N_q étant le nombre d'états de l'automate). Le changement d'état est également régi par une loi de probabilité. Pour un MMC d'ordre 1, la probabilité dite « de transition » de l'état *i* vers l'état *j* ne dépend pas des états antérieurs. En notant s_t l'état de l'automate à l'instant *t*, elle est donc définie par $a_{ij} = P(s_t = j / s_{t-1} = i)$, avec $\Sigma_{j=1}^{N_q} a_{ij} = 1$. L'état initial de l'automate est défini par l'ensemble $\{\pi_i\}_{i\in[1..N_q]}$ avec $\pi_i = P(s_0 = i)$, la probabilité que l'automate soit dans l'état *i* à l'instant *t=0* ($\Sigma_{i=1}^{N_q} \pi_i = 1$). Un modèle de Markov caché[66] est donc entièrement défini par l'ensemble des paramètres $\left\{N_q, \{\pi_i\}, \{a_{ij}\}, \{b_j\}\right\}_{(i,j)\in[1..N_q]^2}$. Il est courant de représenter un MMC sous la forme d'un graphe orienté où chaque nœud est un état et où chaque arc se voit attribuer un coût égal à la probabilité de transition. Une représentation de ce type est effectuée à la Figure 5.2.

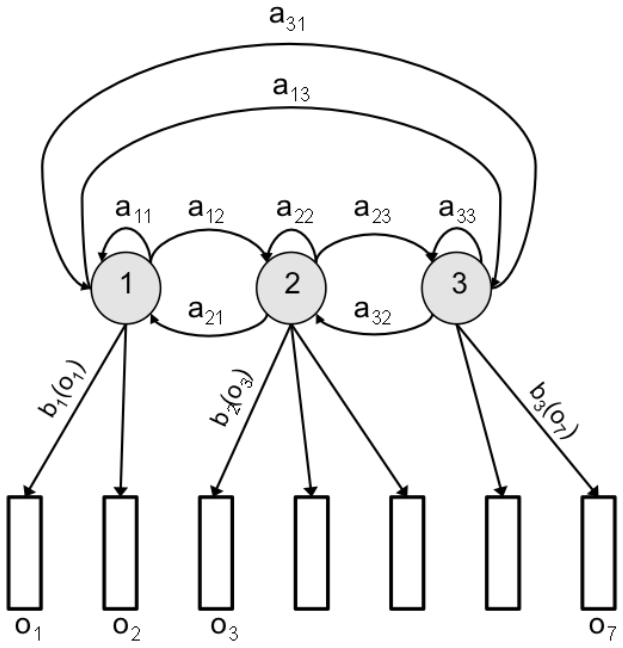


Figure 5.2 : Exemple d'un modèle de Markov caché à trois états (modèle entièrement connecté). La suite d'états *S=1,1,2,2,2,3,3* génère la séquence d'observations $\mathbf{o} = [o_1,...,o_7]$.

Dans le cas d'un MMC modélisant des observations de nature continue (par opposition à discrète), la loi de probabilité d'émission $\{b_j\}_{j\in[1,Nq]}$ est traditionnellement définie comme un

[66] En pratique, seules les observations sont accessibles et la suite d'états que devrait occuper l'automate pour générer ces observations reste inconnue. Voilà pourquoi ce modèle de Markov est qualifié de « caché ».

mélange de gaussiennes[67]. En réutilisant les notations introduites au chapitre précédent pour l'écriture des équations 4.5 et 4.6, on obtient :

$$b_j(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x}, \mu_i, \Sigma_i)$$

avec (Équation 5.3)

$$N(\mathbf{x}, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{-1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right]$$

Pour résoudre le problème formulé par l'équation 5.1, il faut être capable, d'après l'équation 5.2, d'évaluer la vraisemblance $p(\mathbf{x} \mid \mathbf{m})$ d'un modèle $\mathbf{m}$ pour une séquence de N observations $\mathbf{x} = [x_1, ..., x_N]$ donnée, c'est-à-dire la probabilité d'observer la séquence $\mathbf{x}$ sachant le modèle $\mathbf{m}$. Pour une suite d'états $Q=q_1,q_2,\ldots,q_N$ fixée (connue), cette dernière s'écrit:

$$p(\mathbf{x}, Q \mid \mathbf{m}) = \pi_{q_1} b_{q_1}(x_1) \cdot a_{q_1 q_2} b_{q_2}(x_2) \cdot \ldots \cdot a_{q_{(N-1)} q_N} b_{q_N}(x_N) \quad \text{(Équation 5.4)}$$

La suite d'états étant cependant *a priori* inconnue, la vraisemblance globale peut être obtenue en calculant la somme des vraisemblances sur toutes les suites d'états de longueur N possibles, telle que:

$$p(\mathbf{x} \mid \mathbf{m}) = \sum_{Q} p(\mathbf{x}, Q \mid \mathbf{m}) = \sum_{Q} \prod_{t=1}^{N} a_{q_{t-1} q_t} b_{q_t}(x_t) \quad \text{(Équation 5.5)}$$

avec $a_{q_0 q_1} = \pi_{q_1}$

Pour un MMC entièrement connecté à N_q états, le nombre de suites d'états possibles est de l'ordre de $(N_q)^N$. Ceci exclut toute procédure de calcul de la vraisemblance basée sur une procédure purement itérative. Aussi, une approche récursive est traditionnellement utilisée. Cette dernière est basée sur l'algorithme « avant-arrière » (pour *Forward-Backward* en anglais) qui ne sera pas détaillé ici.

L'estimation des paramètres d'un MMC à partir d'une séquence d'observations $\mathbf{x}$ (phase d'apprentissage) est classiquement effectuée à l'aide de l'algorithme de *Baum-Welch*. Cet algorithme itératif qui est une forme particulière de l'algorithme EM introduit au chapitre précédent, effectue une estimation des paramètres au sens du maximum de vraisemblance[68]. A partir d'un modèle initial $\mathbf{m}^n$, il fournit les paramètres d'un nouveau modèle $\mathbf{m}^{n+1}$ tel que $p(\mathbf{x} \mid \mathbf{m}^{n+1}) > p(\mathbf{x} \mid \mathbf{m}^n)$. L'algorithme de *Viterbi*, utilisé notamment dans la phase de décodage décrite ci-après, est souvent utilisé pour l'initialisation des modèles.

[67] Cette loi de probabilité est parfois modélisée par un réseau de neurones.

[68] De façon similaire à ce qui a été présenté au chapitre précédent, on cherchera plutôt à maximiser le logarithme de la vraisemblance (ce qui facilite la manipulation de probabilités dont les valeurs numériques sont faibles).

Le décodage d'une séquence d'observations $\mathbf{x}$ à partir d'un modèle $\mathbf{m}$ (dont les paramètres ont été estimés préalablement) consiste à trouver la suite d'états $\hat{Q}$ la plus à même de générer cette séquence. Cette suite est classiquement définie comme étant celle qui maximise la vraisemblance, telle que :

$$p(\mathbf{x},\hat{Q} \mid \mathbf{m}) = \max_{Q} p(\mathbf{x},Q \mid \mathbf{m}) = \max_{Q} \prod_{t=1}^{N} a_{q_{t-1}q_t} b_{q_t}(x_t) \qquad \text{(Équation 5.6)}$$

Cette équation, qui est une forme sous-optimale de l'équation 5.5 (seule la séquence la plus probable étant conservée), peut être résolue de façon récursive par programmation dynamique, à l'aide de l'algorithme de *Viterbi* (Forney, 1973).

Utilisation des modèles de Markov cachés pour le décodage phonétique

Dans la formulation générale du décodage visuo-phonétique, nous avons postulé un modèle génératif (jusqu'ici noté $\mathbf{m}$) pour chacune des suites phonétiques possibles. Le formalisme des modèles de Markov cachés permet de construire simplement ce modèle. Le MMC de la suite de deux phonèmes A-B peut en effet être obtenu par simple concaténation des MMCs des phonèmes A et B, c'est-à-dire en reliant l'état final du MMC du phonème A à l'état initial du MMC du phonème B.

En reconnaissance de la parole « acoustique », un phonème est traditionnellement modélisé par un MMC « gauche-droite », à trois états (émetteurs). Le choix du nombre d'états tient compte des phénomènes de coarticulation. Le début et la fin d'une réalisation phonétique peut en effet présenter des caractéristiques acoustiques différentes de la partie centrale supposée stationnaire. Le choix de la structure du MMC pour le décodage visuo-phonétique sera discuté à la section 5.2.2.

Pour un modèle donné, c'est-à-dire pour une suite de phonèmes donnée, l'algorithme de *Viterbi* détermine donc la séquence d'états qui fournit la vraisemblance maximale, sachant la séquence d'observations. La suite de phonèmes la plus probable peut donc être déterminée en comparant les vraisemblances obtenues, sur cette même séquence d'observations, pour chacune des suites possibles. Néanmoins, le nombre de suites possibles est potentiellement très grand : il n'est pas possible de toutes les comparer. La solution consiste alors à construire un modèle unique, pouvant générer toutes les suites possibles[69]. La suite phonétique la plus probable est alors définie comme le meilleur chemin dans ce modèle. Ce dernier est obtenu (de façon sous-optimale) par une variante de l'algorithme de *Viterbi*, nommé algorithme « *Token passing* » (que l'on peut traduire par « passage de jeton ») (Young *et al.*, 1989).

[69] Ce modèle pourra être construit à partir d'une « grammaire », un ensemble de règles qui régissent la structure des suites phonétiques autorisées.

Par ailleurs, l'équation 5.2 fait intervenir, en complément de la vraisemblance $p(\mathbf{x} \mid \mathbf{m})$, un terme dit de probabilité *a priori*, noté $P(\mathbf{m})$. Indépendant des observations, ce terme renseigne sur la probabilité d'observer la suite de P phonèmes $\mathbf{m} = [m_1, m_2, ..., m_P]$ dans la langue considérée[70]. Cette probabilité est définie comme le produit des probabilités conditionnelles du dernier phonème de la suite sachant tous les phonèmes précédents, telle que :

$$P(\mathbf{m}) = P(m_1)\prod_{i=2}^{P} P(m_i / m_{i-1}, m_{i-2}, ..., m_1) \qquad \text{(Équation 5.7)}$$

Pour les mêmes raisons que celles précédemment évoquées pour le décodage, il n'est pas possible de calculer cette probabilité pour toutes les suites possibles. Une modélisation du type *n-gramme* est alors utilisée. La probabilité d'apparition d'un phonème ne dépend alors que des *n-1* phonèmes précédents. Dans le cadre de cette étude, une modélisation 2-gramme, plus communément appelée bigramme, est mise en œuvre. La probabilité qu'un phonème m_j succède à un phonème m_i est estimée à l'aide de la relation suivante :

$$P(m_j / m_i) = \frac{f(m_i, m_j)}{f(m_i)}, \text{ si } f(m_i) > 0 \qquad \text{(Équation 5.8)}$$

avec $f(m_i, m_j)$ la fréquence d'apparition de la séquence (m_i, m_j), estimée sur un corpus d'apprentissage.

Enfin, les concepts généraux (brièvement) présentés ici dans le cadre du décodage « phonétique » sont directement transposables au cas du décodage « lexical »[71]. Un modèle de Markov caché d'un mot sera obtenu par concaténation des MMC des phonèmes qui le constituent. La transcription phonétique d'un mot est obtenue soit à partir d'un dictionnaire, soit par règles, à l'aide d'un phonétiseur.

5.2.2. Mise en œuvre

Les manipulations des modèles de Markov cachés (définition, estimation des paramètres et décodage) sont réalisées à l'aide de la boite à outils logicielle HTK[72] (Young, 2005).

Prétraitement des caractéristiques visuelles

Comme il sera discuté ultérieurement, chaque phonème est ici modélisé par un MMC du type « gauche-droite » à 3 états émetteurs. La plus courte suite d'états observables pour ce type d'architecture est obtenue en considérant un automate qui change d'état à chaque unité

[70] Pour simplifier l'énoncé, on confond ici volontairement la suite de phonèmes et le modèle qui la représente.

[71] L'objectif du décodeur lexical est de déterminer à partir d'une séquence d'observations, non pas la suite de phonèmes, mais la suite de mots la plus probable.

[72] HTK : *Hidden markov model ToolKit* - http://htk.eng.cam.ac.uk/

temporelle (donc à chaque observation). La longueur de la plus courte séquence que peut ainsi modéliser ce MMC est donc de trois unités temporelles (donc trois observations). Les flux visuels des bases B1 et B2 étant respectivement cadencés à 29.97 ips et 60 ips, cette longueur minimale est donc de 99 ms pour la base B1 et de 50 ms pour la base B2. Pour pouvoir « artificiellement » modéliser des phonèmes d'une taille inférieure, une stratégie de sur-échantillonnage des données visuelles, similaire à celle décrite au chapitre précédent (section 4.2.1), est adoptée pour les données de la base B1 uniquement. Pour la base B2, une autre stratégie est mise en œuvre. Basée sur le choix de la topologie des MMC, cette dernière sera explicitée ultérieurement.

Chacune des modalités visuelles est décrite par un flux de vecteurs de caractéristiques de dimension 90 (dans l'approche par TCD comme dans l'approche *EigenTongues/EigenLips*). En considérant un ensemble de 40 phonèmes, soit autant de MMC gauche-droite à 3 états (donc 5 transitions possibles), dont la densité de probabilité d'émission est modélisée par un mélange de 30 gaussiennes (matrices de covariance diagonales), le nombre de paramètres à estimer est alors de :

$$\underset{\text{modèles}}{40} \times \underset{\text{états}}{3} \times \underset{\text{gaussiennes}}{30} \times (\underset{\text{moyennes}}{180} + \underset{\text{variances}}{180}) + \underset{\text{modèles}}{40} \times \underset{\text{transitions}}{5} = \underset{\text{paramètres}}{1\,296\,200} \qquad \text{(Équation 5.9)}$$

Afin de réduire ce nombre, plusieurs stratégies de réduction de la dimensionnalité ont été envisagées. Parmi les principales méthodes testées, on citera :

- l'analyse de corrélation canonique, mise en œuvre à l'aide de la procédure décrite à la section 4.2.4 du chapitre précédent.
- l'analyse discriminante linéaire (*Linear Discriminant Analysis* en anglais, ou LDA) (Rao, 2002), très utilisée en reconnaissance audiovisuelle de la parole (Duchnowski *et al.*, 1994; Potamianos *et al.*, 2004).

Contrairement à ce qui a été observé dans les expériences menées au chapitre précédent, la mise en œuvre de ces méthodes de réduction de la dimensionnalité ne s'est pas accompagnée ici d'une amélioration significative des résultats. Les performances obtenues à partir de caractéristiques visuelles « réduites » sont même légèrement inférieures à celles observées à partir des caractéristiques « originales »[73]. Aussi, dans un souci de concision, nous ne détaillerons ni la mise en œuvre, ni les résultats obtenus avec ces méthodes.

[73] La diminution de la performance est de l'ordre de 2 % (voir la section 5.2.3. pour plus d'informations sur le critère utilisé pour l'évaluation des performances du décodeur). On notera toutefois que la réduction de la dimensionnalité permet une réduction considérable du temps nécessaire à l'apprentissage des modèles.

Ensemble de phonèmes

La transcription phonétique du corpus CMU, utilisé lors de la construction des bases de données B1 et B2, est obtenue à l'aide du dictionnaire « CMU *Pronouncing Dictionary* »[74]. Ce dernier décrit la langue anglaise à l'aide de l'ensemble de 40 phonèmes présenté au Tableau 5.1.

Phonème	Exemple	Transcription	Phonème	Exemple	Transcription
aa	*odd*	aa d	l	*lee*	l iy
ae	*at*	ae t	m	*me*	m iy
ah	*hut*	hh ah t	n	*knee*	n iy
ao	*ought*	ao t	ng	*ping*	p ih ng
aw	*cow*	k aw	ow	*oat*	ow t
ay	*hide*	hh ay d	oy	*toy*	t oy
b	*be*	b iy	p	*pee*	p iy
ch	*cheese*	ch iy z	r	*read*	r iy d
d	*dee*	d iy	s	*sea*	s iy
dh	*thee*	dh iy	sh	*she*	sh iy
eh	*ed*	eh d	t	*tea*	t iy
er	*hurt*	hh er t	th	*theta*	th ey t ah
ey	*ate*	ey t	uh	*hood*	hh uh d
f	*fee*	f iy	uw	*two*	t uw
g	*green*	g r iy n	v	*vee*	v iy
hh	*he*	hh iy	w	*we*	w iy
ih	*It*	ih t	y	*yield*	y iy l d
iy	*eat*	iy t	z	*zee*	z iy
jh	*gee*	jh iy	zh	*seizure*	s iy zh er
k	*key*	k iy	sil (silence)		

Tableau 5.1 : Ensemble de phonèmes utilisé pour décrire la langue anglaise (les informations d'accentuation présentes dans le « CMU *Pronouncing Dictionary* » ne sont pas conservées ici)

Topologie du MMC

Chaque phonème est modélisé par un MMC, du type gauche-droite, à 3 états émetteurs. Cette topologie est fixée de façon relativement arbitraire, par analogie avec la topologie standard utilisée en reconnaissance de la parole acoustique. Rappelons que le choix de cette dernière est essentiellement lié à la nécessité de prendre en compte les effets de la coarticulation. Il semble raisonnable de considérer que ces effets sont également présents dans les observations articulatoires mises en jeu ici.

Afin de pouvoir modéliser des phones d'une durée inférieure à 45 ms dans le cas de la base B2 (données visuelles à 60 ips), une transition directe entre l'état initial et l'état final est ajoutée

[74] Accessible à l'adresse suivante : http://www.speech.cs.cmu.edu/cgi-bin/cmudict

(uniquement pour modéliser les données visuelles de cette base). Enfin, pour chacun des modèles, la loi de probabilité d'émission de chacun des états est modélisée par un mélange de gaussiennes dont les matrices de covariance sont diagonales[75].

Combinaison des modalités ultrasonores et vidéo

Dans l'approche directe de la conversion visuo-acoustique présentée au chapitre précédent, une stratégie dite de « fusion au niveau des descripteurs » a été adoptée pour combiner les modalités ultrasonore et vidéo (section 4.2.2). Le formalisme des MMC permet d'envisager d'autres stratégies. Une première alternative, nommée « *Early Integration* » dans (Potamianos *et al.*, 2004) ou MMC multi-flux , consiste à combiner les deux flux visuels au niveau des états du MMC. Ceci se formalise dans la définition de la loi de probabilité d'émission b_j associée à l'état *j* (cas d'un MMC continu et d'une loi du type « mélange de gaussiennes »). En utilisant les notations de l'équation 5.3, cette dernière s'écrit alors :

$$b_j(\left[\mathbf{x}_t^{US}, \mathbf{x}_t^{VIDEO}\right]) = \prod_{S \in \{US, VIDEO\}} \left[\sum_{m_S=1}^{M_S} \alpha_{j_S m_S} N(\mathbf{x}_t^S; \mu_{j_S m_S}; \Sigma_{j_S m_S}) \right]^{\lambda_S} \qquad \text{(Équation 5.10)}$$

avec $\mathbf{x}_t^{US}$ et $\mathbf{x}_t^{VIDEO}$ les vecteurs de caractéristiques visuelles associés respectivement aux modalités ultrasonore (*US*) et vidéo. $M_S = \{M_{US}, M_{VIDEO}\}$ et $\lambda_S = \{\lambda_{US}, \lambda_{VIDEO}\}$ sont respectivement le nombre de gaussiennes utilisées dans le mélange associé à la modalité S, et le poids attribué à cette modalité. Les paramètres M_S et λ_S sont des hyperparamètres qui, dans le cadre de cette étude, sont fixés par validation croisée (le critère d'évaluation de la performance du décodeur sur l'ensemble de validation sera décrit ultérieurement). Pour les déterminer, nous adoptons les contraintes suivantes : $M_{US} = M_{VIDEO}$ et $\lambda_{US} + \lambda_{VIDEO} = 1$[76]. Si les deux modalités visuelles sont ici modélisées de façon indépendante[77], cette stratégie suppose toujours que leur évolution soit « synchrone ». Le phénomène d'anticipation gestuelle précédemment évoqué (désynchronisation de la langue et des lèvres) n'est donc toujours pas pris en considération. D'autres stratégies de combinaison des flux ont été proposées en reconnaissance audio-visuelle de la parole[78]. On citera notamment la technique des « MMC couplés » (Brand *et al.*, 1997),

[75] Contrairement à la modélisation par mélange de gaussiennes décrite au chapitre précédent, l'utilisation de matrices de covariance « pleines » est ici difficilement envisageable car le nombre de paramètres à estimer serait trop important.

[76] Cette seconde contrainte ne trouve pas de motivations de nature théorique dans le formalisme de MMC multi-flux. Elle facilite cependant la recherche des poids optimaux. En pratique, une dizaine de couples sont testés (0.1, 0.9), (0.2, 0.8), etc.

[77] La probabilité d'émission fait ici intervenir un mélange de gaussiennes par modalité, contre un seul mélange pour les deux modalités dans le cas de la stratégie de « fusion au niveau des descripteurs ».

[78] L'objectif est alors de prendre en compte, non pas la désynchronisation des articulateurs entre eux, mais celle (plus importante) qui existe entre le mouvement des lèvres et la réalisation acoustique (signal audio).

qui suppose une synchronisation des deux flux aux limites temporelles de chaque phone, et la stratégie de « *Late Integration* », qui consiste à combiner les résultats obtenus par deux classifieurs entraînés séparément sur chacune des modalités (Adjoudani et Benoît, 1996). Ces deux dernières stratégies, plus difficiles à mettre en œuvre que la stratégie « *Early Integration* », ne sont pas envisagées dans cette étude. Une autre approche est adoptée pour tenter de tenir compte, dans la modélisation, du phénomène d'anticipation gestuelle. Ce dernier est considéré comme étant fortement corrélé au contexte phonétique. Ceci peut se justifier, de façon empirique, en analysant par exemple les gestes articulatoires liés à l'articulation du [v] dans les mots « *several* » ([s eh v r ah l], exemple présenté au chapitre précédent), et « *lived* » ([l ih v d]). Pendant que les lèvres créent le lieu d'articulation du [v] (en se rapprochant), la langue se trouve, dans le premier cas, en position « rétroflexe » pour produire le [r] (anticipation), alors qu'elle reste *a priori*, dans le second cas, dans la position « centrale » nécessaire à l'articulation du phonème précédent [ih] (pas d'anticipation). La composante « ultrasonore » des vecteurs de caractéristiques visuelles associée à ces deux occurrences du phonème [v] sera donc totalement différente d'un cas sur l'autre. L'utilisation d'un modèle différent pour chacun de ces deux contextes paraît alors nécessaire.

Modélisation dépendante du contexte

Afin de prendre en considération les effets liés aux phénomènes de coarticulation et d'anticipation gestuelle, une stratégie de modélisation des classes phonétiques « en contexte » est envisagée. Un phonème est alors modélisé de façon différente en fonction de son contexte « gauche » (phonème précédent) et « droit » (phonème suivant). Le triplet « phonème précédent, phonème courant, phonème suivant » est nommé « triphone ». En décrivant la langue anglaise avec un jeu de 40 phonèmes, le nombre de triphones possibles est de l'ordre de 40^3 = 64 000. Une base de données d'une heure de parole environ ne contient pas assez d'occurrences de chacun des triphones pour pouvoir estimer de façon robuste les paramètres d'un MMC[79]. Pour permettre la mise en œuvre de ces modèles contextuels, nous utilisons une approche basée sur le partage des données d'apprentissage entre les états des MMC associés à des triphones jugés « similaires » (procédure de « *State-tying* » en anglais). Une approche par règles est adoptée pour mesurer cette similarité. Une règle se formule sous la forme d'une « question » qui porte sur le contexte phonétique gauche et droit d'un phonème. Dans le cas de l'exemple décrit précédemment (phonème [v] dans les mots *several* et *lived*), une question intéressante sera par exemple, « Est-ce que le phone suivant est une consonne rétroflexe ? ». Le phonème [v] du mot « *lived* » répondra « non », celui du mot « *several* » répondra « oui », ils ne partageront donc pas les mêmes données d'apprentissage. Pour chaque classe phonétique, ces différentes règles sont combinées dans un arbre de décision binaire. Le partage des données

[79] Le corpus CMU Arctic ne contient que 13.7 % des triphones possibles (voir Tableau 2.2).

s'effectue généralement, non pas au niveau des modèles entiers, mais au niveau des états de ces modèles. Deux états fournissant la même réponse à toutes les questions de l'arbre, partageront les mêmes paramètres. On parle alors « d'états liés ». La mise en place des différents arbres de décision s'effectue à l'aide d'une procédure nommée « *tree-based clustering* » (Young, 1994). Cette dernière, qui ne sera pas détaillée ici, recherche la structure d'arbre qui maximise la vraisemblance des données d'apprentissage sachant ces états liés.

En raison de la nature de nos données visuelles, nous proposons une approche de construction des questions contextuelles qui fasse explicitement référence à la position de la langue et des lèvres. Nous adoptons pour cela la catégorisation de l'espace phonétique proposée par (Browman et Goldstein, 1990) dans le cadre de la phonologie articulatoire (catégorisation également utilisée dans (Livescu, 2005)). Les phonèmes y sont regroupés en fonction de la position et du « degré d'ouverture » supposés des lèvres, de l'apex et du dos de la langue[80]. Par exemple, la configuration articulatoire associée au phonème [sh] est alors décrite en terme de « lèvres en position quelconque, très ouvertes, apex dans la région post-alvéolaire, ouverture critique (caractéristique des fricatives), dos de la langue dans la région palatale, etc. ». Avec cette description, une question contextuelle typique sera: « Est ce que l'articulation du phonème suivant nécessite un déplacement du dos de langue dans la région palatale ? ». Dans la syntaxe adoptée par HTK, cette question s'écrira :

QS "right_context_apex_location_palatal" (+aw,*+ch,*+g,*+jh,*+k,*+ng,*+ow,*+sh,*+uh,*+w,*+z)*

Bien qu'il existe une conversion « directe » entre les deux systèmes de catégorisation, des résultats légèrement meilleurs sont obtenus en utilisant cette description en « gestes articulatoires » plutôt qu'une description plus classique, basée sur le lieu et le mode d'articulation.

Procédure d'apprentissage des MMC visuels

L'initialisation des MMC visuels est basée sur la segmentation au niveau phonétique du flux audio. Cette dernière s'obtient de façon automatique, à l'aide d'une procédure dite d' « alignement forcé ». Un MMC modélisant l'évolution supposée du contenu spectral au cours de la phrase (coefficient mel-cepstraux) est construit à partir de la transcription phonétique associée à cette phrase d'une part (obtenue à l'aide du dictionnaire CMU), et d'un ensemble de MMC (acoustiques) entraînés sur la base multi-locuteurs DARPA TIMIT (Garofolo *et al.*, 1993) d'autre part. La recherche de la séquence d'états qui maximise la vraisemblance de ce modèle sachant les observations acoustiques (effectuée à l'aide de l'algorithme de Viterbi), fournit pour chaque phrase, une segmentation du flux audio au niveau phonétique. Les données visuelles et les données audio étant acquises de façon synchrone, cette segmentation peut donc

[80] Cette catégorisation fait également intervenir la position du vélum et la caractéristique « voisée/non-voisée ».

être utilisée comme segmentation initiale du flux visuel (ultrasonore et vidéo). L'apprentissage des MMC visuels s'effectue ensuite à l'aide d'une procédure classique.

Chacun des modèles est tout d'abord considéré de façon « isolée ». Ses paramètres sont estimés à l'aide de la procédure suivante :

- Pour chacune des classes phonétiques (modélisation indépendante du contexte) :

1) Extraction des segments (visuels) correspondant
2) Initialisation du MMC correspondant (une gaussienne par état) en itérant les deux sous étapes suivantes jusqu'à ce que la vraisemblance du modèle sur l'ensemble des observations ne cesse d'augmenter de façon significative (commande *HInit* de *HTK*).
    - Association de chacune des trames de chacun des segments à un des (trois) états du MMC à l'aide de l'algorithme de *Viterbi* (initialisation par segmentation uniforme dans le cas de la première itération).
    - Mise à jour des paramètres du MMC.
3) Estimation des paramètres du MMC correspondant à l'aide de l'algorithme de *Baum-Welch* (commande *HRest* de *HTK*).

L'apprentissage des MMC visuels se poursuit ensuite par une procédure dite de « ré-estimation en modèles connectés », connue sous le terminologie anglophone de « *embedded training* » (Young, 2005). Cette procédure se compose des étapes suivantes (commande *HERest* de *HTK*):

- Pour chaque phrase :

4) Construction d'un MMC pour la phrase courante, en connectant les MMC phonétiques appropriés (procédure de construction similaire à celle utilisée pour « l'alignement forcé »).
5) Accumulation de statistiques pour tous les MMC à l'aide de l'algorithme *Forward-Backward* (étape « d'estimation »[81]).

- Mise à jour des paramètres de tous les modèles (étape de « maximisation »)

A la différence de la procédure d'apprentissage des modèles « isolés », cette approche ne tient compte que de la suite de phonèmes, et non de la segmentation temporelle. Cette propriété revêt ici une importance particulière car elle permet de remettre en cause la segmentation initiale du flux visuel. Obtenue par alignement forcé du flux audio, cette dernière ne tient pas compte de l'asynchronie entre le geste articulatoire et la réalisation acoustique. Après réitération

[81] Cette version de la procédure de *Baum-Welch*, diffère légèrement de celle utilisée dans le cadre de la ré-estimation des modèles « isolés ». Les deux versions sont entièrement décrites dans (Young, 2005).

de la procédure de « ré-estimation en modèles connectés » (une dizaine de fois en pratique[82]), un réalignement (forcé) du corpus d'apprentissage à l'aide des MMC visuels ainsi estimés est effectué. Cette procédure permet d'obtenir une nouvelle segmentation du flux visuel qui, comme l'illustre la Figure 5.3, est différente de la segmentation du flux audio. Les deux flux sont alors décrits au niveau phonétique de façon « asynchrone ».

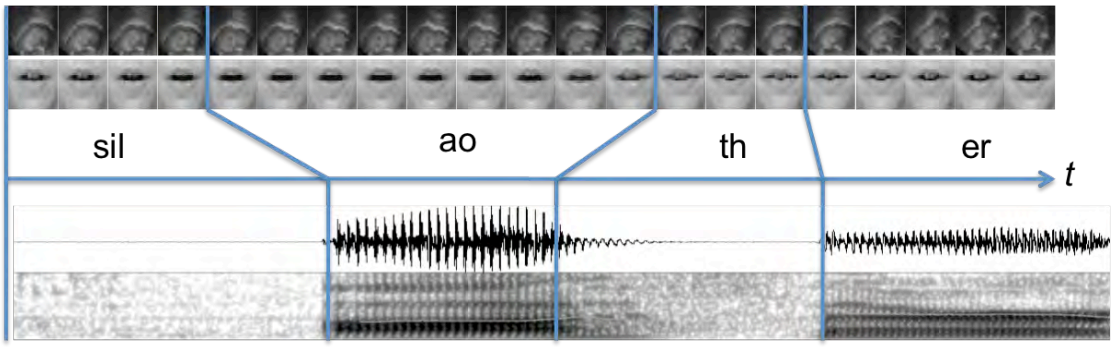


Figure 5.3 : Exemple de segmentation asynchrone des flux visuels et du flux audio au niveau phonétique (première phrase du corpus « *Author of the* … », base B1).

Les modèles « en contexte » (triphones) sont ensuite initialisés par clonage des modèles « hors contexte » précédemment estimés[83]. La procédure de « ré-estimation en modèles connectés » est itérée deux fois avant d'appliquer ensuite l'algorithme de « partage des données » (« *Tree-based clustering* »). Les modèles liés sont ensuite ré-estimés (une dizaine de fois en pratique). Enfin, le nombre de gaussiennes utilisées pour modéliser la densité de probabilité d'émission des états de chacun des modèles est augmenté de façon incrémentale. Entre chaque incrémentation, les modèles sont ré-estimés (cinq fois en pratique). Le nombre optimal de gaussiennes est fixé par validation croisée (le critère utilisé pour mesurer la performance ainsi que le partitionnement de la base de données en ensembles d'apprentissage, de validation et de test seront discutés ultérieurement). Dans le cas des modèles indépendants du contexte, le nombre optimal de gaussiennes est de l'ordre de 30. Dans le cas des modèles dépendant du contexte, ce nombre est de l'ordre de 4[84]. La procédure complète d'apprentissage des MMC visuels, dépendants et indépendants du contexte est schématisée à la Figure 5.4.

[82] Le nombre d'itérations de la procédure de ré-estimation en modèles connectés est déterminé en suivant l'évolution de la vraisemblance moyenne de l'ensemble des modèles sur les observations du corpus d'apprentissage.

[83] Il s'agit par exemple d'attribuer au MMC associé au triphone [*ao-th+er*] (phonème [th], précédé du phonème [ao] et précédant le phonème [er]), les mêmes paramètres que le MMC associé au phonème [th].

[84] On observe de plus une diminution relativement importante des performances lorsque ce nombre augmente.

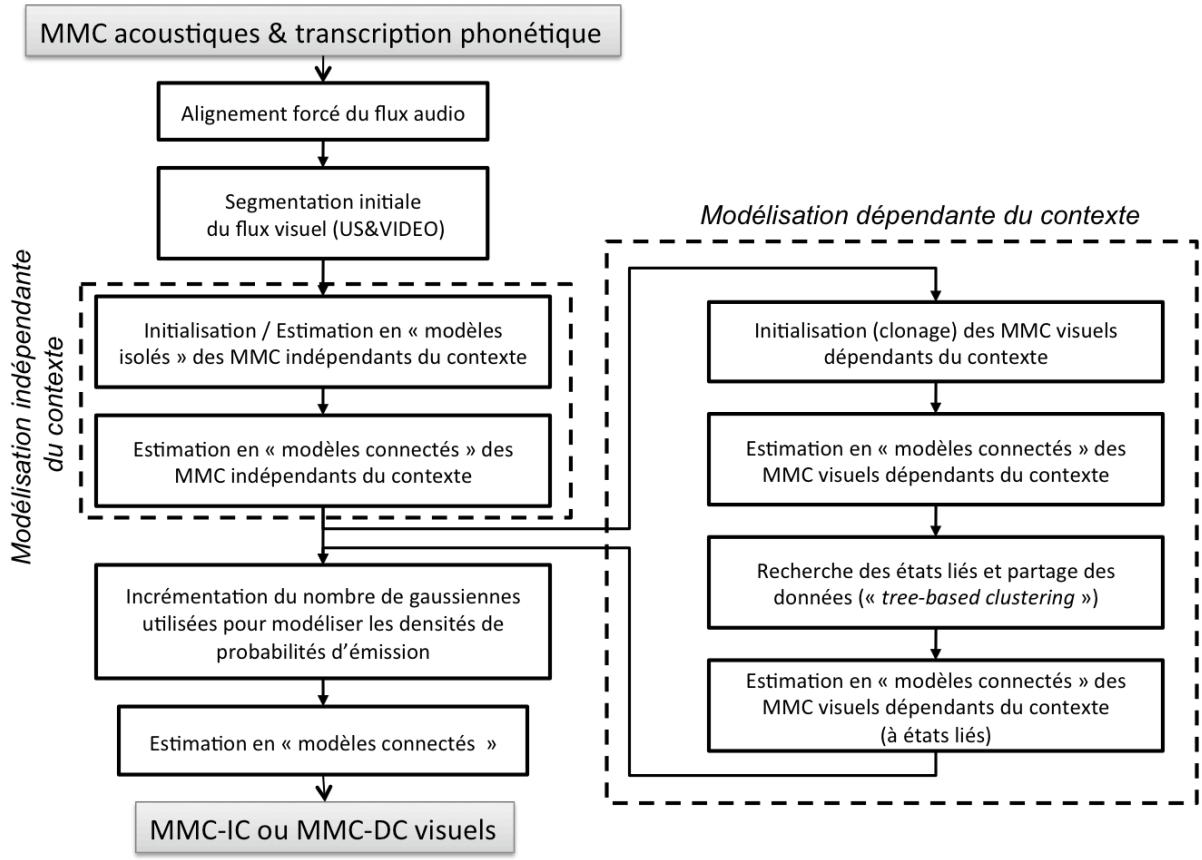


Figure 5.4 : Procédure d'apprentissage des MMC visuels indépendants et dépendants du contexte (IC/DC)

Décodage

Le décodage visuo-phonétique consiste à identifier, dans une séquence inconnue d'observations visuelles, la suite de phonèmes la plus probable, à l'aide des MMC visuels estimés dans la phase d'apprentissage précédemment décrite. C'est cette étape qui permet l'introduction éventuelle de contraintes et informations linguistiques supplémentaires. Ainsi, plusieurs scénarios de décodage sont mis en œuvre. Le premier, qualifié de « non contraint », permet l'obtention de n'importe quelle suite phonétique[85]. Ce scenario, qui n'inclut aucune contrainte linguistique, permet d'évaluer uniquement la capacité des MMC à modéliser l'évolution des caractéristiques visuelles. Le second scénario, qualifié à l'inverse de « contraint », est basé sur un décodage au niveau lexical. Les seules suites phonétiques autorisées sont celles que l'on peut obtenir en combinant les mots extraits d'un dictionnaire de 3000 éléments[86]. Ce scénario impose donc une restriction sur le vocabulaire. On notera cependant que cette restriction n'implique pas systématiquement une limitation dans l'utilisation de l'ICPS (Interface de Communication en Parole Silencieuse), notamment en présence d'homophones. Par exemple, si

[85] Les répétitions adjacentes sont néanmoins exclues (exemple, la suite [sil b ah b sil] est correcte, la suite [sil b ah ah b sil] ne l'est pas).

[86] (dont les mots qui composent les phrases du corpus CMU Arctic)

le locuteur prononce la suite de mots « *ice cream* » et que ces derniers ne font pas partie du dictionnaire, il est *a priori* possible de synthétiser un signal correct en décodant la suite de mots « *I scream* » (si ces derniers sont autorisés), qui fait l'objet d'une transcription phonétique identique. Dans le cadre de la reconnaissance de la parole par MMC, la mise en œuvre de ce type de scénario se formalise dans la définition du graphe de décodage utilisé par l'algorithme « *Token Passing* ». La structure du graphe pour chacun des scénarios est illustrée à la Figure 5.5.

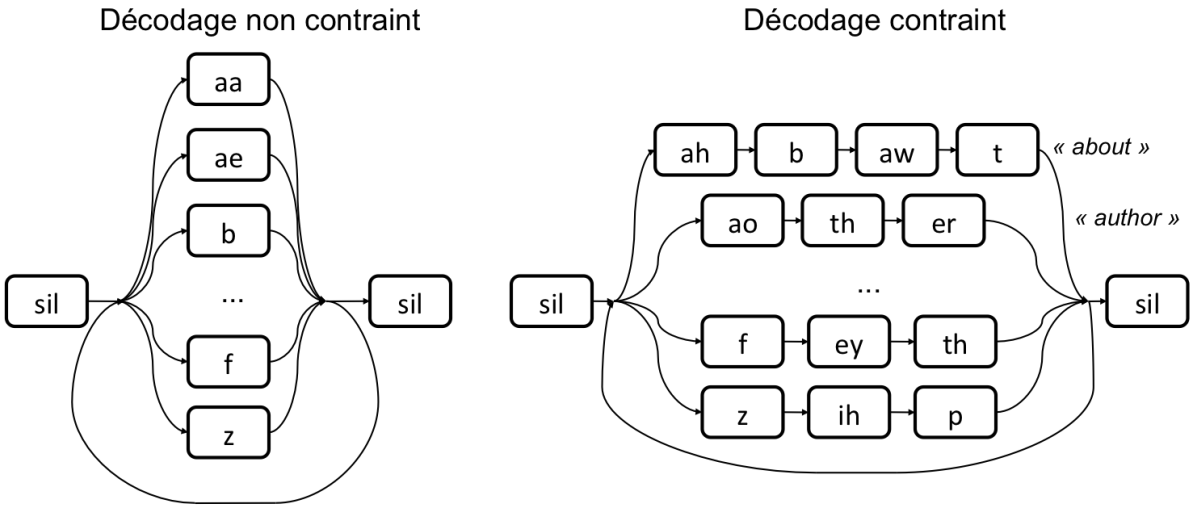


Figure 5.5 : Structure du graphe pour les deux scénarios de décodage visuo-phonétique

L'utilisation d'un modèle de langage est une autre manière, plus souple, d'introduire des informations *a priori* de nature linguistique. Dans le cas du décodage non contraint, un modèle de type bigramme est mis en œuvre au niveau phonétique. Les paramètres de ce modèle sont estimés à partir de la transcription des phrases qui forment l'ensemble d'apprentissage, soit, comme nous le verrons ultérieurement, un millier de phrases environ. Dans le cas du décodage contraint, le modèle bigramme doit se situer au niveau lexical : il doit renseigner sur la probabilité d'occurrence d'un couple de mots[87]. Un corpus d'un millier de phrases (environ 10 000 couples formés à partir d'un dictionnaire de 3000 mots) n'est pas d'une taille suffisante pour permettre une estimation fiable de ces probabilités. L'utilisation d'un corpus d'une taille beaucoup plus importante doit être envisagée. Dans le cadre de cette étude, des expérimentations préliminaires ont été réalisées à partir d'un modèle de langage entraîné sur le corpus CSR (Graff *et al.*, 1995). Cependant, ceci ne s'est accompagné d'aucune augmentation significative des performances. Ce résultat peut éventuellement s'expliquer par la différence de « genre » qui existe entre le corpus CMU Arctic et le corpus CSR, le premier étant basé sur un anglais « littéraire », le second sur un anglais « journalistique » (extraits du *Wall Street Journal* entre 1987 et 1994). De plus, le corpus CMU Arctic ayant initialement été conçu comme support pour l'enregistrement de corpus destinés à la synthèse TTS (Black et Lenzo, 2000), il

[87] Un modèle de langage fondé sur les classes de mots aurait également pu être envisagé, tout comme un modèle plus complexe de type trigramme.

contient un nombre important de phrases que l'on peut qualifier de peu intuitives (donc difficilement modélisables) comme « *Robbery, bribery, fraud.* » ou « *He had a big chimpanzee that was a winner.* ». Aussi, à ce stade de l'étude et pour des données basées sur le corpus CMU Arctic, l'utilisation d'un modèle de langage au niveau lexical, dans le cas du scénario de décodage contraint, a été abandonnée.

Enfin, un décodeur basé sur l'algorithme « *Token Passing* » possède deux hyperparamètres qui ont une influence relativement importante sur ses performances. Le premier est la « pénalité de transition inter-modèle » qui est ajoutée chaque fois qu'un MMC représentant une suite de modèles (de phonèmes ou de mots) « transite » du dernier état du modèle *n-1* au premier état du modèle *n*. Le second est le « facteur d'échelle du modèle de langage » qui pondère la vraisemblance de cette transition par sa probabilité *a priori* (fournie par le modèle de langage). Des valeurs optimales de ces deux paramètres sont obtenues par validation croisée[88].

Critères mis en œuvre pour l'évaluation

Les bases de données B1 et B2 (voir section 2.4.2) sont respectivement partitionnées en 34 et 38 listes de 30 phrases. Une liste est utilisée comme ensemble de test, deux listes comme ensemble de validation (pour l'estimation des hyperparamètres) et les listes restantes comme ensemble d'apprentissage (soit 31 listes pour la base B1 et 35 pour la base B2).

La performance du décodeur visuo-phonétique est évaluée en calculant le taux de reconnaissance phonétique[89], noté T_p, défini par :

$$T_p = \frac{N_p - D - S - I}{N_p} \qquad \text{(Équation 5.11)}$$

où N_p est le nombre de phonèmes dans l'ensemble de test, *D, S, I*, sont respectivement les nombres d'omissions, de substitutions et d'insertions, commises lors du décodage. De façon classique, la mise en correspondance de la chaine phonétique décodée avec la chaine phonétique de référence est effectuée à l'aide d'un algorithme d'alignement de chaines de caractères basé sur la programmation dynamique[90].

En modélisant les réussites (phonèmes correctement identifiés) par une loi binomiale (Montacié et Chollet, 1987), l'intervalle de confiance à 95%, noté $\Delta_{95\%}$, associé au taux de reconnaissance phonétique T_p est définie par (Hogg, 1996) :

[88] Pour information, des valeurs typiques de la (log) pénalité de transition inter-modèle dans le cas du décodage non contraint (en phonèmes) et contraint (en mots) sont respectivement -20 et -150. Le facteur d'échelle du modèle de langage se situe généralement entre 3 et 5 (il s'agit d'un facteur multiplicatif).

[89] On remarquera que la performance du décodeur est évaluée en terme de taux de reconnaissance phonétique pour les deux scénarios de décodage, y compris celui basé sur un décodage au niveau lexical.

[90] Commande HResult de HTK

$$\Delta_{95\%} = [T_p^+ - T_p^-]$$

$$\text{avec } T_p^{\pm} = \frac{T_p + t_{95\%}^2 / 2N_p \pm t_{95\%}\sqrt{T_p(1-T_p)/N_p + t_{95\%}^2/(4N_p^2)}}{1 + t_{95\%}^2 / N_p} \qquad \text{(Équation 5.12)}$$

$$\text{et } t_{95\%} = 1.96$$

Afin d'augmenter la pertinence statistique de nos expériences en réduisant le biais dû à la taille de l'ensemble de test (c'est-à-dire en augmentant N_p), une stratégie de rééchantillonage du type « *jackknife* » est adoptée (Efron, 1981). Cette procédure consiste, ici, à réaliser autant d'expériences que les bases de données contiennent de listes, et à utiliser à chaque expérience une liste de test différente (les listes restantes formant les ensembles de validation et d'apprentissage).

Par ailleurs, il est intéressant de disposer d'une estimation de la performance maximale que l'on peut espérer avec les données et les techniques de modélisation mises en œuvre. En supposant que le pouvoir de discrimination phonétique de la modalité audio est supérieur à celui de la modalité vidéo, cette « borne supérieure » peut donc être définie comme la performance d'un décodeur « acoustico-phonétique », entraîné et évalué sur les mêmes ensembles de données et à l'aide des mêmes procédures. Pour la mise en œuvre de ce décodeur, le signal acoustique est caractérisé par analyse mel-cepstrale[91] (voir section 3.4.1). Un ensemble de MMC acoustiques dépendants du contexte est ensuite entraîné sur le même ensemble d'apprentissage, à l'aide de la même procédure que celle décrite précédemment dans le cas des MMC visuels (voir Figure 5.4). Les performances de ce décodeur acoustico-phonétique sont enfin évaluées à l'aide du même critère que celui décrit précédemment : le taux de reconnaissance phonétique.

5.2.3. Résultats

Plusieurs stratégies ont été présentées dans les sections précédentes pour l'extraction des caractéristiques visuelles, la combinaison des modalités ultrasonore et vidéo et enfin pour la prise en compte du contexte. Afin de les valider, une première série d'expériences est réalisée (Figure 5.6). Les résultats sont présentés et interprétés dans les sections suivantes[92]. Afin d'évaluer uniquement la qualité de la modélisation, les performances du décodeur ne sont pour l'instant évaluées que dans le cas du scénario « non contraint », c'est-à-dire en l'absence d'informations linguistiques *a priori*.

[91] analyse par « fenêtre glissante » effectuée en décalant de 10 ms une fenêtre (de Hanning) couvrant 20 ms de signal, vecteur de caractéristiques acoustiques constitué de 13 coefficients mel-cepstraux, complétés par la valeur de leurs dérivées premières et secondes.

[92] Les expériences décrites ci-après ont fait l'objet des articles (Hueber *et al.*, 2007b; 2008a; 2009a).

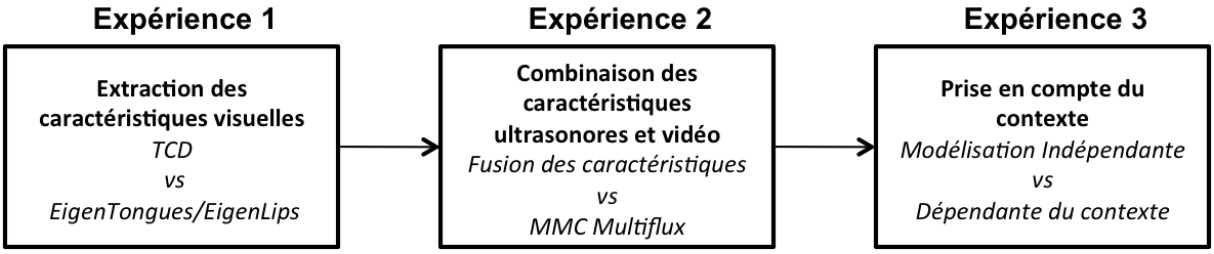


Figure 5.6 : Série d'expériences réalisée dans le but de valider les différentes stratégies proposées pour la mise en œuvre du décodeur visuo-phonétique.

Expérience 1 : Type de caractéristiques visuelles

Le Tableau 5.2 compare les performances du décodeur en fonction de l'approche utilisée pour l'extraction des caractéristiques visuelles (voir section 3.2.3). Une stratégie de « fusion au niveau des caractéristiques » est adoptée pour combiner les modalités ultrasonores et vidéo. Les classes phonétiques sont ici modélisées de façon indépendante du contexte.

	Base B1		Base B2	
	TCD	*EigenTongues/EigenLips*	TCD	*EigenTongues/EigenLips*
T_p (%)	58.5	57.7	58.3	66.9
$\Delta_{95\%}$ (%)	1.0	1.0	1.0	0.9

Tableau 5.2 : Comparaison des performances du décodage visuo-phonétique en fonction du type de caractéristiques visuelles (fusion des caractéristiques ultrasonores et vidéo, modélisation indépendante du contexte, scénario de décodage non-contraint)

Si les deux approches fournissent des résultats comparables dans le cas de la base B1, (écart non significatif), l'approche *EigenTongues/EigenLips* devance nettement l'approche par TCD sur la base B2. Ce résultat, cohérent avec les observations effectuées dans le cas de l'approche directe de la conversion visuo-acoustique, se retrouve également dans les expériences décrites ultérieurement (MMC multi-flux, prise en compte du contexte). Aussi, nous ne présenterons par la suite que les résultats basés sur l'approche *EigenTongues/EigenLips*.

Expérience 2 : Combinaison des modalités ultrasonore et vidéo

Le Tableau 5.3 compare la stratégie de « fusion au niveau de caractéristiques » à celle basée sur l'utilisation de MMC multi-flux (« *Early Integration* »). Les classes phonétiques sont ici modélisées de façon indépendante du contexte.

	Base B1		Base B2	
	Fusion	MMC multi-flux	Fusion	MMC multi-flux
T_p (%)	57.7	59.5	66.9	67.5
$\Delta_{95\%}$ (%)	1.0	1.0	0.9	0.9

Tableau 5.3 : Comparaison des performances du décodage visuo-phonétique en fonction de la stratégie adoptée pour combiner les modalités ultrasonore et vidéo (caractéristiques visuelles du type *EigenTongues/EigenLips*, modélisation indépendante du contexte, scénario de décodage non-contraint)

L'utilisation de MMC multi-flux est à l'origine d'une amélioration des performances. Cette dernière n'est cependant statistiquement significative que dans le cas de la base B1. Rappelons que les valeurs optimales des poids associés à chacune des deux modalités visuelles sont déterminées par validation croisée (minimisation du taux de reconnaissance phonétique sur les deux listes de validation). En moyenne ces valeurs sont de l'ordre de 0.7 pour la modalité ultrasonore et de 0.3 pour la modalité vidéo. Un décodeur basé sur des MMC multi-flux accorde donc, de fait, plus d'importance à la modalité ultrasonore. Ceci semble cohérent avec le fait que, pour l'articulation de nombreux phonèmes, la position de la langue est plus décisive que celle des lèvres. Dans la suite des expériences, la stratégie de fusion des caractéristiques est abandonnée au profit de l'approche par MMC multi-flux.

Expérience 3 : Intérêt de la modélisation dépendante du contexte

Le Tableau 5.4 compare les performances obtenues dans le cadre de modélisations indépendantes et dépendantes du contexte (triphone).

	Base B1		**Base B2**	
	CI	**CD**	**CI**	**CD**
T_p (%)	59.5	65.6	67.5	70.8
$\Delta_{95\%}$ (%)	1.0	0.9	0.9	0.9

Tableau 5.4 : Comparaison des performances du décodage visuo-phonétique en fonction de la prise en compte du contexte dans la modélisation des classes phonétiques (scénario de décodage non-contraint).

La modélisation de classes phonétiques « en contexte » permet donc d'améliorer de façon significative les performances du décodeur (environ 6% d'amélioration absolue). Ce résultat était attendu car, comme nous l'avons précédemment souligné, les phénomènes de co-articulation et d'anticipation gestuelle sont ici mieux pris en compte. Néanmoins, cette amélioration est atteinte au prix d'une modélisation plus complexe, le nombre de paramètres à estimer étant *a priori* plus important que dans le cas d'une modélisation hors contexte. La procédure de partage des données d'apprentissage entre les états des MMC contextuels permet cependant de réduire sensiblement ce nombre. Dans notre cas, le nombre d'états dont il faut réellement estimer les paramètres est réduit à 6 % du nombre d'états total (moyenne sur l'ensemble des phonèmes). Néanmoins, avec 8885 triphones, soit autant de MMC à trois états dont les probabilités d'émission sont modélisées par un mélange de 4 gaussiennes, le nombre de paramètres à estimer est de,

$$\underset{\text{paramètres}}{2\ 305\ 700} = \underset{\substack{\text{partage des paramètres}\\ \text{entre états liés}}}{0.06} \times (\underset{\text{modèles}}{8885} \times \underset{\text{états}}{3} \times \underset{\text{gaussiennes}}{4} \times (\underset{\text{moyennes}}{180} + \underset{\text{variances}}{180}) + \underset{\text{modèles}}{8885} \times \underset{\text{transitions}}{5})$$

(Équation 5.13)

soit environ deux fois plus qu'en modélisation hors contexte (voir l'équation 5.9).

Décodage visuo-phonétique *versus* décodage acoustico-phonétique

Les expériences précédentes ont permis de déterminer les meilleures stratégies à adopter pour l'apprentissage des modèles visuels (caractéristiques visuelles du type *EigenTongues/EigenLips*, modélisation basée sur des MMC multi-flux, modélisation dépendante du contexte). Le Tableau 5.5 compare à présent la performance du décodeur visuo-phonétique à une « borne supérieure », c'est-à-dire à la performance d'un décodeur acoustico-phonétique entraîné et évalué sur les mêmes données et à l'aide des mêmes procédures.

	Base B1		Base B2	
	Visuel	**Acoustique**	**Visuel**	**Acoustique**
T_p (%)	65.6	82.4	70.8	81.6
$\Delta_{95\%}$ (%)	0.9	0.8	0.9	0.8

Tableau 5.5 : Décodage visuo-phonétique versus décodage acoustico-phonétique (modélisation par MMC-multiflux (visuel), dépendants du contexte (visuel et acoustique), scénario de décodage non contraint)

Ainsi, la performance d'un décodeur phonétique basé sur les modalités visuelles est de l'ordre de 80 % de celle d'un décodeur basé sur la modalité audio. Par ailleurs, pour cette expérience comme pour les précédentes, et de façon similaire à ce qui a été observé dans le cas de la conversion visuo-acoustique « directe », les performances observées sont systématiquement meilleures sur la base B2 que sur la base B1. Il est difficile de déterminer avec certitude l'origine de ces écarts de performance car ces corpus sont basés sur des locutrices différentes. Cependant, les performances du décodage à partir de la modalité audio étant quasiment identiques pour les deux bases, ces écarts ne proviennent *a priori* pas de la quantité de données disponibles pour l'apprentissage (légèrement supérieure pour la base B2). On peut donc les expliquer par la différence de qualité des données visuelles, entre les deux bases (les données de la base B2 présentant une meilleure résolution spatiale et temporelle que celles de la base B1).

Contributions respectives des modalités ultrasonore et vidéo

L'expérience décrite ci-après vise à comparer les quantités d'informations véhiculées par les deux modalités visuelles. Le Tableau 5.6 présente les performances de deux décodeurs visuo-phonétiques, le premier n'exploitant que les données ultrasonores, le second, que les données vidéo. Ces expérimentations sont réalisées sur la base B2 (qui présente les données vidéo avec la meilleure résolution spatiale et temporelle).

	Ultrason et vidéo	**Ultrason**	**Video**
T_p (%)	70.8	62.2	43.6
$\Delta_{95\%}$ (%)	0.9	1.0	1.0

Tableau 5.6 : Comparaison des performances du décodeur phonétique dans le cas où une seule des deux modalités visuelles n'est utilisée (base B2, approche *EigenTongues/EigenLips,* modélisation dépendante du contexte, scénario de décodage non contraint).

De façon attendue, la modalité ultrasonore apporte plus d'informations sur la production que la modalité vidéo.

Scénarios de décodage

L'introduction d'une étape préalable de décodage phonétique dans le processus de conversion visuo-acoustique est motivée par la volonté d'introduire des informations linguistiques *a priori*. Dans le cas du scénario de décodage dit non contraint, ces informations sont fournies par le modèle de langage (ML) au niveau phonétique (bigramme) dont la mise en œuvre a été décrite précédemment. Dans le cas du scénario de décodage « contraint », cet apport d'information prend la forme d'une limitation sur le vocabulaire et le décodage est réalisé au niveau lexical (à l'aide d'un dictionnaire de 3000 mots contenant notamment les mots du corpus CMU Arctic). Les performances obtenues dans le cadre de ces différents scénarios sont présentées dans le Tableau 5.7 (où figurent également les nombres d'omissions *D*, d'insertions *I*, et de substitutions *S*).

Base	**B1**			**B2**		
Scénario	**Libre**	**Libre+ML phonétique**	**Contraint**	**Libre**	**Libre+ML phonétique**	**Contraint**
T_p (%)	65.6	66.2	**74.7**	70.8	71.4	**83.3**
$\Delta_{95\%}$ (%)	0.9	1.0	0.8	0.9	0.9	0.8
D	4294	3397	3964	4412	4543	3363
S	6279	6377	3613	5289	5103	2389
I	1397	1987	1232	1393	1227	590
N	34693			37970		

Tableau 5.7 : Comparaison des performances du décodeur visuo-phonétique en fonction du scénario de décodage mis en œuvre (caractéristiques visuelles du type *EigenTongues/EigenLips*, modélisation dépendante du contexte par MMC multi-flux).

Dans le cas du scénario non contraint, l'apport du modèle de langage apparaît comme très faible (amélioration de la performance à peine supérieure à la largeur de l'intervalle de confiance). Le modèle bigramme au niveau phonétique est donc ici peu informatif. Rappelons que ce dernier est entraîné (et évalué) sur le corpus CMU Arctic qui, construit de sorte à couvrir l'espace de diphones, peut faire apparaître des séquences phonétiques « exotiques » difficilement modélisables.

De façon attendue, la restriction du vocabulaire (scénario contraint) est à l'origine d'une nette amélioration des performances (de l'ordre de 10 % pour les deux bases de données). Dans les deux bases, le nombre de substitutions est notamment réduit de plus de 40 % par rapport au scénario non contraint. Pour analyser en détail la nature des erreurs (restantes), les matrices de confusions associées au décodage des bases B1 et B2, dans le cas du scénario contraint, sont présentées à la Figure 5.7. De façon prévisible, les phonèmes qui ne diffèrent que par leurs caractéristiques de voisement et de nasalité (sauf en cas de contact linguovélaire) sont ceux qui sont le plus souvent confondus. Ainsi, les principales erreurs de substitution s'effectuent au sein

des groupes ([p], [b], [m]), ([t], [d], [n]), ([f], [v]), ([s],[z]), ([k], [g]), ([ch], [jh]), ([th], [dh]). Le manque d'informations sur la position de l'apex peut par ailleurs expliquer les erreurs commises au sein du groupe ([t], [d], [s], [z], [sh]). Des erreurs sont également commises sur les voyelles. Le phénomène de réduction est très certainement à l'origine de la substitution de nombreuses d'entres elles par la voyelle « centrale » [ah]. Les diphtongues pour lesquelles le point d'articulation varie constamment au cours de la réalisation grâce notamment à un mouvement continu de la langue (glissement ou *glide* en anglais), sont parfois décodées comme une suite de deux voyelles « stables ». Ceci peut expliquer les substitutions au sein des groupes ([ey], [ah]), ([oy], [iy]) et ([ow], [ao]). Enfin, certaines voyelles très proches sont également confondues, comme [uh] et [uw] d'une part et [iy] et [ih] d'autre part.

Le scénario de décodage contraint donne l'accès à un niveau de description linguistique encore supérieur, le niveau lexical. Comme nous le verrons ultérieurement, ce niveau de description sera utilisé dans le cadre de la synthèse du signal de parole, pour la génération du contenu prosodique « cible ». Aussi, il apparaît nécessaire d'évaluer la performance du décodeur, non plus uniquement en terme de taux de reconnaissance phonétique, mais en terme de taux de reconnaissance « en mots ». Ce dernier, noté ici T_m est obtenu de façon classique, à l'aide de l'équation 5.11, en ne considérant plus le niveau phonétique, mais le niveau lexical[93]. La valeur de ce taux sur chacune des deux bases de données est présentée au Tableau 5.8.

	Base B1	**Base B2**
T_m (%)	48.7	61.6
$\Delta_{95\%}$ (%)	2.1	2.0

Tableau 5.8 : Taux de reconnaissance « en mots » (caractéristiques visuelles du type *EigenTongues/EigenLips*, modélisation dépendante du contexte par MMC multi-flux, dictionnaire de 3000 mots, aucun modèle de langage n'est ici utilisé).

De façon attendue, ce dernier reste relativement faible en l'absence de modèle de langage. Néanmoins, avec un taux de reconnaissance phonétique de l'ordre de 75 % pour la base B1 et de plus de 80 % pour la base B2, le scénario de décodage contraint est le scénario privilégié pour la seconde étape de l'approche indirecte de la conversion visuo-acoustique : la synthèse du signal de parole.

[93] T_m est nommé *Accuracy* dans HTK. En notant WER le « *Word Error Rate* », on obtient $T_m = 1 - WER$.

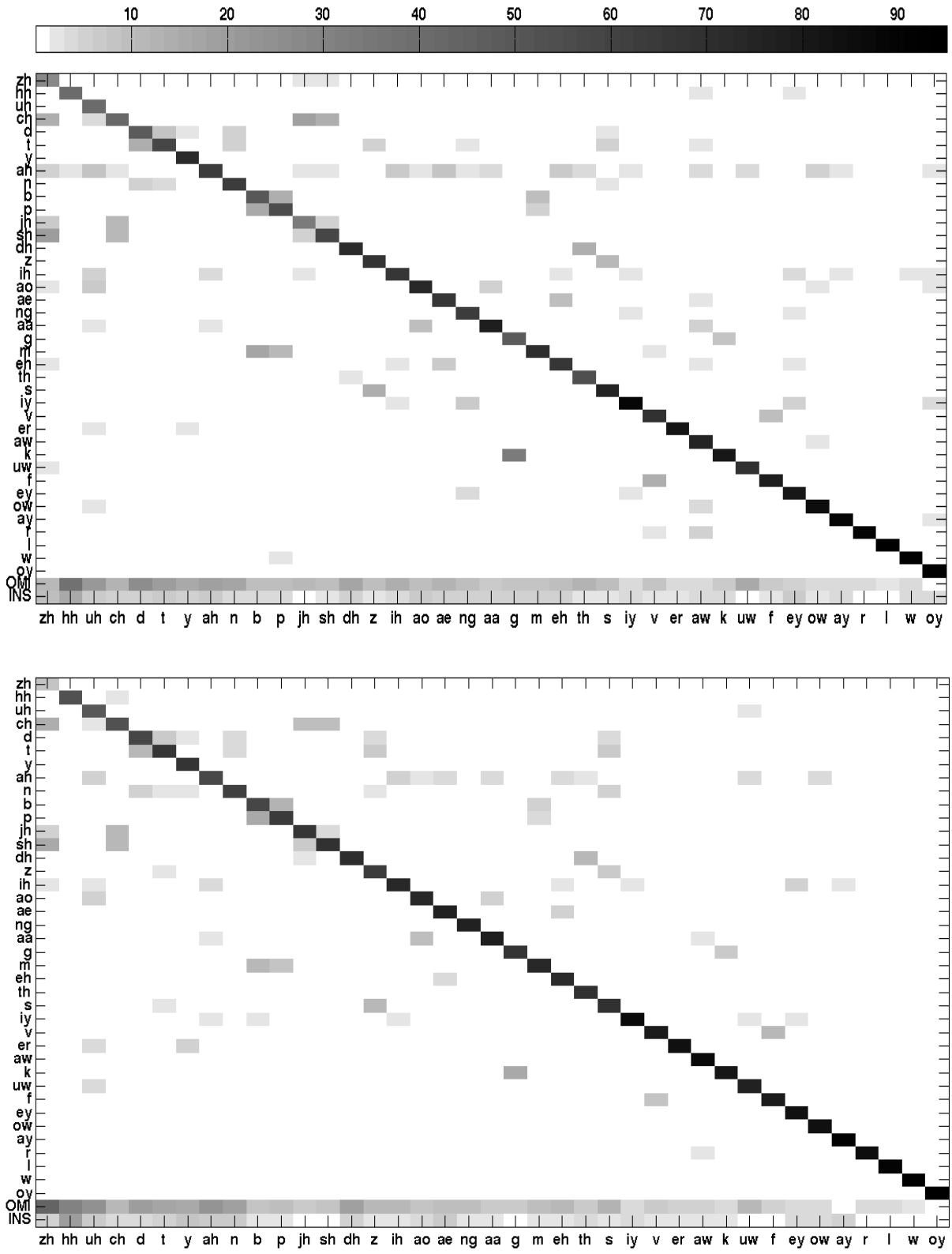


Figure 5.7 : Matrices de confusion du décodage visuo-phonétique dans le cas du scénario contraint (en haut, base B1, en bas, base B2, en abscisse, la référence, en ordonnée, la prédiction, OMI : omission, INS : insertion).

5.3. Synthèse du signal de parole

Un schéma général de la procédure mise en œuvre pour la synthèse du signal de parole à partir du flux de caractéristiques visuelles et de la suite phonétique décodée est présenté à la Figure 5.8. Les différentes étapes de cette procédure sont décrites dans les sections suivantes.

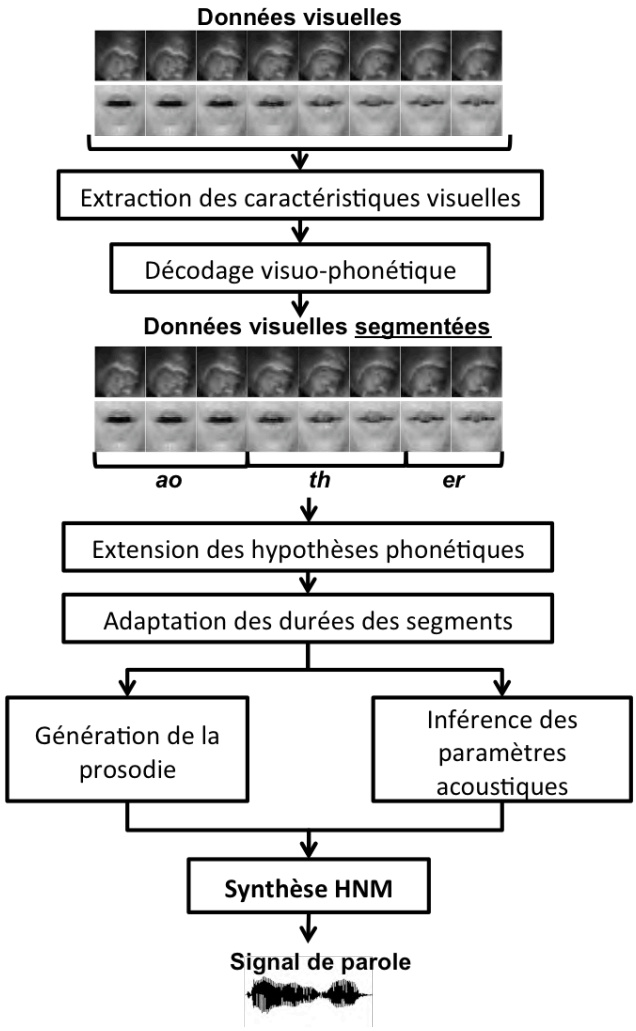


Figure 5.8 : Synthèse du signal de parole à partir du flux visuel décodé au niveau phonétique

5.3.1. Extension des hypothèses phonétiques

Dans le cadre de l'approche indirecte de la conversion visuo-acoustique, l'étape de décodage phonétique est introduite en amont du processus de synthèse. Cette architecture, dont les avantages ont été évoqués précédemment (intégration d'informations supplémentaires de nature linguistique), présente néanmoins un inconvénient important ; la qualité du signal de synthèse est *a priori* entièrement dépendante de la performance du décodeur phonétique. En d'autres termes, une erreur dans la suite phonétique décodée se retrouvera irrémédiablement dans le signal de synthèse. Or, d'après les expériences présentées à la section précédente, entre 20 % et 30 % des phonèmes décodés restent erronés. Il apparaît donc nécessaire d'introduire, dans la procédure de synthèse, un mécanisme permettant de remettre en cause certaines des hypothèses

du décodeur phonétique. La procédure que nous adoptons ici consiste à remplacer la suite phonétique décodée par un « treillis de possibilités ». Il s'agit de disposer, pour un même segment, de plusieurs alternatives phonétiques. Pour ce faire, deux approches ont été envisagées. La première consiste à adopter, lors du décodage, une stratégie connue sous le nom de « *N-best* ». Cette dernière consiste à conserver, lors du parcours du graphe de décodage (par l'algorithme « *Token Passing* »), non pas « le meilleur » mais les « *N* meilleurs » chemins (au sens du maximum de vraisemblance). Cependant, les *N* suites phonétiques obtenues à l'aide de cette stratégie ne présentent en pratique que de faibles différences[94]. Une seconde approche est alors mise en œuvre. Cette dernière est une approche « par règles », qui considère, comme alternative possible au phonème décodé, les phonèmes jugés *a priori* similaires (ou très proches) du point de vue de la position de la langue et des lèvres. Ces « sosies labiaux et linguaux » sont listés dans le Tableau 5.9.

[p], [b], [m]	[k], [g], [ng]
[t], [d], [n]	[ch], [jh]
[f], [v]	[th], [dh]
[s], [z]	[uh], [uw]
[sh], [zh]	[ih], [iy]

Tableau 5.9 : Regroupement des phonèmes en « sosies labiaux et linguaux » pour la création du treillis phonétique

La Figure 5.9 illustre cette seconde approche sur un exemple.

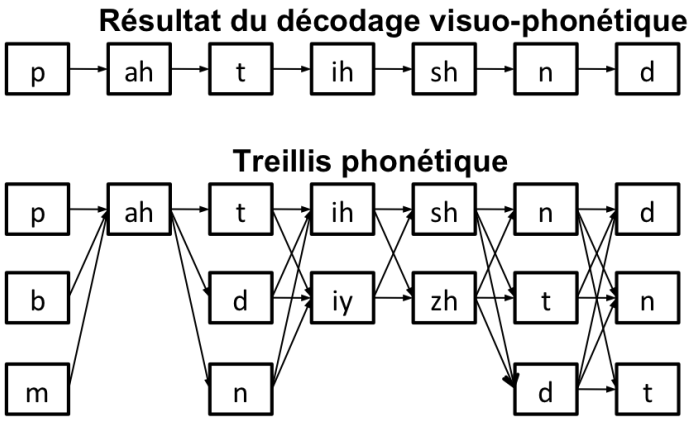


Figure 5.9 : Exemple d'un treillis phonétique généré à l'aide de l'approche « par règles », basée sur l'utilisation des regroupements en sosies labiaux et linguaux (décodage initial du mot « *petitioned* »)

[94] Par exmple, les vingt meilleures hypothèses ne diffèrent que d'un ou deux phonèmes seulement. Pour obtenir de véritables alternatives, il faudrait conserver, lors du parcours du graphe de décodage, un nombre beaucoup plus important de chemins sous-optimaux, ce qui est difficile à implémenter en pratique.

5.3.2. Adaptation de la durée des segments

Afin de tenir compte de la désynchronisation entre le geste articulatoire et la réalisation acoustique, les flux visuels et audio ont été décrits, au niveau phonétique, de façon « asynchrone » (voir Figure 5.3). Le « rythme articulatoire » étant différent du « rythme acoustique », la segmentation temporelle d'un flux visuel obtenue lors du décodage visuo-phonétique ne peut *a priori* pas être utilisée directement pour la synthèse du signal audio. Pour pouvoir la corriger, il est nécessaire de modéliser (ou tout du moins de quantifier) cette asynchronie visuo-acoustique.

Dans (Bailly *et al.*, 2008) et (Govokhina, 2008), une procédure pour l'apprentissage et l'utilisation d'un « modèle de décalage » (ou « *phasing model* ») entre les limites temporelles du geste articulatoire est proposée dans le cadre de l'animation des visages parlants artificiels (approche *PHMM* pour « *Phased Hidden Markov Model* »)[95]. L'objectif est d'augmenter le réalisme de ces visages parlants en générant des trajectoires articulatoires et acoustiques qui respectent l'asynchronie audiovisuelle naturelle. Dans le cadre de cette étude, aucune contrainte de synchronisation entre le signal audio synthétisé et la séquence visuelle de test n'étant fixée, une approche légèrement différente est adoptée.

En phase d'apprentissage, la durée moyenne des occurrences de chacun des phonèmes est calculée dans le flux visuel et dans le flux acoustique. Les résultats de cette estimation sont présentés à la Figure 5.10.

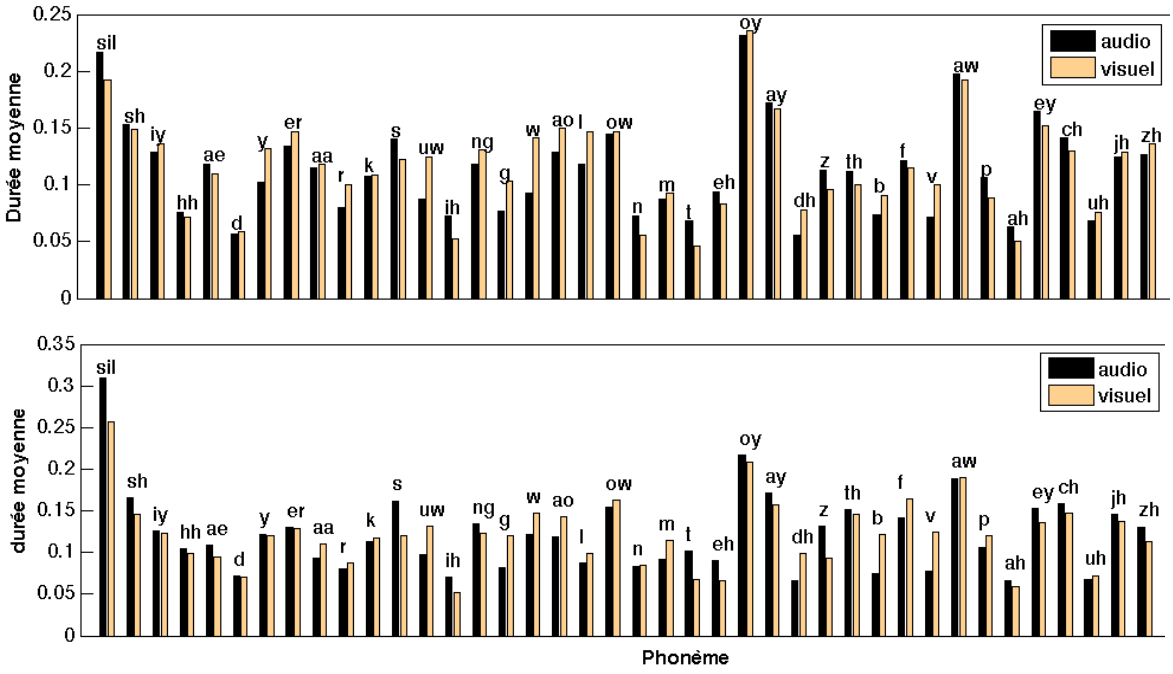


Figure 5.10 : Durée moyenne (en secondes) des phonèmes dans le flux audio et dans le flux visuel (en haut la base B1, en bas, la base B2)

[95] On consultera notamment le chapitre 5 de (Govokhina, 2008) (page 87 à 91).

Pour chaque classe phonétique, un « taux de déformation temporelle » est ensuite obtenu en calculant le rapport de ces durées moyennes (dans la Figure 5.10, cela revient, pour chaque classe phonétique, à diviser la hauteur de la barre jaune par la hauteur de la barre noire). En phase de test, la longueur de chaque phonème identifié dans le flux visuel est alors multipliée par le taux de déformation temporelle correspondant. Le segment décodé est alors soit compressé (taux < 1), soit dilaté (taux > 1) ; une nouvelle décomposition temporelle est ainsi générée.

5.3.3. Inférence des paramètres acoustiques – Approche par sélection d'unités

Dans le cadre de l'approche indirecte de la conversion visuo-acoustique, la génération du signal acoustique, à partir du flux de données visuelles d'une part et du treillis d'hypothèses phonétiques d'autre part, s'appuie sur la technique d'analyse-synthèse « Harmonique plus Bruit » ou HNM, décrite au chapitre 3 (section 3.4.2). La première technique proposée dans cette étude pour l'inférence des paramètres HNM, est basée sur approche « concaténative » (ou approche « par sélection d'unités »)[96].

La base de données d'apprentissage étiquetée au niveau phonétique peut être perçue comme un dictionnaire « d'unités multimodales » où chaque unité associe une réalisation acoustique à une réalisation articulatoire[97]. L'approche proposée consiste à chercher, dans ce dictionnaire, la suite d'unités qui maximise la similarité avec la séquence de test dans le domaine visuel, tout en minimisant les discontinuités acoustiques dans le signal qui sera synthétisé. Pour déterminer cette suite, nous proposons une procédure inspirée de celle proposée par (Hunt et Black, 1996) pour la synthèse de la parole à partir du texte (système « *Text-To-Speech* » ou TTS).

L'unité linguistique privilégiée ici pour la synthèse du signal (donc pour la construction du dictionnaire) est le diphone. Pour le flux visuel, un « diphone » est simplement défini ici comme la réunion de deux phones adjacents[98]. Soit $\mathbf{o} = [\mathbf{o}_1,...,\mathbf{o}_L]$ la séquence visuelle de test constituée de L observations, $\tau = [\tau_1,...,\tau_T]$ (avec $\tau_1 = 1$ et $\tau_T = L$) la décomposition temporelle de $\mathbf{o}$ au niveau phonétique fournie par le décodeur visuo-phonétique et T le nombre de phonèmes identifiés, la suite T_τ des T diphones visuels « cibles » est définie par :

$$T_\tau = \{T_1, \ ..., \ T_T\}$$
$$\text{avec } T_n = \left[\mathbf{o}_{\tau_n}, \ ..., \ \mathbf{o}_{\tau_{n+2}}\right], \ n \in [1 \ .. \ T] \qquad \text{(Équation 5.14)}$$

En revanche, pour la composante acoustique des unités du dictionnaire, un « véritable » diphone est constitué en ne conservant que la partie « centrale » d'un regroupement de deux

[96] Cette technique a notamment fait l'objet des articles (Hueber *et al.*, 2007c; Hueber *et al.*, 2008b).

[97] Notons qu'en raison du caractère asynchrone de la description au niveau phonétique du flux visuel et du flux acoustique, ces réalisations peuvent être de longueurs différentes.

[98] Il ne s'agit donc pas véritablement d'un « diphone », mais plutôt d'un « biphone ».

phones adjacents. Cette dernière est déterminée en recherchant, pour les phones « gauche » et « droite », l'instant où la stabilité spectrale est maximale[99]. Chaque diphone acoustique est représenté dans le dictionnaire par sa séquence de paramètres HNM.

La procédure de sélection d'unités consiste à rechercher, dans le dictionnaire, la séquence de diphones $U = \{U_1, ..., U_T\}$ qui maximise la ressemblance avec la séquence de test, tout en limitant les distorsions introduites en concaténant les composantes « acoustiques » de *U*. En considérant le dictionnaire comme un graphe où chaque nœud est occupé par un diphone, cette séquence optimale d'unités correspond au chemin de coût minimum. Ce dernier est déterminé à l'aide de l'algorithme de *Viterbi* (Forney, 1973). Le coût affecté au choix d'une unité s'exprime comme la somme pondérée de deux coûts, un « coût de cible » noté C^t, évalué dans notre cas dans le domaine visuel, et un « coût de concaténation », noté C^c, évalué (de façon plus classique) dans le domaine acoustique[100].

Le coût de cible mesure la similarité entre le diphone cible identifié dans la séquence visuelle de test, et un des diphones du dictionnaire. Les deux unités (notées ici T_k et U_k) étant *a priori* de longueurs différentes, le coût de cible est défini par :

$$C^t(U_k, T_k) = DTW(U_k, T_k) \qquad \text{(Équation 5.15)}$$

où $DTW(a,b)$ est la distance cumulée sur le chemin d'alignement (non linéaire) des séquences *a* et *b*, obtenu par « déformation temporelle dynamique », ou « *Dynamic Time Warping* »[101]. Le coût de concaténation mesure la discontinuité spectrale due à la concaténation des composantes acoustiques des diphones U_k et U_{k+1}. Il est défini par :

$$C^c(U_k, U_{k+1}) = D\left(MFCC\left(U_k^{END}\right), MFCC\left(U_{k+1}^1\right)\right) \qquad \text{(Équation 5.16)}$$

où D est la distance euclidienne et $MFCC(U_k^l)$ le vecteur de coefficients mel-cepstraux de l'unité U_k à l'instant *l* (l'instant noté END correspond à la fin de l'unité). Le coût de concaténation entre deux unités « naturellement adjacentes » dans le corpus est défini comme nul.

Cette procédure de sélection d'unités est résumée à la Figure 5.11. La procédure de synthèse du signal à partir des unités sélectionnées sera décrite ultérieurement.

[99] Cet instant est défini comme celui où la norme (euclidienne) du vecteur de dérivées des coefficients mel-cepstraux est minimale.

[100] Les poids associés au coût de cible et au coût de concaténation sont déterminés manuellement.

[101] L'algorithme DTW, très utilisé en traitement de la parole, est décrit en détail dans (Boite *et al.*, 2000).

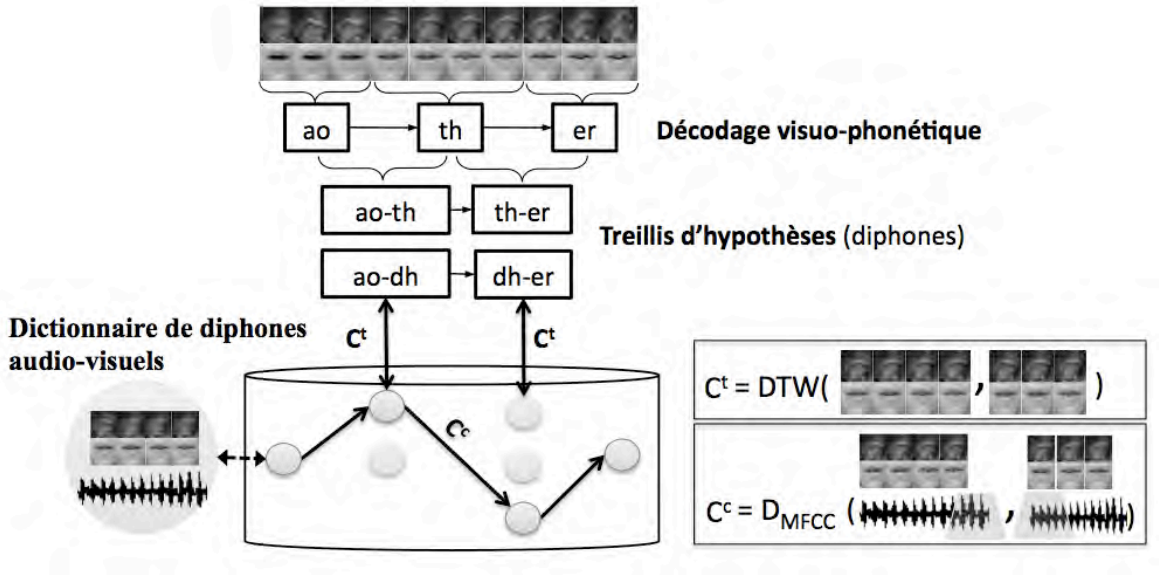


Figure 5.11 : Inférence des paramètres acoustiques – Approche par sélection d'unités

5.3.4. Inférence des paramètres acoustiques – Approche stochastique

La seconde approche proposée pour l'inférence des paramètres HNM à partir du treillis d'hypothèses phonétiques est basée sur l'utilisation de la technique dite de « synthèse par modèles de Markov Cachés » ou « synthèse par modèles stochastiques ».

L'utilisation de MMC pour l'inférence de paramètres acoustiques a été initialement proposée par (Donovan, 1996). Cette technique s'est ensuite considérablement développée dans les années 2000, notamment sous l'impulsion du groupe de travail HTS[102] (Tokuda *et al.*, 2000). La synthèse par MMC est une technique toujours à ce jour en pleine évolution, qui fait l'objet d'une abondante littérature. Aussi, nous n'en présenterons que les principes généraux.

Principe

On note $\mathbf{m}$ le MMC représentant la suite de phonèmes dont on veut générer une réalisation acoustique (construit par concaténation de MMC acoustiques au niveau phonétique). En définissant la séquence d'observation recherchée $\hat{\mathbf{x}}$ (c'est-à-dire celle que l'on souhaite générer) comme celle qui maximise la vraisemblance du modèle $\mathbf{m}$ parmi toutes les séquences d'observations $\mathbf{x}$ possibles, le problème de l'inférence peut se formaliser sous la forme suivante :

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{m}) \qquad \text{(Équation 5.17)}$$

[102] pour « *HMM-based Speech Synthesis* ». Le groupe de travail HTS développe la boîte à outils du même nom, qui est une adaptation (un *patch*) de HTK, disponible sur http://hts.sp.nitech.ac.jp/. Les expérimentations décrites ci-après utilisent cette boîte à outils.

En notant Q la séquence d'états que doit suivre le modèle $\mathbf{m}$ pour générer une séquence d'observations $\mathbf{x}$, et d'après l'équation 5.5 (section 5.2.1), cette équation devient :

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} \sum_{Q} p(\mathbf{x}, Q \mid \mathbf{m}) \\ &= \arg\max_{\mathbf{x}} \sum_{Q} p(\mathbf{x} \mid Q, \mathbf{m}) P(Q \mid \mathbf{m})\end{aligned} \qquad \text{(Équation 5.18)}$$

De façon similaire à ce que nous avons présenté précédemment (section 5.2.1), cette équation peut être résolue de façon sous-optimale (voir l'équation 5.6) en ne considérant que la séquence d'états la plus probable $\hat{Q}$, tel que :

$$\begin{aligned}&\hat{Q} = \max_{Q} P(Q \mid \mathbf{m}) \\ &\text{d'où } \hat{\mathbf{x}} \approx \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{m}, \hat{Q})\end{aligned} \qquad \text{(Équation 5.19)}$$

La séquence $\hat{Q}$ détermine la durée pendant laquelle le modèle reste dans un état particulier et par conséquent, la durée des phones synthétisés. Soit D_i la variable aléatoire associée à la durée de l'état i du modèle $\mathbf{m}$, P_i la probabilité de cette variable, d_i une de ces réalisations, et K le nombre d'états de $\mathbf{m}$; la probabilité que le modèle évolue suivant la séquence d'état Q s'écrit :

$$P(Q \mid \mathbf{m}) = \prod_{i=1}^{K} P_i(D_i = d_i) \qquad \text{(Équation 5.20)}$$

Dans le cas d'un MMC, la loi de probabilité des durées des états est une loi géométrique :

$$P_i(D_i = d_i) = a_{ii}^{d_i - 1}(1 - a_{ii}) \qquad \text{(Équation 5.21)}$$

avec a_{ii} la probabilité « de transition » de l'état *i* vers l'état *i*. Cependant, ce type de loi ne permet pas de modéliser correctement la structure temporelle d'un phone (Pylkkönen et Kurimo, 2004). Si la modélisation des durées ne semble pas être un point crucial en reconnaissance de la parole, elle constitue en revanche un enjeu majeur en synthèse. Aussi, les modèles de Markov Cachés sont remplacés par des modèles dit « semi-Markoviens » (HSMM pour « *Hidden semi-Markov Model* »[103]) pour lesquels il est possible de spécifier explicitement le type de loi de probabilité utilisée pour modéliser la durée des états. Dans (Yoshimura *et al.*, 1998) (et dans HTS), une loi normale est utilisée, telle que :

$$P_i(D_i = d_i) = N(d_i; \mu_i; \sigma_i) \qquad \text{(Équation 5.22)}$$

[103] On notera que l'algorithme de *Baum-Welch* (section 5.2.1), ne peut pas être utilisé pour l'estimation des paramètres d'un modèle semi-markovien. D'autres algorithmes doivent alors être mis en œuvre. Ces derniers ne seront pas détaillés ici. On notera simplement qu'ils sont moins optimaux que la procédure de *Baum Welch*.

Une fois la séquence d'état $\hat{Q}$ déterminée, la génération de la séquence d'observation $\hat{\mathbf{x}}$ peut alors être envisagée. Lorsque la densité de probabilité d'émission d'un état du modèle $\mathbf{m}$ est modélisée par une loi normale[104], l'observation qui maximise la vraisemblance pour cet état est simplement un vecteur égal à l'espérance mathématique de cette loi. Comme l'illustre la Figure 5.12, la séquence d'observations qui maximise la vraisemblance du modèle évoluant suivant la séquence d'état $\hat{Q}$ est donc une fonction constante par morceaux. Présentant une discontinuité à chaque changement d'état, une telle séquence n'est pas acceptable. Pour obtenir des trajectoires continues, Tokuda propose un nouvel algorithme d'inférence basé sur le respect de la dynamique des paramètres (Tokuda *et al.*, 2000). Il introduit ainsi le concept de « *Trajectory HMM* ».

Cette approche suppose qu'une observation $\mathbf{x}_t$ de dimension M est composé d'un vecteur de paramètres statiques, noté $\mathbf{c}_t$, et de deux vecteurs de paramètres dynamiques, notés $\Delta^{(1)}\mathbf{c}_t$ et $\Delta^{(2)}\mathbf{c}_t$ respectivement vecteurs de dérivées premières et secondes de $\mathbf{c}_t$, tels que:

$$\begin{aligned}
&\mathbf{x}_t = [\mathbf{c}_t^T, \Delta^{(1)}\mathbf{c}_t^T, \Delta^{(2)}\mathbf{c}_t^T]^T \\
&\mathbf{c}_t = [c_t(1), c_t(2), ..., c_t(M)]^T \\
&\Delta^{(d)}\mathbf{c}_t = \sum_{\tau=-L_-^{(d)}}^{L_+^{(d)}} w^{(d)}(\tau)\mathbf{c}_{t+\tau} \text{ avec } d = 1,2
\end{aligned} \qquad \text{(Équation 5.23)}$$

où les $w^{(d)}(\tau)$ sont les coefficients utilisés dans l'algorithme de discrétisation des dérivées d'ordre d (T est l'opérateur de transposition)[105]. En considérant une séquence $\mathbf{x}$ de N observations telle que $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_N]$, l'équation 5.23 peut s'écrire sous forme matricielle :

$$\mathbf{x} = W\mathbf{c} \quad \text{avec}$$

$$\begin{cases}
\mathbf{c} = [\mathbf{c}_1^T, ..., \mathbf{c}_N^T]^T \\
W = [W_1, ..., W_N]^T \otimes I_{M\times M} \\
W_t = [w_t^{(0)}, w_t^{(1)}, w_t^{(2)}] \\
w_t^{(d)} = \left[\underbrace{0, ..., 0}_{t-L_-^{(d)}-1}, w^{(d)}(-L_-^{(d)}), w^{(d)}(L_+^{(d)}), \underbrace{0, ..., 0}_{T-(t+L_+^{(d)})} \right]^T \\
L_-^{(0)} = L_+^{(0)} = 0 \text{ et } w^{(0)} = 1
\end{cases} \qquad \text{(Équation 5.24)}$$

On définit ensuite la matrice $M \in \mathfrak{M}_{3M,K}(\mathbb{R})$ telle que $M = [\mu_1^T, ..., \mu_K^T]$ où $\mu_k \in \mathfrak{M}_{3M,1}(\mathbb{R})$ est l'espérance de la loi normale associée à l'état k (K est le nombre d'états du modèle $\mathbf{m}$). De

[104] Le cas d'un modèle semi-markovien pour lequel les (lois de) probabilités d'émission sont modélisées par un mélange de gaussiennes n'est pas envisagé ici.

[105] Typiquement, on utilise $w^{(1)} = [-1/2, 0, 1/2]$ et $w^{(2)} = [1, -2, 1]$.

façon symétrique, on définit également la matrice $U = [U_1,...,U_K]$ où $U_k \in \mathfrak{M}_{3M\times 3M}(\mathbb{R})$ est la matrice de covariance (diagonale) de la loi normale associée à l'état k. La recherche de la séquence d'observation qui maximise la vraisemblance du modèle (ou ici la log-vraisemblance), tout en respectant les conditions sur sa structure formulées à l'équation 5.23, s'effectue en résolvant l'équation suivante :

$$\frac{\delta \log p(W\mathbf{c} \mid \hat{Q}, \lambda)}{\delta \mathbf{c}} = 0 \qquad \text{(Équation 5.25)}$$

dont Tokuda montre que la solution est :

$$\mathbf{c} = (W^T U^{-1} W)^{-1} W^T U^{-1} M \qquad \text{(Équation 5.26)}$$

Comme l'illustre la Figure 5.12, la séquence de paramètres ainsi obtenue ne présente pas de discontinuité et peut donc être utilisée pour la synthèse du signal. On a cependant reproché à cette méthode de produire des trajectoires « trop lisses ». Pour faire face à ce problème, Toda propose dans (Toda et Tokuda, 2007) de pénaliser, lors de l'inférence, les trajectoires présentant une variance globale trop faible (une valeur cible étant préalablement estimée sur l'ensemble d'apprentissage).

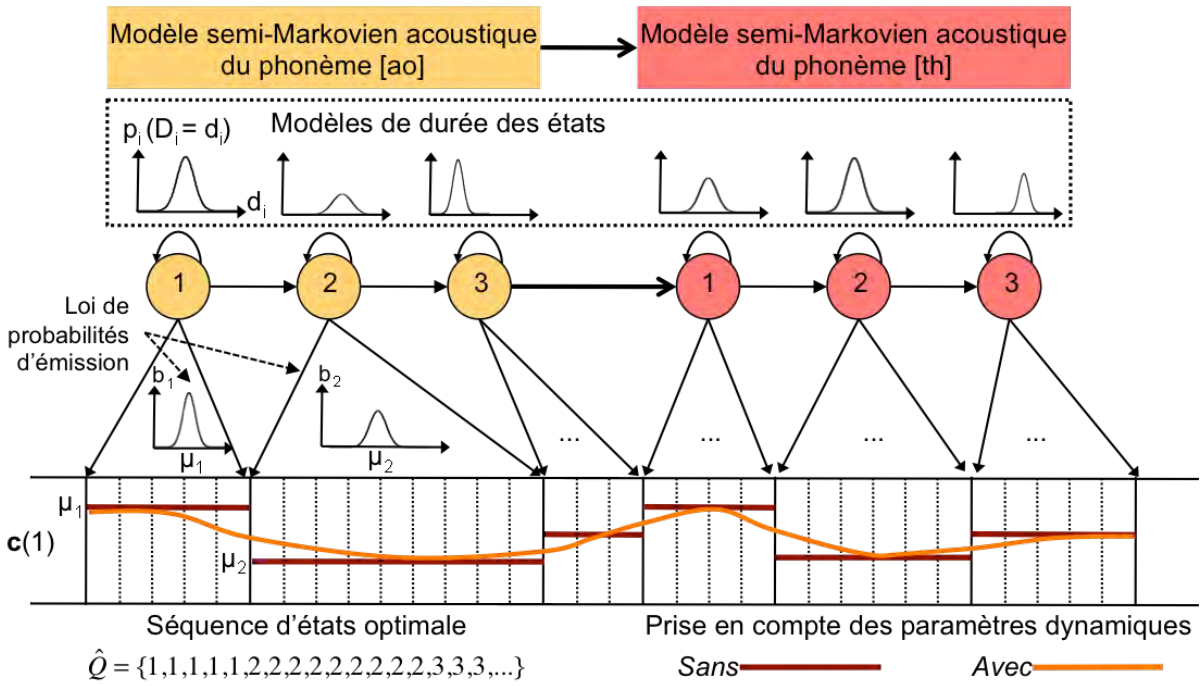


Figure 5.12 : Inférence d'une trajectoire acoustique à l'aide de modèles semi-Markoviens (illustration dans le cas de la synthèse de chaîne phonétique [ao]-[th], cas d'un vecteur de paramètres monodimensionnel)

Mise en œuvre

Cette technique d'inférence par modèles stochastiques est mise en œuvre, dans le cadre de cette étude, pour la génération des paramètres HNM à partir du treillis d'hypothèses phonétiques.

La procédure d'analyse HNM, décrite à la section 3.4.2, fournit pour chaque phrase, un flux de vecteurs acoustiques de dimension 31 dont la structure est rappelée ci-après:

$$\underbrace{\boxed{f_0}\boxed{G_h}\boxed{lsf_{h1}, lsf_{h2}-lsf_{h1}, \ldots, lsf_{h12}-lsf_{h11}}}_{\text{Partie Harmonique } \mathbf{P}_{harm}}\underbrace{\boxed{G_b}\boxed{lsf_{b1}, lsf_{b2}-lsf_{b1}, \ldots, lsf_{b16}-lsf_{b15}}}_{\text{Partie Bruit } \mathbf{P}_{bruit}}$$

Chaque vecteur acoustique est complété par les dérivées premières et secondes de chacune des caractéristiques (ce qui fournit un vecteur de dimension 93). Contrairement aux modélisations précédentes, une partie des paramètres mis en jeu ici n'est définie que sur certaines portions du signal acoustique. Il s'agit de la fréquence fondamentale et des coefficients LSF décrivant la partie harmonique (ainsi que leurs dérivées premières et secondes) qui ne sont non nuls que sur les trames voisées. Pour modéliser l'évolution de ces paramètres, nous utilisons l'approche MSD-HMM (pour « *Multi-Space Probability Distribution HMM* ») décrite dans (Tokuda *et al.*, 2002). Le flux de paramètres acoustiques est alors interprété comme 7 flux distincts, tel que :

$$\underbrace{\boxed{f_0}\boxed{\Delta^{(1)} f_0}\boxed{\Delta^{(2)} f_0}\boxed{\mathbf{P}_{harm}}\boxed{\Delta^{(1)}\mathbf{P}_{harm}}\boxed{\Delta^{(2)}\mathbf{P}_{harm}}}_{\text{Modélisation par MSD-HMM}}\underbrace{\boxed{[\mathbf{P}_{bruit}, \Delta^{(1)}\mathbf{P}_{bruit}, \Delta^{(2)}\mathbf{P}_{bruit}]}}_{\text{Modélisation par modèles semi-Markoviens}}$$

Un ensemble de modèles acoustiques (à trois états), dépendant du contexte (triphones) est ensuite entraîné selon la procédure décrite précédemment pour l'apprentissage des MMC visuels (en sautant cependant l'étape d'incrémentation du nombre de gaussiennes utilisées pour modéliser les densités de probabilités d'émission, voir Figure 5.4).

La procédure d'inférence décrite à la section précédente est basée sur la construction d'un modèle « gauche-droite » qui représente la chaîne phonétique à synthétiser. Cette chaîne doit être définie de façon univoque et il n'est *a priori* pas possible de considérer, pendant l'inférence, plusieurs hypothèses phonétiques alternatives (du moins dans la formulation standard de la synthèse par MMC considérée ici). Le treillis d'hypothèses phonétiques décrit à la section 5.3.1 ne peut donc pas être utilisé ici « en l'état ». Une autre approche est alors adoptée. Cette dernière consiste à représenter par un même modèle acoustique les phonèmes dont les réalisations acoustiques sont jugées « proches » comme [uw] et [uh], [iy] et [ih] et certains « sosies labiaux et linguaux » : [p] et [b], [t] et [d], [k] et [g], [ch] et [jh]. En pratique, ceci s'effectue simplement en changeant, d'une part, les étiquettes des segments de l'ensemble d'apprentissage, et, d'autre part, celles des segments identifiés par le décodeur visuo-phonétique. Le but de ces regroupements est d'effectuer une synthèse que nous qualifions ici de « floue ». L'objectif étant de profiter du caractère « lisse » des trajectoires inférées par l'approche

stochastique afin de limiter l'impact d'une erreur « locale » sur l'intelligibilité globale de la phrase.

5.3.5. Génération de la prosodie

Les deux techniques d'inférence des paramètres HNM décrites précédemment sont basées sur une description segmentale du signal de parole (au niveau phonétique). Aucune caractéristique de type « suprasegmentale » renseignant notamment sur son contenu prosodique[106] n'est prise en compte. En effet, pour la sélection des unités dans l'approche concaténative, comme pour la définition des modèles acoustiques dans l'approche stochastique, l'étiquette phonétique est l'unique descripteur symbolique utilisé[107]. Aussi, il apparaît peu probable que le signal synthétisé à partir des paramètres HNM, sélectionnés ou inférés, présente des caractéristiques prosodiques acceptables. Une approche simple est alors adoptée pour définir un contenu prosodique plus « naturel ».

Définir le contenu prosodique d'un signal de synthèse revient principalement à déterminer d'une part le placement des accents et, d'autre part, l'intonation globale sur la phrase. De façon générale, un positionnement correct des syllabes accentuées est un élément très important pour l'intelligibilité du signal de synthèse. Ceci est d'autant plus vrai pour les langues dites « à accent libre » comme l'anglais, pour lesquelles la position de l'accent dépend notamment du mot prononcé[108]. Le placement automatique des accents nécessite donc une description du signal au niveau lexical. Ce niveau est notamment accessible en sortie du décodeur visuo-phonétique dans le cas du scénario « contraint »[109]. La suite de mots correspondant à la chaîne phonétique décodée est alors envoyée dans le module de « génération de l'intonation » du système TTS *Festival* (Black et Lenzo, 2000). L'approche adoptée par ce dernier pour déterminer les syllabes devant être accentuées ne sera pas détaillée ici. On notera cependant que cette dernière est basée sur l'extraction puis l'analyse de descripteurs symboliques contextuels par un arbre de décision du type CART (« *classification and regression tree* »). L'intonation globale sur la phrase est ensuite déterminée, par ce même module, à l'aide d'une approche par règles, dans laquelle la courbe mélodique est construite à partir de « points cibles » qui définissent la hauteur de chaque syllabe.

---

[106] Le terme « prosodie » renvoie ici principalement à l'intonation (contour mélodique) de la phrase et aux variations de la fréquence fondamentale au niveau syllabique (accentuation).

[107] Un système TTS fait généralement intervenir des descripteurs symboliques de plus haut niveau qui « expliquent » le contenu prosodique (tels que la position de la syllabe dans le mot, dans la phrase, etc.). Dans un système par concaténation d'unités, ces descripteurs sont pris en compte dans l'algorithme de sélection. Dans un système basé sur la synthèse par MMC, ces descripteurs sont considérés comme des marqueurs contextuels et sont pris en compte lors de la définition des modèles.

[108] Par exemple, les mots « *diplomat* », « *diplomacy* » et « *diplomatic* » reçoivent respectivement l'accent sur leur première, seconde et troisième syllabe (Boite *et al.*, 2000).

[109] Dans ce scénario, les suites phonétiques autorisées sont celles que l'on peut construire en combinant les 3000 mots du dictionnaire (voir section 5.2.2).

Il apparaît de façon claire que cette approche n'est optimale que dans le cas où la suite de mots décodée est correcte, ce qui au regard des performances obtenues sur la tâche de décodage au niveau lexical, reste relativement rare (taux de reconnaissance en mots de l'ordre de 61 % sur la base B2). Néanmoins, et c'est l'hypothèse qui est faite ici, lorsque la structure syntaxique de la phrase décodée est assez proche de celle de la phrase originale, la courbe d'évolution de la fréquence fondamentale obtenue par cette approche peut être considérée comme une bonne candidate pour la synthèse du signal. La Figure 5.13 illustre cela sur l'exemple de la phrase « *Now animals do not like mockery* » qui dans nos expérimentations est décodée comme « *No animals to hot like mockery* ».

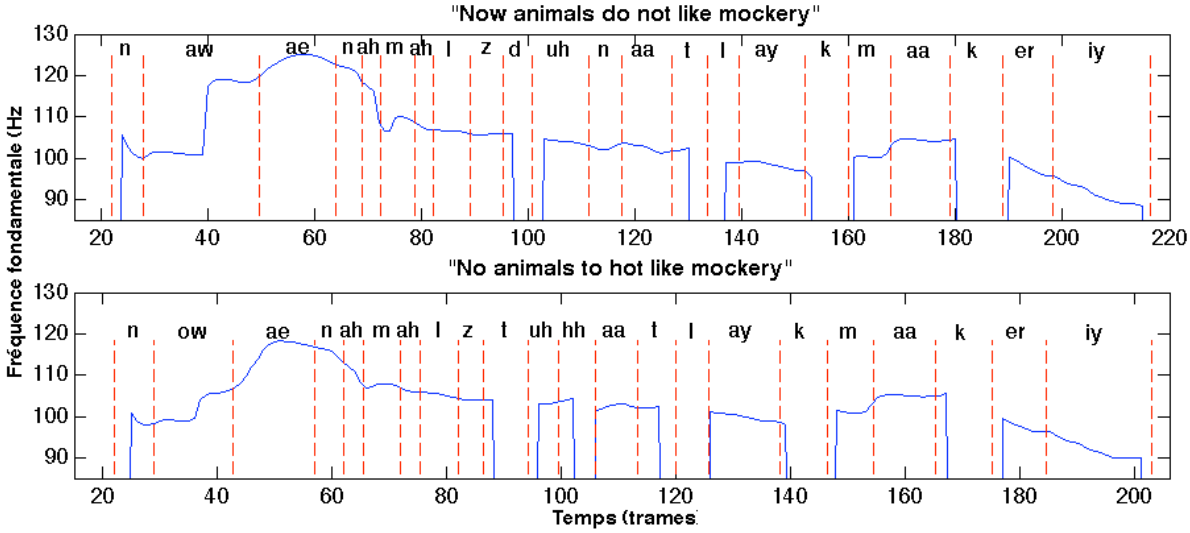


Figure 5.13 : Génération de la courbe d'évolution de la fréquence fondamentale à l'aide du système TTS *Festival* à partir du décodage de la séquence visuelle au niveau lexical. En haut, la courbe obtenu à partir de la suite de mots « correcte », en bas, à partir de la suite de mots « décodée » (qui comporte plusieurs erreurs).

5.3.6. Synthèse du signal

Dans le cas de l'approche par concaténation d'unités, la durée de chaque diphone sélectionné est adaptée pour être en cohérence avec les durées décodées dans le flux visuel (ces dernières étant elles-mêmes corrigées par la procédure décrite à la section 5.3.2). Cette opération d'étirement (ou de compression) s'effectue par ré-échantillonnage (interpolation linéaire), dans le domaine temporel, de la séquence de paramètres HNM. La séquence de paramètres du signal à synthétiser est ensuite obtenue par concaténation des séquences de paramètres des unités sélectionnées. Un lissage des paramètres HNM autour des instants de concaténation est également effectué. Dans l'approche stochastique, la durée souhaitée pour chacun des phonèmes est spécifiée au moment de l'inférence. Aucune adaptation *a posteriori* n'est donc nécessaire. Pour chacun des phonèmes, la fréquence fondamentale inférée est ensuite adaptée de telle sorte que sa valeur moyenne sur le segment soit égale à la valeur moyenne de la courbe « cible », obtenue à l'aide de l'approche décrite dans la section précédente. Enfin, le

signal est généré selon le schéma de synthèse HNM précédemment décrit à la Figure 3.14 (« phase de synthèse »).

5.3.7. Evaluation

Dans la présente section, la qualité du signal de parole obtenu par conversion visuo-acoustique indirecte est discutée en terme d' « intelligibilité » et de « qualité vocale ». La notion d'intelligibilité fait ici référence à la capacité d'un auditeur à comprendre le message émis par « l'utilisateur de l'ICPS ». Le message perçu pouvant être décrit comme le résultat de l'application du « bruit de conversion visuo-acoustique » sur le message original, il s'agit, en d'autres termes, d'estimer la tolérance de l'auditeur à ce bruit. La mesure de l'intelligibilité peut se réaliser de différentes manières, à l'aide, par exemple, d'un test subjectif (test MOS, *Mean Opinion Scores*), d'un test d'intelligibilité segmentale (test de rimes à paires minimales) ou bien, comme c'est la cas ici, d'un test par transcription. La « qualité vocale » est usuellement définie comme ce qui différencie deux productions vocales ayant le même contenu lexical. Il s'agit d'une notion complexe qui fait intervenir différentes caractéristiques, à différents niveaux de description du signal. Ainsi, interviennent dans l'évaluation de la qualité vocale, des aspects liés à l'intonation, à l'intensité, au rythme, à l'articulation, et enfin à la qualité sonore (Payri, 2000).

Evaluation qualitative

Une première évaluation « qualitative » de l'intelligibilité et de la qualité vocale d'un signal de synthèse obtenu par conversion visuo-acoustique « indirecte », est ici effectuée à partir de la série d'exemples sonores présentée sur la page Internet associée à ce manuscrit. A l'écoute de ces derniers, on constate tout d'abord que la qualité vocale et l'intelligibilité du signal sont bien supérieures à celles obtenues dans le cas de l'approche directe de la conversion visuo-acoustique (présentée au chapitre précédent). Néanmoins, l'intelligibilité du signal de parole reste fortement corrélée avec le taux de reconnaissance phonétique. Des signaux générés à partir d'une chaine phonétique contenant plus de 25 % d'erreurs (taux de reconnaissance phonétique inférieur à 75 %) ne sont généralement pas intelligibles. Le type d'erreur commise est également important. Certaines erreurs de substitution ont un impact relativement limité sur l'intelligibilité globale, comme celles commises entre certains sosies labiaux ([p], [b]) ([t], [d]), [f], [v], etc.), certains phonèmes proches ([uh], [uw]), ou certaines diphtongues (lorsque ces dernières sont décodées comme une séquence de deux voyelles distinctes). A l'inverse, d'autres erreurs de substitution sont plus perturbantes, notamment celles commises entre les phonèmes qui ne se différencient que par leur caractéristique de nasalité ([m] *versus* [p] ou [b], [n] *versus* [t] ou [d]). Plus fréquentes, les erreurs d'omission et surtout d'insertion, ont un impact également important sur l'intelligibilité du signal. Une seule erreur d'insertion placée à un endroit « stratégique » de la chaîne phonétique décodée, suffit parfois à rendre le signal de synthèse difficilement intelligible. Par ailleurs, la qualité vocale du signal varie fortement en

fonction de l'approche utilisée pour l'inférence des paramètres HNM. Pour illustrer la discussion sur ce point, les paramètres HNM de deux exemples sonores[110], obtenus après décodage visuo-phonétique de la même séquence visuelle, sont présentés respectivement à la Figure 5.14 et à la Figure 5.15. Le premier exemple est basé sur l'approche d'inférence par sélection d'unités (section 5.3.3), le second sur l'approche stochastique (section 5.3.4). Basée sur l'utilisation de « vrais » segments de parole, l'approche concaténative permet d'obtenir un signal au contenu spectral riche, qui préserve les détails de l'articulation (on observera à la Figure 5.14 l'aspect « bruité » des coefficients LSF). Il s'agit de l'approche qui, à ce jour (et dans ce contexte particulier), est à l'origine du résultat le plus « naturel ». Cependant, cette qualité s'obtient au prix d'une mauvaise robustesse face aux erreurs de décodage visuo-phonétique. En effet, lorsque le parcours du treillis d'hypothèses par l'algorithme de sélection d'unités ne permet pas de la corriger (voir sections 5.3.1 et 5.3.3), une erreur de décodage introduit, dans le signal de synthèse, un segment « inapproprié » qui vient limiter fortement l'intelligibilité et la qualité vocale de ce dernier. L'approche stochastique, comme l'illustre la Figure 5.15, fournie un résultat cohérent avec les considérations théoriques présentées précédemment (section 5.3.4), c'est-à-dire des trajectoires acoustiques « lisses » et ce, de façon beaucoup plus prononcée que dans le cas de l'approche concaténative. Cette propriété rend cette approche moins sensible aux erreurs de décodage visuo-phonétique. En effet, contrairement à l'approche concaténative, un phonème « aberrant » entre deux phonèmes correctement décodés, représentera une cible acoustique « lointaine », qui en raison des contraintes sur la dynamique, ne pourra qu'être approchée, mais non complètement atteinte. L'impact de cette erreur de décodage sera donc réduit. Néanmoins, la conséquence de cette « continuité acoustique » est une moins bonne restitution des détails de l'articulation et, en comparaison avec l'approche concaténative, la qualité vocale du signal obtenu par une approche stochastique semble (ici) plus faible.

[110] Ces deux exemples sonores sont accessibles à la section « Exemple 6 » de la page Internet (la phrase cible est « *Now animals do not like mockery* »).

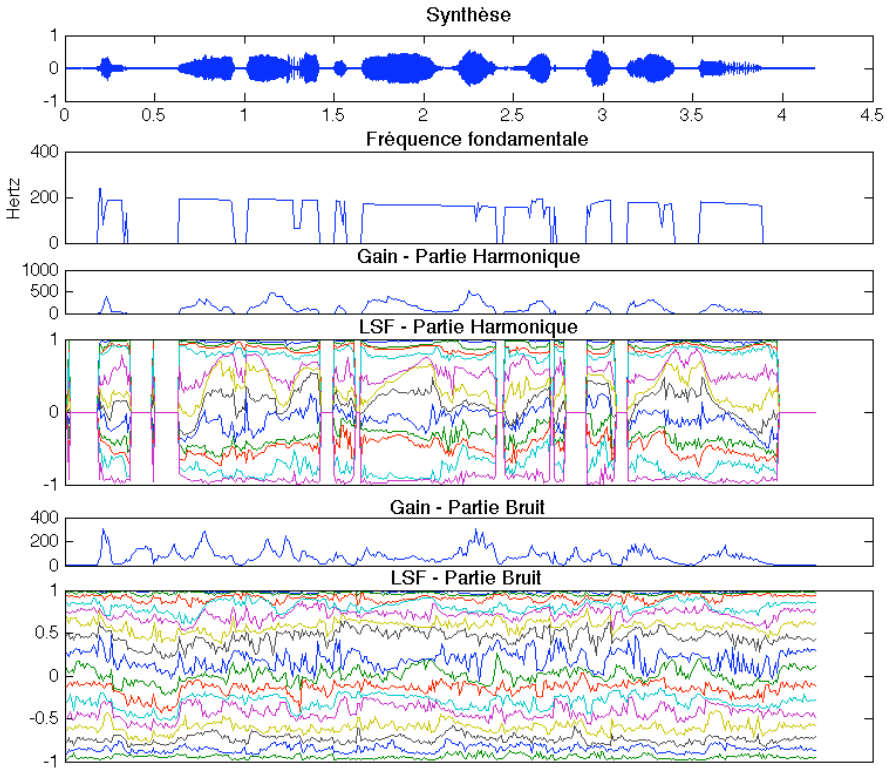


Figure 5.14 : Exemple d'un signal de synthèse obtenu dans le cas de l'approche indirecte de la conversion visuo-acoustique, cas de l'inférence des paramètres HNM par sélection d'unités. De haut en bas, la forme d'onde, la fréquence fondamentale « cible » (voir section 5.3.5) puis successivement, gain du filtre AR et coefficients LSF pour les parties harmonique et bruit.

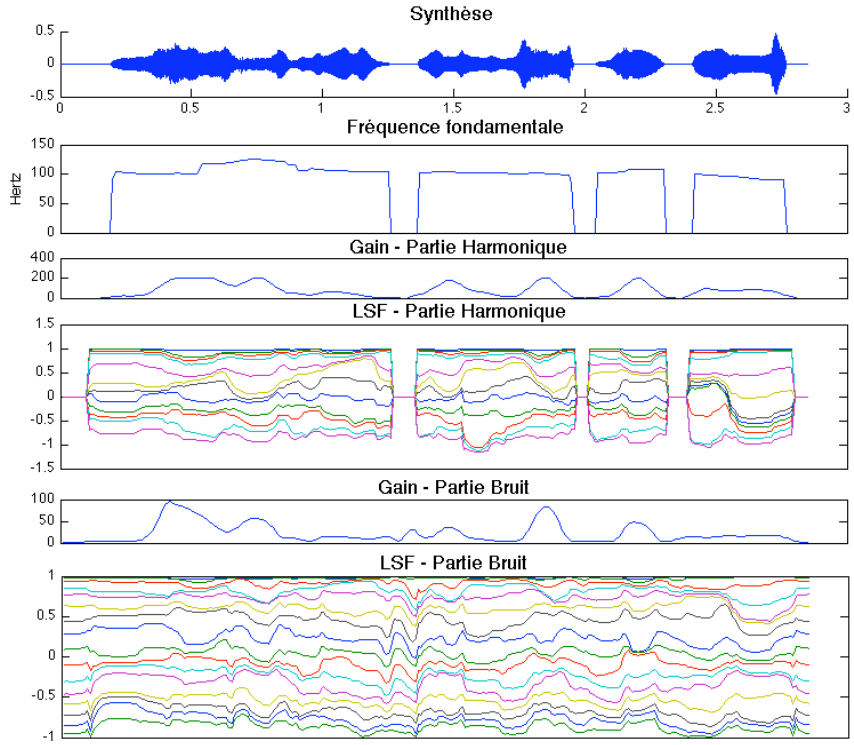


Figure 5.15 : Exemple d'un signal de synthèse obtenu dans le cas de l'approche indirecte de la conversion visuo-acoustique, cas de l'inférence des paramètres HNM par synthèse stochastique, à partir de la même séquence visuelle que celle utilisée à la Figure 5.14.

Evaluation « quantitative » (perceptive)

Afin d'évaluer « quantitativement » l'approche indirecte de la conversion visuo-acoustique, un test d'intelligibilité est mis en œuvre. Trois objectifs particuliers sont poursuivis :

1) Evaluer la qualité du « module de synthèse » de façon indépendante de celle du décodeur visuo-phonétique (pour lequel nous disposons d'un critère d'évaluation objectif qui est le taux de reconnaissance phonétique).
2) Comparer les deux approches d'inférence de paramètres HNM (compromis « robustesse aux erreurs de décodage *versus* qualité vocale »)
3) Evaluer la « véritable » performance du système, c'est-à-dire l'intelligibilité d'un signal de synthèse obtenu à partir d'une chaîne phonétique comportant un certain nombre d'erreurs.

Le test mis en œuvre est un « test par transcription ». Le sujet se voit présenter un exemple sonore qu'il doit retranscrire par écrit. Le sujet réalise le test à distance, en remplissant le formulaire en ligne[111] dont une capture d'écran est présentée à la Figure 5.16.

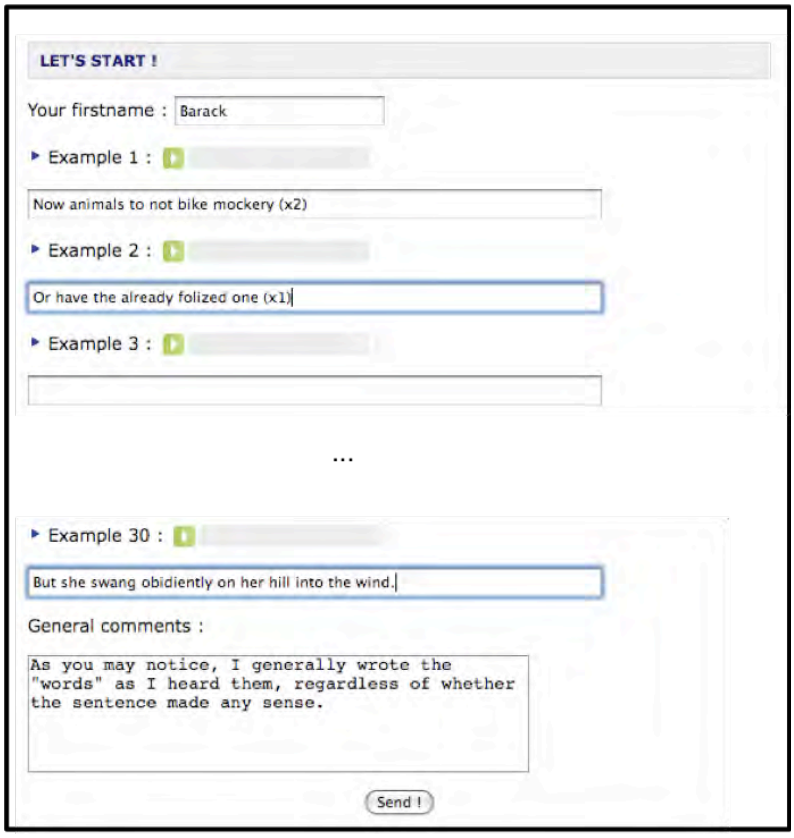


Figure 5.16 : Capture d'écran du formulaire en ligne mis en œuvre pour les tests d'intelligibilité.

Il est suggéré au sujet d'effectuer le test soit en utilisant un casque, soit dans un environnement « serein » dans le cas d'une écoute par haut-parleurs (il reste toutefois impossible de s'assurer que

[111] Ce formulaire est accessible depuis la page Internet associée à ce manuscrit.

cette condition soit respectée). L'unique consigne donnée au sujet est d'écouter puis de retranscrire immédiatement chaque exemple sonore. Si le sujet a besoin de N écoutes supplémentaires, il doit ajouter à la fin de la transcription la mention « (xN) ». Il n'est par ailleurs nullement spécifié au sujet si les phrases qui lui sont présentées, comme celles qu'il doit construire, sont porteuses ou non de sens (on remarquera à ce titre le commentaire laissé par le sujet à la Figure 5.16). Pour construire l'ensemble des stimuli, 60 phrases sont choisies aléatoirement dans les ensembles de test successifs fournis par la procédure de *jacknife* (voir 5.2.2). Puis on sélectionne, parmi ces 60 phrases, un ensemble de 15 phrases qui respectent les deux critères suivants :

- Etre « courte » (une dizaine de mots au maximum) afin de réduire les erreurs dues uniquement aux difficultés potentielles de mémorisation du sujet.
- Présenter un taux de reconnaissance phonétique de l'ordre de celui observé sur l'ensemble du corpus, dans le cas du scénario de décodage contraint (soit 75 % pour la base B1 et 80 % pour la base B2, voir section 5.2.3).

Quatre ensembles de stimuli sont ensuite construits avec ces 15 phrases :

- Ensemble E1 : 15 exemples sonores générés après décodage visuo-phonétique (cas du scénario de décodage contraint) puis synthèse à l'aide de l'approche concaténative.
- Ensemble E2 : 15 exemples sonores générés après décodage visuo-phonétique dans le cas où la transcription phonétique est connue (alignement forcé), puis synthèse à l'aide de l'approche concaténative.
- Ensemble E3 : 15 exemples sonores générés après décodage visuo-phonétique (cas du scénario de décodage contraint) puis synthèse à l'aide de l'approche stochastique.
- Ensemble E4 : 15 exemples sonores générés après décodage visuo-phonétique dans le cas où la transcription phonétique est connue (alignement forcé), puis synthèse à l'aide de l'approche stochastique.

Quatorze sujets ont participé au test (auditeurs naïfs et non rémunérés pour cette tâche). Sept se sont vu présenter successivement les ensembles E1 et E2 (synthèse concaténative), les sept autres les ensembles E3 et E4 (synthèse stochastique), soit 30 phrases à transcrire pour chacun des sujets.

Le critère retenu pour la mesure de l'intelligibilité est le « taux de reconnaissance en mots », T_m, utilisé classiquement pour évaluer un système de transcription automatique de la parole, défini par :

$$T_m = \frac{N_m - D - S - I}{N_m} \qquad \text{(Équation 5.27)}$$

où N_m est le nombre de mots dans chacun des ensembles de stimuli (soit 110 mots ici). L'alignement de la transcription effectuée par le sujet avec la transcription de référence s'effectue par programmation dynamique, de façon similaire à ce qui a été présenté pour la

mesure du taux de reconnaissance phonétique. Les résultats obtenus sont résumés dans les graphiques présentés à la Figure 5.17.

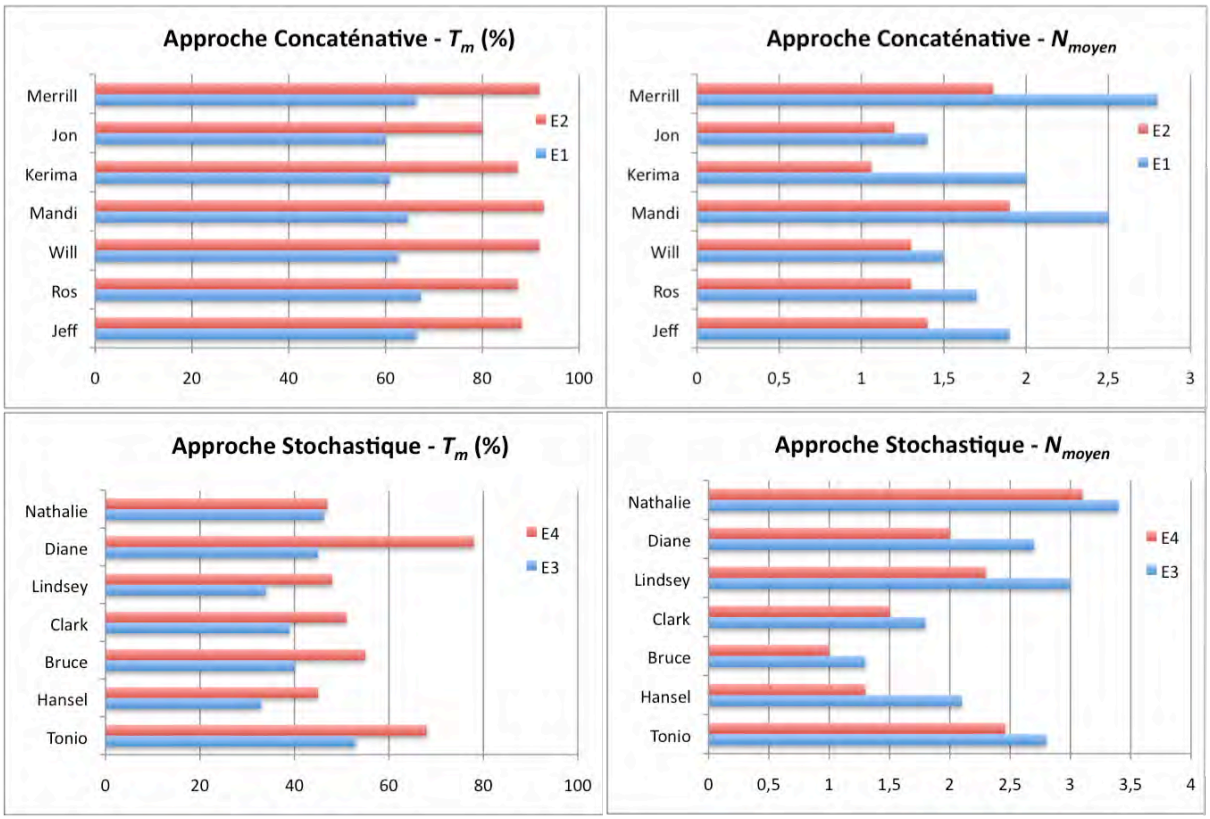


Figure 5.17 : Résultats du test d'intelligibilité (à gauche, le taux de reconnaissance en mots T_m, à droite, le nombre moyen d'écoutes N_{moyen}). En rouge, cas des signaux synthétisés à partir d'une suite phonétique correcte (T_p = 100 %), en bleu, signaux synthétisés lorsque $T_p \approx 80$ % (performance « véritable » du système)

Le tâche de transcription des ensembles E2 et E4 (pour lesquels T_p = 100 %) permet de satisfaire l'objectif (1) précédemment mentionné, à savoir l'évaluation du module de synthèse de façon indépendante de celle du décodeur visuo-phonétique. L'intelligibilité moyenne (mesurée par T_m) est de 88 % pour l'approche concaténative et de 56 % pour l'approche stochastique, et ce, avec un nombre moyen d'écoutes respectivement de 1.4 et 1.9 (la différence de performance observée entre les deux approches sera discutée plus loin). La transcription fournie par les sujets reste donc imparfaite malgré un contenu segmental (phonétique) correct. Une partie des erreurs commises peut tout d'abord s'expliquer par la présence d'artéfacts de synthèse. Rappelons que, dans le cas de l'approche concaténative, le dictionnaire d'unités est construit à partir d'un enregistrement d'un peu moins d'une heure de parole, ce qui est assez peu en comparaison avec un système TTS standard et ne permet pas de couvrir de façon optimale l'espace acoustique. De plus, l'étiquetage du corpus d'apprentissage au niveau phonétique est ici effectué de façon automatique. Aucune correction manuelle visant à corriger les éventuelles erreurs d'alignement n'est effectuée. L'autre partie des erreurs semble liée au rythme de la parole produite. Ce dernier

n'apparaît généralement pas naturel. Ceci semble indiquer que la procédure décrite à la section 5.3.2 ne permet pas d'adapter correctement les durées décodées dans la séquence visuelle.

La comparaison des performances obtenues sur les ensembles E1 et E2 d'une part, avec celles obtenues sur les ensembles E3 et E4 d'autre part, permet de satisfaire l'objectif (2) précédemment mentionné, à savoir la comparaison des approches concaténative et stochastique. Avec un taux de reconnaissance en mots de 65 % et 88 % sur les ensembles E1 et E2 contre 41 % et 58 % sur les ensembles E3 et E4, les signaux de parole générés à l'aide de l'approche concaténative semblent plus intelligible que ceux générés à l'aide de l'approche stochastique. La « mauvaise » performance de l'approche stochastique peut néanmoins surprendre au regard des performances généralement atteintes par un système standard comme HTS. Néanmoins, l'utilisation qui est ici faite de la technique de synthèse par MMC, est assez particulière. Rappelons que l'approche stochastique est mise en œuvre dans le but d'effectuer une synthèse que nous avons qualifiée de « floue ». Lors de l'apprentissage des modèles comme lors de l'inférence, l'espace acoustique est partitionné non pas en classes phonétiques mais, de façon plus « grossière », en classes de « sosies labiaux et linguaux » (voir section 5.3.4). Par ailleurs, dans l'approche stochastique, la procédure d'inférence des paramètres acoustiques ne dépend que de la suite d'étiquettes phonétiques fournie par le décodeur (et de la décomposition temporelle). Une fois le décodage visuo-phonétique effectué, le flux visuel n'est plus utilisé, contrairement à l'approche concaténative qui le met à contribution lors de la sélection des unités (dans la définition du « coût de cible » notamment, voir section 5.3.3).

Enfin, conformément à l'objectif (3) précédemment mentionné, ce test permet d'évaluer la performance « véritable » du système. Il s'agit de mesurer l'intelligibilité d'un signal de parole synthétisé à partir d'une séquence phonétique qui contient un certain nombre d'erreurs. Rappelons que ces dernières peuvent être soit « rectifiées », dans le cas de l'approche concaténative (utilisation d'un des chemins alternatifs du treillis d'hypothèses phonétiques), soit « atténuées », dans le cas de l'approche stochastique (technique dite de « synthèse floue »). Lorsque la performance du décodeur visuo-phonétique sur la séquence de test est de l'ordre 80 % (performance moyenne du système sur l'ensemble du corpus), dans le cas d'une synthèse basée sur l'approche concaténative, l'intelligibilité moyenne de la parole reconstituée est de l'ordre de 65 % (taux de reconnaissance « en mots » moyen, évalué sur les transcriptions fournies par les différents sujets). Ainsi, bien que certaines phrases soient remarquablement bien reconstruites, le système n'apparaît pas capable de produire de façon systématique, à ce jour, une parole totalement intelligible. Comme nous l'avons mentionné précédemment, l'intelligibilité du signal de synthèse dépend du nombre et du type d'erreurs de décodage commis. Cependant, elle dépend également de la qualité de son contenu prosodique, qui semble encore insuffisante. Si l'intonation globale sur la phrase semble acceptable, le placement des accents, notamment lexicaux, reste trop souvent imprécis. Rappelons que ces derniers sont déterminés suite au décodage de la séquence visuelle au niveau lexical (voir 5.3.5). Or, si le taux de reconnaissance

phonétique est en moyenne de l'ordre de 80 %, le taux de reconnaissance « en mots » plafonne à 49 % pour la base B1, et à 61 % pour la base B2 (voir section 5.2.3). Cette performance semble être suffisante pour générer une intonation globale acceptable mais reste encore trop faible pour espérer un placement correct des accents.

Enfin, ce résultat doit cependant être interprété à la lumière de la difficulté d'un tel test. Bien que, en toute rigueur, les phrases mises en œuvre dans ce test ne peuvent pas être qualifiées de « sémantiquement imprévisibles », on soulignera néanmoins la quasi impossibilité pour l'auditeur de tirer ici pleinement profit de ses capacités naturelles de suppléance mentale[112] . Or, au regard du caractère mal posé de la parole silencieuse précédemment mentionné, c'est justement cette dernière qui semble pouvoir compenser l'incomplétude du message reconstitué par conversion visuo-acoustique. Aussi, bien que cela reste à être mis en œuvre, une évaluation basée sur un contenu linguistique plus « réaliste », devrait révéler une intelligibilité supérieure à celle mesurée ici.

[112] Sur une phrase comme « *I noticed blood spouting from Kerfoot's left hand.* », il apparaît en effet difficile pour l'auditeur d'inférer un contenu lexical mal identifiable dans le signal de synthèse, uniquement en s'aidant d'informations contextuelles.

Conclusion générale et Perspectives

Rappel de la problématique

Ce travail de thèse s'inscrit dans la récente thématique de recherche sur la réalisation « d'interfaces de communication en parole silencieuse », dispositifs capables d'interpréter une parole normalement articulée mais non vocalisée. Les principales applications visées par ce type d'interface sont l'aide aux patients laryngectomisés (suite à un cancer du larynx), la communication parlée dans un environnement où le silence est requis, ou à l'inverse, dans un lieu extrêmement bruyant. Le dispositif envisagé dans cette étude combine deux systèmes d'imagerie pour capturer l'activité de l'appareil vocal : l'échographie, qui renseigne principalement sur les mouvements de la langue, et la vidéo, qui est utilisée pour saisir l'activité labiale. Le problème posé est celui de la « conversion visuo-acoustique », c'est-à-dire la synthèse d'un signal de parole « acoustique », uniquement à partir d'un flux de données « visuelles » (images ultrasonores et vidéo).

Contributions principales et résultats obtenus

Les différentes approches proposées pour réaliser cette conversion visuo-acoustique utilisent l'apprentissage artificiel pour modéliser les liens entre le geste articulatoire et la réalisation acoustique. Ces approches par apprentissage nécessitent tout d'abord l'acquisition de données d'étude. Ce point, abordé dans le second chapitre de ce document, a fait l'objet de plusieurs contributions. Un système de positionnement de la sonde ultrasonore par rapport à la tête du locuteur, basé sur l'utilisation combinée de deux capteurs inertiaux (accéléromètres 3-axes) a tout d'abord été conçu. Ce système, utilisé aujourd'hui pour repositionner le référentiel « tête-sonde » entre deux sessions d'acquisition de données, pourra notamment être mis en œuvre dans un futur prototype embarqué. Puis, un système d'acquisition permettant l'enregistrement simultané des flux audio et visuels a été développé. Ce dernier, nommé *Ultraspeech*, est basé sur la synchronisation des différents capteurs (ultrasonore, vidéo, audio et inertiaux) par voie logicielle ; il permet, à ce jour, l'enregistrement des flux ultrasonore et vidéo à 60 ips. Deux bases de données, contenant chacune environ une heure de parole continue, en langue anglaise (corpus CMU Arctic), ont enfin été constituées. La première est construite à l'aide du système d'acquisition du *Vocal Tract Visualization Lab* (flux visuels cadencés à 30 ips), la seconde, à l'aide du système *Ultraspeech* (flux visuels cadencés à 60 ips).

La seconde étape vers la conversion visuo-acoustique est l'extraction de l'information pertinente dans les données enregistrées. Présentées au troisième chapitre, différentes techniques d'analyse ont été mises en œuvre dans ce but. Une procédure de réduction du bruit de *speckle* dans les images ultrasonores a tout d'abord été suggérée. Puis, une première approche de caractérisation des images ultrasonores, basée sur l'extraction du contour de la surface supérieure de la langue, a été envisagée. Ne prenant pas en compte les autres structures présentes dans

l'image, et jugée peu robuste face aux altérations fréquentes que cette dernière peut subir, cette approche « par segmentation » a cependant été abandonnée. Pour le codage des images ultrasonores comme pour celui des images vidéo, deux autres approches ont alors été mises en œuvre. Prenant en compte la totalité de la région d'intérêt dans le codage, ces dernières sont qualifiées « d'approches globales ». La première est basée sur l'utilisation de la transformée en cosinus discrète (TCD), la seconde sur l'analyse en composantes principales (approche *EigenTongues/EigenLips*). Dans les diverses expériences menées, c'est cette seconde approche qui est à l'origine des meilleurs résultats. Enfin, deux procédures « d'analyse-synthèse » du signal acoustique ont été présentées, la décomposition mel-cepstrale complexe, et la modélisation « Harmonique plus Bruit » (HNM).

Introduite au quatrième chapitre, la première approche proposée pour réaliser la conversion visuo-acoustique, est qualifiée de « directe ». Cette approche est basée sur l'apprentissage d'une « fonction de transformation » qui convertit (directement) les caractéristiques visuelles en caractéristiques acoustiques. La synthèse du signal de parole est ici réalisée à l'aide d'un modèle « source-filtre » (filtre MLSA). Un prétraitement des caractéristiques, basé sur l'analyse de corrélation canonique des flux visuels et acoustique, a tout d'abord été proposé. Deux approches ont ensuite été comparées pour l'inférence des paramètres du filtre (coefficients mel-cepstraux). La première est basée sur une modélisation par réseau de neurones, la seconde sur la modélisation de l'espace conjoint « visuo-acoustique » par un mélange de gaussiennes. Pour construire la fonction d'excitation du filtre, un classifieur neuronal est mis en œuvre pour prédire la caractéristique de voisement, et un modèle du type mélange de gaussiennes est utilisé pour inférer la fréquence fondamentale des trames voisées. Bien que certaines caractéristiques acoustiques, comme les premiers coefficients mel-cepstraux ou la caractéristique de voisement, semblent pouvoir être prédits de façon assez précise, les résultats obtenus par cette approche « directe », restent globalement insuffisants pour permettre la synthèse d'un signal de parole intelligible. Nous avons alors supposé que cette « mauvaise » performance pouvait en grande partie s'expliquer par le caractère « mal posé » du problème que constitue la conversion visuo-acoustique. En l'absence d'activité laryngée, un même geste articulatoire observé peut en effet correspondre à plusieurs réalisations acoustiques différentes. Il est alors apparu nécessaire d'introduire dans la modélisation des informations supplémentaires permettant de lever certaines de ces « ambiguïtés articulatoires ». Dans ce but, nous avons proposé d'effectuer la conversion visuo-acoustique non plus au niveau du « signal », mais à un niveau linguistique supérieur, le niveau phonétique. Une nouvelle approche a donc été proposée.

Cette dernière, objet du cinquième chapitre, est qualifiée « d'approche indirecte » car elle introduit, en amont du processus de synthèse, une étape intermédiaire de décodage dit « visuo-phonétique ». La mise en œuvre de cette étape nécessite en phase d'apprentissage, la segmentation des flux visuels au niveau phonétique. En raison de l'asynchronie naturelle entre le geste articulatoire et la réalisation acoustique, les flux visuels et audio doivent être décrits de

façon « asynchrone ». Une procédure de segmentation a été proposée à cet effet. Chaque classe de segments visuels est modélisée par un modèle de Markov cachés (MMC). Pour combiner les deux modalités visuelles, une stratégie de « fusion au niveau des états du MMC » (MMC multi-flux) s'est avérée plus performante qu'une stratégie de « fusion au niveau des caractéristiques ». Afin de prendre en compte le phénomène de coarticulation, et plus particulièrement, le phénomène d'anticipation gestuelle (« désynchronisation » de la langue et des lèvres), une modélisation des classes phonétiques « en contexte » a été effectuée. En phase de test, deux scénarios ont été introduits pour le décodage d'une séquence visuelle. Le premier, qualifié de « non contraint », autorise le décodage de n'importe quelle suite phonétique et n'inclut donc aucune contrainte linguistique. Dans le second, qualifié à l'inverse de « contraint », les suites phonétiques autorisées sont celles que l'on peut construire en combinant les mots extraits d'un dictionnaire de 3000 éléments. Ce scenario cherche donc à résoudre certaines des « ambiguïtés articulatoires » précédemment mentionnées, en s'aidant d'informations de nature lexicale. La performance du décodeur visuo-phonétique constitue l'un des résultats principaux de ce travail. On retiendra notamment que sur les données visuelles acquises à l'aide du système d'acquisition *Ultraspeech*, traitées à l'aide de l'approche *EigenTongue/EigenLips*, le taux de reconnaissance phonétique (40 classes) est de 71.4 % dans le cas du scénario de décodage non-contraint, et de 83.3 % dans le cas du scénario contraint.

Dans l'approche indirecte de la conversion visuo-acoustique, la synthèse du signal est réalisée à l'aide de l'approche « Harmonique plus Bruit ». Deux méthodes ont été proposées pour l'inférence des paramètres HNM à partir de la chaîne phonétique décodée. La première est basée sur une approche « concaténative », la seconde sur une approche « stochastique » (synthèse par MMC). La prise en compte d'informations linguistiques de haut niveau ne permettant pas de lever toutes les ambiguïtés articulatoires, une procédure d'expansion de la chaine phonétique décodée en un treillis d'hypothèses, a été proposée. Dans le cadre de l'approche concaténative, il s'agit de compléter l'approche probabiliste jusqu'ici adoptée (décodeur par MMC), par une approche « données » ou « *data-driven* » (car basée sur l'exploration du dictionnaire). Dans l'approche stochastique, cette procédure a permis d'envisager une synthèse qualifiée de « floue », puisque basée sur l'apprentissage d'un MMC acoustique non pas pour chaque classe phonétique, mais pour chaque classe de « sosies labiaux et linguaux ». Pour mesurer l'intelligibilité du signal de parole généré par « conversion visuo-acoustique indirecte », un test d'intelligibilité basé sur une tâche de transcription a été mis en œuvre. La qualité de la transcription est évaluée en calculant le « taux de reconnaissance en mots » (critère classiquement utilisé pour évaluer la performance d'un système de transcription automatique). Le meilleur taux, obtenu à l'aide de l'approche concaténative, est à ce jour de 65 % (cas du scénario de décodage contraint). Bien qu'encore insuffisante pour envisager un système totalement fonctionnel, cette performance est néanmoins un des principaux résultats de ce travail.

Perspectives générales

Ce travail fournit un cadre général pour la réalisation d'un système de reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal. Ce cadre fait intervenir quatre étapes principales : l'acquisition des données, l'extraction des caractéristiques, la conversion visuo-acoustique et la synthèse du signal. Pour chacune de ces étapes, des solutions alternatives ou complémentaires à celles proposées dans ce manuscrit peuvent être envisagées.

Un sondage de la cavité buccale par ultrasons ne fournit qu'une information partielle sur l'activité articulatoire. La couche d'air entre la surface de la langue et le palais ne permet notamment pas de capturer l'activité du vélum (sauf en cas de contact linguo-vélaire). Ceci explique notamment les confusions effectuées par le décodeur visuo-phonétique entre les phonèmes ([p], [b], [m]) d'une part et ([t], [d], [n]), d'autres part. Le velum étant la réunion des fibres de plusieurs muscles, la capture de son activité par électromyographie pourra être envisagée.

Pour la caractérisation des images ultrasonores, des expérimentations préliminaires ont été menées sur l'utilisation des paramètres d'un « modèle actif d'apparence » (Cootes *et al.*, 1998). Cette méthode qui, d'une certaine manière, fait la synthèse des approches par segmentation et des approches globales, n'a pour l'instant pas permis d'obtenir des résultats supérieurs à ceux fournis par l'approche *EigenTongues*. Ceci peut notamment s'expliquer par la difficulté que constitue la tâche de segmentation manuelle des images utilisées pour l'apprentissage du modèle. Il est en effet difficile d'identifier et de « suivre » d'une image à l'autre un « point caractéristique » sur la surface de la langue. Des techniques d'estimation du flot optique, à l'aide par exemple des méthodes du type « *block matching* » ou des méthodes variationnelles, pourront être mises en œuvre dans ce but.

Un certain nombre de techniques peuvent également être envisagées pour améliorer la performance du décodage visuo-phonétique. Bien que partiellement intégré par la prise en compte du contexte dans la modélisation des classes phonétiques, le phénomène d'anticipation gestuelle reste toujours susceptible d'induire le décodeur en erreur. Une modélisation des flux visuels par MMC asynchrones (Gravier *et al.*, 2002), ou par réseaux bayésiens dynamiques pourrait permettre de mieux prendre en compte les dépendances temporelles entre les mouvements des différents articulateurs (Livescu, 2005). Par ailleurs, les différentes expériences menées dans ce travail ont montré la nécessité d'intégrer dans la conversion visuo-acoustique des connaissances linguistiques *a priori*. Ceci s'est traduit par l'utilisation d'un modèle de langage au niveau phonétique (bigramme), et par une restriction sur le vocabulaire (cas du scénario de décodage contraint). La structure particulière du corpus CMU Arctic n'a pas permis, dans cette étude, d'utiliser efficacement un modèle de langage au niveau lexical. Elément fondamental d'un système de transcription de la parole « acoustique », la mise en œuvre d'un tel modèle devra néanmoins être envisagée (éventuellement sur un corpus au contenu linguistique plus « réaliste »).

Plusieurs améliorations peuvent également être envisagées pour l'étape de synthèse du signal. Afin d'en améliorer le « rythme », une procédure plus performante d'adaptation de la durée des phonèmes décodés dans le flux visuel devra être envisagée. Cette dernière pourra, par exemple, utiliser un taux de déformation temporelle (voir section 5.3.2) qui dépend du contexte phonétique (comme dans (Govokhina, 2008)). Enfin, lors de l'évaluation de l'ICPS, il pourra être intéressant d'exploiter pleinement l'aspect multimodal de la perception de la parole, mis notamment en évidence par l'effet McGurk (McGurk, 1976). Il s'agirait alors de présenter au « destinataire » de la communication, non plus uniquement le signal de parole reconstitué, mais également l'image originale des lèvres de « l'émetteur ».

Etude des spécificités de l'articulation silencieuse

Dans ce travail, les expériences effectuées dans le cadre de la conversion visuo-acoustique sont basées sur des données acquises en « parole vocalisée », et non en « articulation silencieuse ». Ce dernier mode de production, qui sera celui utilisé dans l'application finale, peut cependant faire l'objet de certaines spécificités. En effet, l'articulation silencieuse interrompt la boucle de rétroaction entre la production et la perception, qui permet d'ajuster le geste articulatoire en fonction de la réalisation acoustique perçue. La caractérisation des spécificités de l'articulation silencieuse constitue sans aucun doute un axe de recherche à part entière, qui devra faire l'objet d'études dédiées. Quelques résultats préliminaires ont néanmoins été obtenus dans le cadre de ce travail de thèse. Ces derniers sont présentés ci-après.

La première expérience a portée sur le décodage visuo-phonétique des 60 phrases enregistrées en articulation silencieuse (voir section 2.4.2) à l'aide des modèles (MMC) entraînés sur les données visuelles acquises en parole vocalisée. Le taux de reconnaissance phonétique mesuré s'est alors avéré être environ 20 % inférieur à celui observé sur les données visuelles correspondant aux mêmes phrases, mais acquises en parole vocalisée (scénario de décodage non-contraint). Bien qu'assez peu significatif compte tenu de la (trop petite) taille de l'ensemble de test, ce resultat semble néanmoins suggérer qu'il est difficile d'utiliser des modèles visuels estimés sur de la « parole vocalisée » pour décoder de la « parole silencieuse ». La mise en œuvre de techniques d'adaptation devra donc être envisagée.

Une seconde expérience, basée sur une tâche de reconnaissance de mots isolés, a été effectuée. Il a été demandé à la locutrice de la base B2, de prononcer quatre fois les 39 éléments qui composent l'alphabet radio international[113], tout d'abord en parole vocalisée, puis en parole silencieuse. Les deux ensembles de données visuelles ainsi constitués ont été analysés à l'aide de l'approche *EigenTongues/EigenLips* (les modalités ultrasonore et vidéo sont ici combinées à l'aide d'une stratégie de fusion au niveau des caractéristiques). La procédure adoptée pour la

[113] Alpha, Bravo, Charlie, Delta, Echo, Foxtrot, Golf, Hotel, India, Juliet, Kilo, Lima, Mike, November, Oscar, Papa, Quebec, Romeo, Sierra, Tango, Uniform, Victor, Whisky, X-ray, Yankee, Zulu, Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Decimal, Stop

reconnaissance est une approche par « comparaison d'exemples » (Boite *et al.*, 2000). Il s'agit de déterminer, parmi tous les exemples du dictionnaire, celui dont la distance par rapport à l'exemple de test est minimale. Cette distance est calculée, dans l'espace des caractéristiques visuelles, à l'aide de l'algorithme DTW. Pour les deux ensembles de données, les trois premières occurrences de chacun des 39 mots sont utilisées pour former le dictionnaire, la quatrième occurrence est utilisée pour le test. Une synthèse des résultats obtenus (taux de mots correctement identifiés) est présentée dans le tableau ci-dessous.

		Exemples du dictionnaire	
		Parole vocalisée	Articulation silencieuse
Exemples de test	Parole vocalisée	89 %	65 %
	Articulation silencieuse	73 %	83 %

Tableau 6.1 : Articulation silencieuse *versus* parole vocalisée, comparaison des performances obtenues sur une tâche de reconnaissance de « mots isolés », en fonction du mode de production considéré pour la constitution de l'ensemble de test et du dictionnaire d'exemples.

On observe que la performance est maximale lorsque le dictionnaire est basé sur le même mode de production que les données de test (83 % contre 73 % dans le cas de l'articulation silencieuse). Ce résultat préliminaire semble confirmer la nécessité de prendre en compte, dans la mise en œuvre de la conversion visuo-acoustique, les spécificités de ce mode de production particulier qu'est l'articulation silencieuse.

Enfin, pour pouvoir réaliser une véritable « interface » de communication, la construction d'un prototype embarqué capable d'effectuer la conversion visuo-acoustique en « temps réel », devra être envisagée. Un tel système pourrait notamment permettre à l'utilisateur d'adapter sa production à l'écoute du signal reconstitué (diffusé par exemple dans une oreillette), optimisant ainsi lui même, la qualité de sa « transmission silencieuse ».

Références

Adjoudani, A., and Benoît, C. (**1996**). "On the Integration of Auditory and Visual Parameters in an HMM-based ASR," in *Speechreading by humans and machines*, edited by D. G. S. a. M. E. Hennecke (Springer, Berlin, Germany), pp. 461-471.

André-Obrecht, R., Jacob, B., and Parlangeau, N. (**1997**). "Audiovisual speech recognition and segmental master slave HMM," in *European Tutorial Workshop on Audio-Visual Speech Processing* (Rhodes, Greece), pp. 49–52.

Anon (**1969**). "IEEE Recommended Practice for Speech Quality Measurements," IEEE Transactions on Audio and Electroacoustics **17**, 225-246.

Aron, M., Ferveur, N., Kerrien, E., Berger, M. O., and Laprie, Y. (**2007**). "Acquisition and synchronization of multimodal articulatory data," in *Interspeech* (Antwerpen, Belgium), pp. 1398-1401.

Bailly, G., Govokhina, O., Breton, G., Elisei, F., and Savariaux, C. (**2008**). "The trainable trajectory formation model TD-HMM parameterized for the LIPS 2008 challenge," in *Interspeech* (Brisbane, Australia), pp. 2318-2321.

Barras, C. (1996). Reconnaissance de la parole continue: Adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés (PhD, Informatique, Université Paris VI, Paris).

Black, A. W., and Lenzo, K. (**2000**). "Building voices in the Festival speech synthesis system."

Boersma, P., and Weenink, D. (**2009**). "Praat: doing phonetics by computer (Version 5.1.13) (Computer program, http://www.fon.hum.uva.nl/praat/)."

Boite, R., Bourlard, H., Dutoit, T., Hancq, J., and Leich, H. (**2000**). *Traitement de la parole* (Presses Polytechniques et Universitaires Romandes, Lausanne).

Bonnin, A., Broussouloux, C., Convard, J.-P., Legmann, P., Seguin, G., and Blery, M. (**2004**). *Echographie* (Masson, Paris).

Bos, J. C., and Tack, D. W. (**2005**). "Audio Display Hardware Investigation for Future Dismounted Soldier Computer Systems (Technical Report from Defence R&D Canada and Human systems Inc)," (Toronto, Canada).

Brand, M., Oliver, N., and Pentland, A. (**1997**). "Coupled hidden Markov models for complex action recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)* (San Juan, Puerto Rico), pp. 994-999.

Bredin, H. (2007). Vérification de l'identité d'un visage parlant. Apport de la mesure de synchronie audiovisuelle face aux tentatives délibérées d'imposture (PhD, Département Traitement du Signal et des Images, Telecom Paris, Paris).

Bregler, C., and Konig, Y. (**1994**). "“Eigenlips” for robust speech recognition," in *ICASSP* (Adelaide, SA, Australia), pp. 669-672.

Browman, C. P., and Goldstein, L. (**1990**). "Gestural specification using dynamically-defined articulatory structures," Journal of Phonetics **18**, 299-320.

Brown, D. R., Ludwig, R., Pelteku, A., Bogdanov, G., and Keenaghan, K. (**2004**). "A novel non-acoustic voiced speech sensor," Measurement science & technology **15**, 1291-1302.

Brumberg, J. S., Nieto-Castanon, A., Guenther, F. H., Bartels, J. L., Wright, E. J., Siebert, S. A., Andreasen, D. S., and Kennedy, P. R. (**2008**). "Methods for construction of a long-term human brain machine interface with the Neurotrophic Electrode," in *Neuroscience Meeting Planner, Program No. 779.5* (Washington, DC).

Burnett, G. C., Gable, T. J., Holzrichter, J. F., and Ng, L. C. (**1997**). "Voiced excitation functions calculated from micropower impulse radar information," The Journal of the Acoustical Society of America **102**, 3168-3168.

Chen, Y., Chu, M., Chang, E., Liu, J., and Liu, R. (**2003**). "Voice Conversion with Smoothed GMM and MAP Adaptation " in *Eurospeech* (Geneva, Switzerland), pp. 2413-2416.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (**1998**). "Active Appearance Models," in *European Conference on Computer Vision* (Freiburg, Germany), pp. 484-498.

Davidson, L. (**2006**). "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance," Journal of the Acoustical Society of America **120**, 407-415.

Del Pozo, A. (2008). Voice Source and Duration Modelling for Voice Conversion and Speech Repair (PhD, Machine Intelligent Laboratory, Cambridge University Engineering Department, Cambridge).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B **39**, 1-38.

Denby, B., Oussar, Y., Dreyfus, G., and Stone, M. (**2006**). "Prospects for a silent speech interface using ultrasound imaging," in *ICASSP* (Toulouse, France), pp. 365-368.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilber, J. M., and Brumberg, J. S. (**2009**). "Silent speech interfaces," Speech Communication in Press (doi:10.1016/j.specom.2009.1008.1002).

Denby, B., and Stone, M. (**2004**). "Speech synthesis from real time ultrasound images of the tongue," in *ICASSP* (Montreal, Canada), pp. 685-688.

Donovan, R. (1996). Trainable Speech Synthesis (PhD, Engineering Department, Machine Intelligence Lab, University of Cambridge, Cambridge).

Doval, B. (1994). Estimation de la fréquence fondamentale des signaux sonores (PhD, Université de Paris 6, Paris).

Dreyfus, G., Samuelides, M., Martinez, J., Gordon, M., Badran, F., Thiria, S., and Hérault, L. (**2008**). *Réseaux de neurones - Méthodologies et applications* (Eyrolles).

Duchnowski, P., Meier, U., and Waibel, A. (**1994**). "See me, hear me: integrating automatic speech recognition and lip-reading," in *ICSLP* (Yokohama, Japan), pp. 547-550.

Dumont, A., and Calbour, C. (**2002**). *Voir la parole* (Masson).

Dupont, S., and Luettin, J. (**2000**). "Audio-Visual Speech Modeling for Continuous Speech Recognition," IEEE Transactions on Multimedia **2**, 141-151.

Efron, B. (**1981**). "Nonparametric Estimates of Standard Error - the Jackknife, the Bootstrap and Other Methods," Biometrika **68**, 589-599.

Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E., and Chapman, P. M. (**2008**). "Development of a (silent) speech recognition system for patients following laryngectomy," Medical Engineering & Physics **30**, 419-425.

Fisher, C. G. (**1968**). "Confusions among visually perceived consonants," Journal of Speech and Hearing Research **11**, 796-804.

Forney, G. D. (**1973**). "Viterbi Algorithm," Proceedings of the IEEE **61**, 268-278.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1993**). "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-100354."

Govokhina, O. (2008). Modèles de génération de trajectoires pour l'animation de visages parlants (PhD, Département de Parole et Cognition du laboratoire GIPSA-Lab INP Grenoble, Grenoble).

Graff, D., Rosenfeld, R., and Paul, D. (**1995**). "CSR-III Text," Linguistic Data Consortium, Philadelphia.

Gravier, G., Potamianos, G., and Neti, C. (**2002**). "Asynchrony modeling for audio-visual speech recognition," in *International Conference on Human Language Technology Research* (San Diego, California).

Gray, H., Goss, C., M. (**1973**). *Anatomy of the human body* (LEA & FEBIGER, Philadelphia).

Hasegawa, T., and Ohtani, K. (**1992**). "Oral image to voice converter-image input microphone," in *ICCS/ISITA* (Singapore), pp. 617-620.

Hayakawa, K., Takeda, S., Kawabe, K., and Shimura, T. (**1989**). "Acoustic characteristics of PVA gel," in *IEEE Ultrasonics Symposium* (Montréal, Canada), pp. 969-972.

Heracleous, P., Nakajima, Y., Lee, A., Saruwatari, H., and Shikano, K. (**2003**). "Accurate hidden Markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation," in *IEEE Workshop on Automatic Speech Recognition and Understanding* (St. Thomas, Virgin Islands, USA), pp. 73-76.

Hieronymus, J. (**1993**). "ASCII Phonetic Symbols for the World's Languages: Worldbet," AT&T Bell Laboratories, Technical Memo **23**.

Hogg, R. V., Tanis, E.A., (**1996**). *Probability and statistical inference (5th ed.)* (Prentice Hall College, NJ).

Hueber, T. (2006). Synthèse de la parole à partir d'imagerie ultrasonore et optique de l'appareil vocal (Thèse de Master, INSA Lyon, Lyon).

Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., and Stone, M. (**2007a**). "Eigentongue feature extraction for an ultrasound-based silent speech interface," in *ICASSP* (Honolulu, USA), pp. 1245-1248.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (**2007b**). "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips," in *Interspeech* (Antwerp, Belgium), pp. 658-661.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (**2008a**). "Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface," in *Interspeech* (Brisbane, Australia), pp. 2032-2035.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (**2008b**). "Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips," in *Interspeech* (Brisbane, Australie), pp. 2028-2031.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (**2009a**). "Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface," in *Interspeech* (Brighton, UK), pp. 640-643.

Hueber, T., Chollet, G., Denby, B., Stone, A., and Zouari, L. (**2007c**). "Ouisper: Corpus Based Synthesis Driven by Articulatory Data," in *International Congress of Phonetic Sciences* (Saarbrücken, Germany), pp. 2193-2196.

Hueber, T., Chollet, G., Denby, B., and Stone, M. (**2008c**). "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," in *International Seminar on Speech Production* (Strasbourg, France), pp. 365-369.

Hueber, T., Dubois, R., Roussel, P., Denby, B., and Dreyfus, G. (**2009b**). "Dispositif de reconstitution de la parole par sondage ultrasonore de l'appareil phonatoire (Brevet, Numéro de dépot 09 04444) " (France).

Hunt, A. J., and Black, A. W. (**1996**). "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP* (Atlanta, USA), pp. 373-376.

Imai, S. (**1983**). "Cepstral analysis synthesis on the mel frequency scale," in *ICASSP* (Boston, USA), pp. 93-96.

Imai, S., and Furuichi, C. (**1988**). "Unbiased estimator of log spectrum and its application to speech signal processing," in *EURASIP* (Grenoble, France), pp. 203-206.

Imai, S., Sumita, K., and Furuichi, C. (**1983**). "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," Electronics and Communications in Japan (Part I: Communications) **66**, 10-18.

Itakura, F. (**1975**). "Line spectrum representation of linear predictor coefficients of speech signals," Journal of the Acoustical Society of America **57**, 35-35.

Jorgensen, C., Lee, D. D., and Agabon, S. (**2003**). "Sub auditory speech recognition based on EMG signals," in *International Joint Conference on Neural Networks*, pp. 3128-3133.

Jou, S.-C. S., Schultz, T., and Waibel, A. (**2007**). "Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture," in *ICASSP* (Honolulu, USA), pp. 401-404.

Jou, S. C., Maier-Hein, L., Schultz, T., and Waibel, A. (**2006**). "Articulatory feature classification using surface electromyography," in *ICASSP*, pp. 605-608.

Kain, A. (2001). High-resolution voice transformation (PhD, OGI School of Science & Engineering, Oregon Health & Science University).

Kröger, B. J., and Birkholz, P. (**2007**). "A Gesture-Based Concept for Speech Movement Control in Articulatory Speech Synthesis " in *Verbal and Nonverbal Communication Behaviours*, edited by S. B. Heidelberg (Berlin), pp. 174-189.

Krstolovic, S. (2001). Speech Analysis with Production Constraints (PhD, Département d'Electricité, Ecole Polytechnique Fédérale de Lausanne, Lausanne).

Le Huche, F., Allali, A. (**2001**). *La Voix. Anatomie et physiologie des organes de la voix et de la parole* (Masson, Paris).

Livescu, K. (2005). Feature-based pronunciation modeling for automatic speech recognition (PhD, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science., Massachusetts Institute of Technology).

Maier-Hein, L., Metze, F., Schultz, T., and Waibel, A. (**2005**). "Session independent non-audible speech recognition using surface electromyography," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 331-336.

Mallat, S. (**2001**). *Une exploration des signaux en ondelettes* (Broché).

Matsui, K., Hara, N. (**1999**). "Enhancement of esophageal speech using formant synthesis," in *ICASSP* (Phoenix, USA), pp. 81-84.

Matthews, I., Potamianos, G., and Neti, C. (**2001**). "A Comparison Of Model And Transform-Based Visual Features For Audio-Visual LVCSR," in *IEEE International Conference on Multimedia and Expo* (Tokyo, Japan), pp. 210-214.

McGurk, H. a. M., John (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.

Mielke, J., Baker, A., Archangeli, D., and Racy, S. (**2005**). "Palatron: A Technique for Aligning Ultrasound Images of the Tongue and Palate," Coyote Papers **14**, 97-108.

Miller, A. (**2008**). "Click Cavity Formation and Dissolution in IsiXhosa: Viewing Clicks with High-Speed Ultrasound," in *International Seminar on Speech Production* (Strasbourg, France), pp. 137-140.

Montacié, C., and Chollet, G. (**1987**). "Systèmes de référence pour l'évaluation d'applications et la caractérisation de bases de données en reconnaissance automatique de la parole," in *JEP* (Hammamet, Tunisie), pp. 323-326.

Nakajima, Y., Kashioka, H., Campbell, N., and Shikano, K. (**2006**). "Non-audible murmur (NAM) recognition," IEICE Transactions on Information and Systems **E89d**, 1-8.

Nakajima, Y., Kashioka, H., Shikano, K., and Campbell, N. (**2003**). "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *ICASSP* (Hong Kong), pp. 708-711.

Nakamura, K., Toda, T., Saruwatari, H., and Shikano, K. (**2007**). "Impact of Various Small Sound Source Signals on Voice Conversion Accuracy in Speech Communication Aid for Laryngectomees," in *Interspeech* (Anvers, Belgium), pp. 2517-2520.

Payri, B. (2000). Perception de la voix parlée : Cohérence du timbre du locuteur (Phd, LIMSI, Université de Paris 11, Paris).

Perona, P., and Malik, J. (**1990**). "Scale-Space and Edge-Detection Using Anisotropic Diffusion," Ieee Transactions on Pattern Analysis and Machine Intelligence **12**, 629-639.

Petajan, E. D. (1984). Automatic lipreading to enhance speech recognition (speech reading) (PhD, University of Illinois, Champaign, IL, USA).

Potamianos, G., Graf, H. P., and Cosatto, E. (**1998**). "An Image Transform Approach for HMM Based Automatic Lipreading," in *ICIP* (Chicago, USA), pp. 173-177.

Potamianos, G., and Neti, C. (**2001**). "Improved ROI and within frame discriminant features for lipreading," in *ICIP* (Thessaloniki, Greece), pp. 250-253.

Potamianos, G., Neti, C., Luettin, J., and Matthews, I. (**2004**). "Audio-visual automatic speech recognition: An overview " in *Issues in Visual and Audio-Visual Speech Processing*, edited by E. V.-B. G. Bailly, and P. Perrier (MIT Press).

Potard, B. (2008). Inversion acoustique-articulatoire avec contraintes. (PhD, LORIA, Université Henri Poincaré - Nancy 1, Nancy).

Pylkkönen, J., and Kurimo, M. (**2004**). "Duration modeling techniques for continuous speech recognition," in *Interspeech* (Jeju Island, Korea), pp. 385-388.

Quatieri, T. F., Brady, K., Messing, D., Campbell, J. P., Campbell, W. M., Brandstein, M. S., Weinstein, C. J., Tardelli, J. D., and Gatewood, P. D. (**2006**). "Exploiting non-acoustic sensors for speech encoding," IEEE Transactions on Audio Speech and Language Processing **14**, 533-544.

Rabiner, L., and Juang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice Hall PTR).

Rao, C. R. (**2002**). *Linear Statistical Inference and its Applications, 2nd Edition* (Wiley).

Rogozan, A., and Deléglise, P. (**1998**). "Adaptive fusion of acoustic and visual sources for automatic speech recognition," Speech Communication **26**, 149-161.

Stone, M. (**2005**). "A guide to analysing tongue motion from ultrasound images," Clinical Linguistics & Phonetics **19**, 455 - 501.

Stone, M., and Davis, E. P. (**1995**). "A head and transducer support system for making ultrasound images of tongue jaw movement," Journal of the Acoustical Society of America **98**, 3107-3112.

Stuart, B. (**1939**). "Throat microphone - Patent 2165124," (United States, United States).

Stylianou, I. (1990). Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification (PhD, Signal et Image, ENST Paris, Paris).

Sugie, N., and Tsunoda, K. (**1985**). "A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production," IEEE Transactions on Biomedical Engineering **32**, 485-490.

Suppes, P., Han, B., and Lu, Z. L. (**1998**). "Brain-wave recognition of sentences," Proceedings of the National Academy of Sciences (USA) **95**, 15861-15866.

Suppes, P., Lu, Z. L., and Han, B. (**1997**). "Brain wave recognition of words," Proceedings of the National Academy of Sciences (USA) **94**, 14965-14969.

Tauber, C. (2005). Filtrage anisotrope robuste et segmentation par B-spline snake: application aux images echographique (PhD, INP Toulouse, Toulouse).

Toda, T., Black, A. W., and Tokuda, K. (**2008**). "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," Speech Communication **50**, 215-227.

Toda, T., Saruwatari, H., and Shikano, K. (**2001**). "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *ICASSP* (Salt Lake City, UT, USA), pp. 841 - 844.

Toda, T., and Tokuda, K. (**2007**). "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," IEICE Transactions on Information and Systems **E90-D**, 816-824.

Toda, T., and Tomoki, S. (**2005**). "NAM-to-Speech Conversion with Gaussian Mixture Models," in *Interspeech* (Lisbon, Portugal), pp. 1957-1960.

Tokuda, K., Mausko, T., Miyazaki, N., and Kobayashi, T. (**2002**). "Multi-space probability distribution HMM," IEICE Transactions on Information and Systems **E85-D**, 455-464.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (**2000**). "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP* (Istanbul, Turkey), pp. 1315-1318.

Tran, V.-A., Bailly, G., Loevenbruck, H., and C., J. (**2008a**). "Improvement to a NAM captured whisper-to-speech system," in *Interspeech* (Brisbane, Australia), pp. 1465-1498.

Tran, V.-A., Bailly, G., Loevenbruck, H., and Toda, T. (**2008b**). "Predicting F0 and voicing from NAM-captured whispered speech," in *Speech Prosody* (Campinas, Brazil), pp. 107-110.

Turk, M., and Pentland, A. (**1991**). "Eigenfaces for Recognition," Journal of Cognitive Neuroscience **3**, 71-86.

Vanegas, O., Tanaka, A., Tokuda, K., and Kitamura, T. (**1998**). "HMM-based visual speech recognition using intensity and location normalization," in *ICSLP* (Sydney, Australia), pp. 289–292.

Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T., and Waibel, A. (**2006**). "Sub-word unit based non-audible speech recognition using surface electromyography," in *Interspeech* (Pittsburgh, USA), pp. 1487-1490.

Wester, M. (2006). Unspoken speech - speech recognition based on surface electroencephalography (PhD, Institut für Theoretische Informatik, Universität Karlsruhe, Karlsruhe, Germany).

Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-Lehouillier, H., Vatikiotis-Bateson, E., and Hailey, D. S. (**2005**). "The Haskins optically corrected ultrasound system (HOCUS)," Journal of Speech, Language, and Hearing Research. **48**, 543-553.

Wrench, A., Scobbie, J., and Linden, M. (**2007**). "Evaluation of a helmet to hold an ultrasound probe," in *Ultrafest IV* (New York, USA).

Wrench, A. A., and Scobbie, J. M. (**2006**). "Spatio-temporal inaccuracies of video-based ultrasound images of the tongue," in *ISSP* (Ubatuba, Bresil), pp. 451-458.

Yingyoung, Q. (**1990**). " Replacing tracheoesophageal voicing sources using LPC synthesis," Journal of the Acoustical Society of America **88**, 1228-1235.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (**1998**). "Duration Modeling For HMM-Based Speech Synthesis," in *Interspeech* (Sydney, Australia), pp. 29-32.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. (**2005**). "The HTK Book," (http://htk.eng.cam.ac.uk/).

Young, S. J., Odell, J. J., Woodland, P. C. (**1994**). "Tree-based state tying for high accuracy acoustic modeling," in *Workshop on Human Language Technology* (Plainsboro, NJ), pp. 307-312.

Young, S. J., Russell, N. H., and Thornton, J. H. S. (**1989**). "Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems (technical report)," (Cambridge University Engineering Department).

Yu, Y. J., and Acton, S. T. (**2002**). "Speckle reducing anisotropic diffusion," IEEE Transactions on Image Processing **11**, 1260-1270.

Annexe A

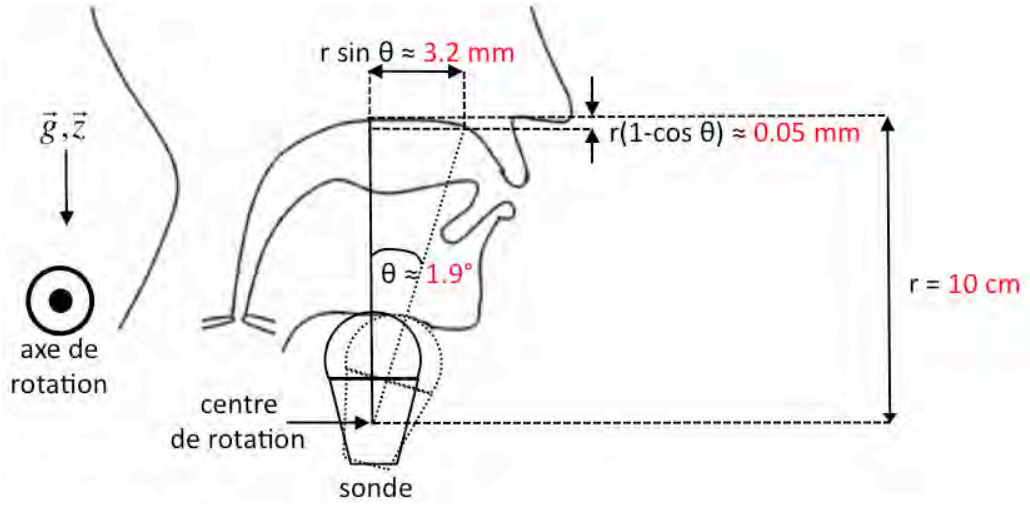
Précision de la méthode de repositionnement de la sonde ultrasonore basée sur l'utilisation de capteurs inertiels

Les deux capteurs utilisés pour le repositionnement de la sonde ultrasonore par rapport à la tête du locuteur (entre deux sessions d'acquisition de données) sont des accéléromètres 3-axes ADXL330 (*Analog Devices*). Chaque capteur est intégré sur une carte de développement conçue par la société *Phidgets* (référence 1059), qui facilite notamment l'interfaçage avec un micro-ordinateur (sur bus USB). D'après la documentation fournie par le constructeur, l'écart-type du bruit de mesure, noté σ_g, est de 1.9 mG pour les axes X et Y et de 2.9 mG pour l'axe Z (axe colinéaire au vecteur gravitationnel). Pour une utilisation en mesure d'inclinaisons, les erreurs angulaires maximales correspondantes, notées $\Delta\theta$, sont les suivantes :

	Erreurs angulaires maximales (en degré)
Axe X & Axe Y	$\Delta\theta_x = \Delta\theta_y = 3.5°$
Axe Z	$\Delta\theta_z = 4.4°$

Afin de diminuer l'écart-type de ce bruit, nous moyennons les mesures d'accélération sur une fenêtre glissante de 30 échantillons, soit 0.5 secondes, le capteur fournissant une mesure toutes les 16 ms (60 Hz). En notant $\hat{\sigma}_g$ l'écart-type du bruit de mesure résultant, on obtient $\hat{\sigma}_g = \sigma_g / \sqrt{30}$, soit une erreur angulaire maximale égale à 1.5° pour les axes X et Y, et 1.9° pour l'axe Z.

L'impact de ce bruit de mesure sur le repositionnement de la sonde par rapport à la tête (référencée par la position du palais) est illustré par la figure ci-dessous (illustration dans le cas où un seul accéléromètre est utilisé) :



En considérant une distance « sonde-palais » égale à 10 cm, l'erreur de repositionnement est de l'ordre de 0.05 mm dans l'axe « sonde-palais », et de 3.2 mm dans l'axe qui lui est perpendiculaire dans le plan sagittal médian.

Annexe B - Publications

Revues

[1] Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips", Speech Communication, à paraître.

[2] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S. "Silent speech interfaces", Speech Communication, à paraître.

Brevet

[3] Hueber, T., Dubois, R., Roussel, P., Denby, B., and Dreyfus, G. (2009b). "Dispositif de reconstitution de la parole par sondage ultrasonore de l'appareil phonatoire (Brevet, Numéro de dépot 09 04444)" (France).

Chapitres de livres

[4] Hueber, T., Denby, B. (2009). "Analyse du conduit vocal par imagerie ultrasonore", L'imagerie médicale pour l'étude de la parole, Alain Marchal, Christian Cavé, Traité Cognition et Traitement de l'Information, IC2, Hermes Science.

[5] Chollet, G., Landais, R., Hueber, T., Bredin, H., Mokbel, C., Perrot, P., Zouari, L. (2007). "Some Experiments in Audio-Visual Speech Processing", Advances in Nonlinear Speech Processing, vol 4885, Springer, pp. 28-56.

Conférences internationales avec actes et comité de lecture

[6] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., and Stone, M. (2007). "Eigentongue feature extraction for an ultrasound-based silent speech interface," in ICASSP (Honolulu, USA), pp. 1245-1248.

[7] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2007). "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips," in Interspeech (Antwerp, Belgium), pp. 658-661.

[8] Hueber, T., Chollet, G., Denby, B., Stone, A., and Zouari, L. (2007). "Ouisper: Corpus Based Synthesis Driven by Articulatory Data," in International Congress of Phonetic Sciences (Saarbrücken, Germany), pp. 2193-2196.

[9] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2008). "Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface," in Interspeech (Brisbane, Australia), pp. 2032-2035.

[10] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2008). "Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips," in Interspeech (Brisbane, Australie), pp. 2028-2031.

[11] Hueber, T., Chollet, G., Denby, B., and Stone, M. (2008). "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," in International Seminar on Speech Production (Strasbourg, France), pp. 365-369.

[12] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2009). "Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface," in Interspeech (Brighton, UK), pp. 640-643.

Les articles [6] à [12] sont joints à ce manuscrit.

EIGENTONGUE FEATURE EXTRACTION FOR AN ULTRASOUND-BASED SILENT SPEECH INTERFACE

T.Hueber[1,3], G.Aversano[3], G.Chollet[3], B.Denby[1,2], G.Dreyfus[1], Y.Oussar[1], P.Roussel[1], M.Stone[4]

[1]Laboratoire d'Electronique, Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI-Paristech), 10 rue Vauquelin, 75231 Paris Cedex 05 France ; thomas.hueber@gmail.com
[2]Université Pierre et Marie Curie – Paris VI, B.C. 252, 4 place Jussieu, 75252 Paris Cedex 05, France ; denby@ieee.org
[3]Laboratoire Traitement et Communication de l'Information, Ecole Nationale Supérieure des Télécommunications (ENST-Paristech), 46 rue Barrault, 75634 Paris Cedex 13
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore MD 21201 USA

ABSTRACT

The article compares two approaches to the description of ultrasound vocal tract images for application in a "silent speech interface," one based on tongue contour modeling, and a second, global coding approach in which images are projected onto a feature space of *Eigentongues*. A curvature-based lip profile feature extraction method is also presented. Extracted visual features are input to a neural network which learns the relation between the vocal tract configuration and line spectrum frequencies (LSF) contained in a one-hour speech corpus. An examination of the quality of LSF's derived from the two approaches demonstrates that the eigentongues approach has a more efficient implementation and provides superior results based on a normalized mean squared error criterion.

Index Terms— image processing, speech synthesis, neural network applications, communication systems, silent speech interface

1. INTRODUCTION

There has been significant interest recently in the notion of a "silent speech interface (SSI)" – a portable device used as an alternative to tracheo-oesophageal speech for larynx cancer patients, for situations where silence must be maintained, or for voice communication in noisy environments. Approaches based on electromyography [1], a non-audible murmur microphone [2], and ultrasound and optical imagery ([3], [4]) have appeared in the literature.

We present here results of a visuo-acoustic SSI study based on a one-hour corpus comprising ultrasound and optical imagery of the vocal tract. The use of a corpus of this size – which was motivated by the desire to interface to a concatenative speech synthesizer – has led to the development of robust feature extraction techniques in order to accommodate the wide variety of articulator configurations appearing in the corpus. In particular, an *Eigentongues* approach has been introduced in order to address the problem of ultrasound frames in which the tongue images poorly. Section 2 of the article details data acquisition and ultrasound image preprocessing, while section 3 describes the feature extraction techniques used in the image (ultrasound and optical) and speech signal analyses. Modeling of the link between visual and acoustic features is introduced in section 4, along with experimental results.

2. DATA ACQUISITION AND PREPROCESSING

2.1. Data acquisition

Data were taken using a 30 Hz ultrasound machine and the Vocal Tract Visualization Lab HATS system [5], which maintains acoustic contact between the throat and the ultrasound transducer during speech. A lip profile image is embedded into the ultrasound image, as shown in figure 1.

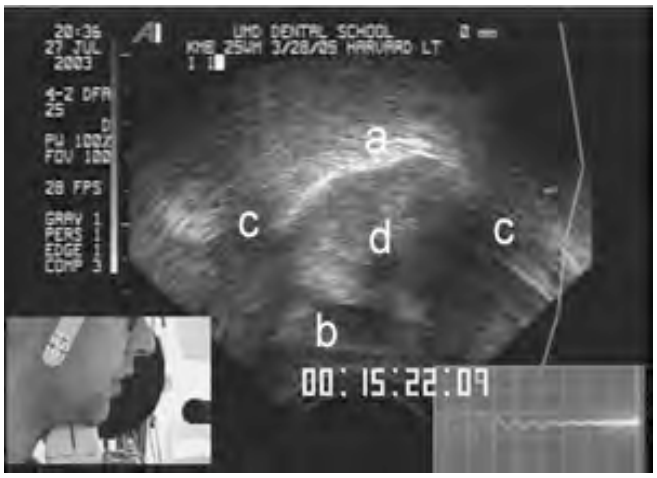


Figure 1. Example of an ultrasound vocal tract image with embedded lip profile : (a) tongue surface ; (b) hyoid bone ; (c) hyoid and mandible acoustic shadows ; (d) muscle, fat and connective tissue within the tongue.

The speech dataset used consists of 720 sentences, organized in 10 lists, from the IEEE/Harvard corpus [6], spoken by a male native American English speaker. The IEEE sentences were chosen because they are constructed to have roughly equal intelligibility across lists and all have approximately the same duration, number of syllables, grammatical structure and intonation. After cleaning the database, the resulting speech was stored as 72473 JPEG frames and 720 WAV audio files sampled at 11025 Hz.

2.2. Ultrasound image preprocessing

In order to select a region of interest, the ultrasound images are first reduced to a 50 (radial angle) by 50 (azimuthal angle) semi-polar grid. To decrease the effects of speckle, the reduced images are filtered using the anisotropic diffusion filter proposed by Yu [7]. This iterative process introduces intra-region smoothing while inhibiting inter-region smoothing [8], *via* a local coefficient of variation [9], so that speckle is removed without destroying important image features. A typical result after these two preprocessing steps is illustrated in figure 2(a).

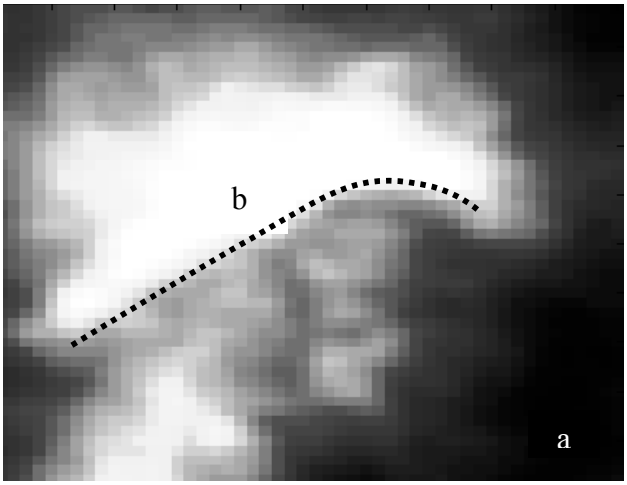


Figure 2. Reduced and filtered ultrasound image (a) and tongue surface contour fit by a 4th order spline (b)

3. FEATURE EXTRACTION

3.1. Ultrasound image feature extraction

3.1.1. Tongue contour extraction

As in [3] and [4], our first approach considers the tongue surface to be the only ultrasound image information relevant to the prediction of speech characteristics. Tongue contour candidate points are defined as maxima of the smoothed vertical intensity gradient. Then, in the present work, a Least Median Square (LMS, [10])-based spline interpolation method, tolerating up to 50% outlier points, is used in order to retain only relevant tongue contour candidates; this is an improvement over the contour extraction method implemented in [3] and [4].

A typical tongue contour is shown in figure 2(b). Due to refraction, however, the tongue surface will be poorly imaged when the tongue surface is at angles nearly parallel to the ultrasound beam, as in the case of the phoneme /i/ for example. The contour extraction described previously fails in such frames – which are found to constitute some 15 % of our database – since the tongue surface is simply no longer visible in them. These "outlier frames" are detected automatically using the area of the convex hull of intensity gradient maxima. Below, we present a more global feature extraction approach which provides a solution to the missing contour problem.

3.1.2. Eigentongue feature extraction

The second approach features the use of Principal Component Analysis (PCA), or Karhunen-Loève expansion, for describing the ultrasound images. The first step is to create a finite set of orthogonal images, which constitutes, up to a certain accuracy, a subspace for the representation of all likely tongue configurations. These images are referred to as *Eigentongues*, a term inspired by the *Eigenface* method of Turk and Pentland [11]. The first three *Eigentongues*, obtained after a PCA on 1000 reduced and filtered ultrasound images, are shown in figure 3.

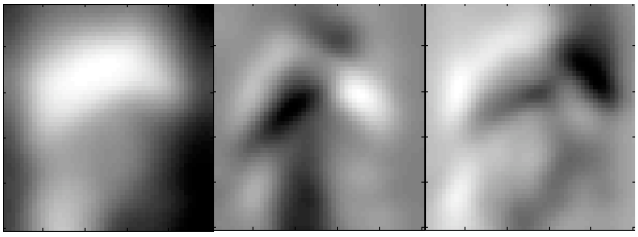


Figure 3. The first three *Eigentongues* (1-3 from left to right)

Once the set of *Eigentongues* has been created, the images of subsequent tongue configurations can be represented quite compactly in terms of their projections onto the set of *Eigentongues*, as shown in figure 4.

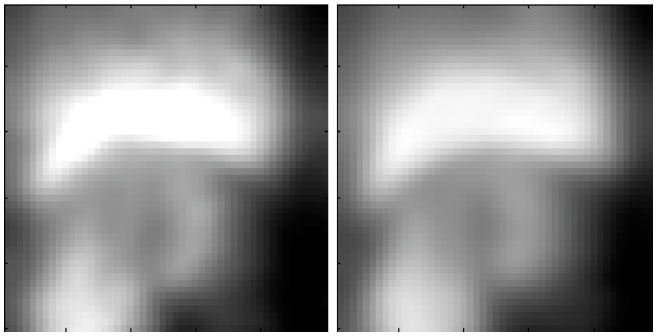


Figure 4. A reduced ultrasound image (left) and its re-synthesis (right) using 20 *Eigentongue* components

The *Eigentongue* components encode the maximum amount of relevant information in the images, mainly tongue position, of course, but also other structures such as the hyoid bone, muscles, etc.

3.2. Optical image feature extraction

The optical image feature extraction consists of a description of the lip profile. We propose an algorithm based on the observation of Attneave that information along a visual contour is concentrated in regions of high curvature, rather than distributed uniformly [12]. The lip edge profile is easily extracted using the Sobel method. The curvature of this two-dimensional curve is then computed using the Turning Angle introduced by Feldman [13]. Upper/lower lip and commissure positions coincide with extrema of the curvature, as shown as figure 5, while the values of the curvature at these points give local lip shape information.

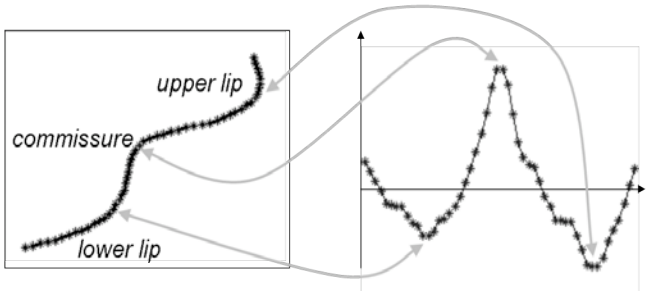


Figure 5. Lip profile description using curvature computation (left: lip contour; right: curvature of lip contour)

3.3. Speech signal description

For each 33 ms audio frame (dictated by the 30 Hz ultrasound rate), twelve LSF's are calculated using a pre-accentuation filter, linear predictive coding and a Hann window with a half-frame overlap. The robustness of LSF coefficients is known to assure the stability of the LPC filter [14]. A voiced/unvoiced flag and fundamental frequency (for voiced frames) are also computed, using a simple autocorrelation-based method. These last two features are not used in the visuo-acoustic modeling which follows, but allow a qualitative, audible comparison of our different results, if desired, *via* LPC synthesis using the predicted autoregressive filter coefficients and a realistic excitation function.

4. VISUO-ACOUSTIC MODELING

Our first feature extraction method, described in sections 3.2 and 3.1.1, provides 15 features per frame, including 9 for the lips (position and curvature of upper/lower lips and commissure) and 6 for the tongue (4^{th} order spline coefficients and interval of definition). The second, *Eigentongue* method gives 29 features per frame, the first 20 *Eigentongue* components plus lip features. A multilayer perceptron (MLP) is used to perform the mapping between these input visual features and the 12 LSF's [15]. A separate network is used for each LSF in order to limit the number of adjustable parameters in the model. A total of 71502 frames are used for training, with an independent set of 971 frames for model selection and validation. We now compare the LSF prediction obtained from the two methods. Because each LSF is defined upon its own interval, we introduce a normalized measure of the quality of the prediction α, along with an estimate of its standard deviation ε [16]:

$$\alpha = \frac{100}{\sqrt{N}} \cdot \frac{\sqrt{\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2}}{|\, y_{\max} - y_{\min} \,|} \qquad \varepsilon = \alpha \sqrt{\frac{1}{2N}}$$

where N is the number of examples in the validation database, y are the true LSF's, and $\tilde{y}$ the predicted LSF's.

4.1. Comparing Contour and *Eigentongue* approaches

For the tongue contour method, in order to obtain reasonable training results, the "outlier frames" for which the automatic contour extraction algorithm (described in section 3.1.1) failed were removed from the training set. As the *Eigentongue* feature extraction approach does not restrict relevant information to a specific structure, no outlier is generated when that structure is not imaged, and thus all frames may be used in the visuo-acoustic modeling with this method. Columns 1 and 2 of Table 1 compare the results of the two approaches.

LSF number	Tongue Contour	*Eigentongue*	*Eigentongue* + history
	Parameter α (% of total dynamic range)		
1	18.7 ± 0.4	16.9 ± 0.4	16.4 ± 0.4
2	16.1 ± 0.4	14.4 ± 0.3	13.7 ± 0.3
3	14.3 ± 0.3	12.4 ± 0.3	12.3 ± 0.3
4	13.1 ± 0.3	11.8 ± 0.3	10.8 ± 0.2
5	14.2 ± 0.3	11.5 ± 0.3	11.9 ± 0.3
6	13.1 ± 0.3	11.8 ± 0.3	10.6 ± 0.2
7	15.7 ± 0.4	13.7 ± 0.3	12.6 ± 0.3
8	13.1 ± 0.3	11.8 ± 0.3	12.1 ± 0.3
9	14.6 ± 0.3	12.8 ± 0.3	12.4 ± 0.3
10	12.9 ± 0.3	11.2 ± 0.2	11.2 ± 0.2
11	14.5 ± 0.3	13.7 ± 0.3	11.4 ± 0.2
12	16.3 ± 0.4	14.5 ± 0.3	14.4 ± 0.3

Table 1. Comparison of tongue contour based modeling and *Eigentongue* based modeling. Quoted errors, ε, are estimates of the standard deviation of α using a Gaussian assumption

The table shows that LSF's 4, 6, 8 and 10 are the best predicted by tongue contour and lip profile features, and that using *Eigentongues* provides an improvement in overall prediction quality which is small, but statistically

significant. The filtering step described in section 2.2 is in fact not essential for the *Eigentongue* feature extraction, as image regions of high intensity variability will be associated with the higher order *Eigentongues*, which are not used. Similar results are obtained using *Eigentongues* obtained from unfiltered images.

4.2. Introducing 'history' into the input variables

The use of *Eigentongues* allows all of the video frames to participate in the training, which is not the case for the contour method due to the missing frames. We can then in a simple way take account of the intrinsically dynamic nature of speech production in our visuo-acoustic modeling by providing the training algorithm, at frame *n*, with the *Eigentongue* and lip variables of frames *n*-1 and *n*-2, as well. An additional small improvement in the prediction of LSF's 2, 4, 6, 7 and 11 is seen, as compared to the static modeling.

5. CONCLUSION AND PERSPECTIVES

A new turning-angle algorithm for the description of lip profiles has been introduced, which, because curvature-based, should hopefully make the method robust against the variability of lip shapes between speakers. Two methods for feature extraction from ultrasound images have been presented and compared. The visuo-acoustic modeling with *Eigentongues* gives better results than those obtained using tongue contours as input. The *Eigentongue* method is easier to implement, appears to take more information into account, and is not prone to failures due to instrumental effects, thus allowing the dynamic nature of speech to be taken into account in a natural way. It could be interesting, however, in future work, to combine the two approaches in the context of active appearance models [17]. The model we propose is at present able to predict an acoustical description of speech with errors ranging from 11% to 16%. Whether this performance is adequate for application in an SSI will only become apparent once a concatenative speech synthesis model using our predicted quantities as inputs has been experimented. The elaboration of such a test, as well as the use of alternative dynamic process modeling techniques (Hidden Markov Models, Time Delay Neural Networks [18]) are currently underway.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge useful discussions with Isabelle Bloch. This work was supported in part by CNFM (Comité National de Formation en Microélectronique).

7. REFERENCES

[1] C. Jorgensen, D. D. Lee, and S. Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals," *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, pp. 3128-3133, 2003.

[2] Y. Nakajima, P. Heracleous, H. Saruwatari and K. Shikano, "A Tissue-conductive Acoustic Sensor Applied in Speech Recognition for Privacy," *Smart Objects & Ambient Intelligences Oc-EUSAI 2005*, pp. 93-98, 2005.

[3] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone, "Prospects for a Silent Speech Interface Using Ultrasound Imaging," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

[4] B. Denby and M. Stone, "Speech Synthesis from Real Time Ultrasound Images of the Tongue," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.

[5] M. Stone, "A Guide to Analysing Tongue Motion from Ultrasound Images," *Clinical Linguistics and Phonetics*, pp. 359-366, 2003.

[6] IEEE, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225-246, 1969.

[7] Y. Yu and S. T. Acton, "Speckle Reducing Anisotropic Diffusion," *IEEE Transactions on Image Processing*, vol. 11, pp. 1260-1270, 2002.

[8] P. Perona and J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 629-639, 1990.

[9] J. Lee, "Digital Image Enhancement and Noise Filtering by Use of Local Statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, pp. 165-168, 1980.

[10] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, New York, USA, John Wiley & Sons, Inc., 1987.

[11] M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586-591, 1991.

[12] F. Attneave, "Some Informational Aspects of Visual Perception," *Psych. Review*, vol. 61, pp. 183-193, 1954.

[13] J. Feldman and M. Singh, "Information Along Contours and Object Boundaries," *Psych. Review*, vol. 112, pp. 243-252, 2005.

[14] G. Kang and L. Fransen, "Application of Line-Spectrum Pairs to Low-Bit-Rate Speech Encoders," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tampa, USA, 1985.

[15] G. Dreyfus, *Neural Networks: Methodology and Applications*, Springer, New York, 2005.

[16] H. Cramér, *Mathematical Methods of Statistics,* Princeton University Press, Princeton, 1999.

[17] M. B. Stegmann, R. Fisker, B. K. Ersbøll, H. H. Thodberg, L. Hyldstrup, "Active Appearance Models: Theory and Cases," *Proc. 9th Danish Conference on Pattern Recognition and Image Analysis*, vol. 1, pp. 49-57, AUC Press, 2000.

[18] T. Hanazawa, A. Waibel, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition Using Time Delay Neural Networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 328-339, 1989.

Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips

Thomas Hueber [1,3], *Gérard Chollet* [3], *Bruce Denby* [1,2], *Gérard Dreyfus* [1], *Maureen Stone* [4]

[1]Laboratoire d'Electronique, Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI-Paristech), 10 rue Vauquelin, 75231 Paris Cedex 05 France
[2]Université Pierre et Marie Curie – Paris VI, 4 place Jussieu, 75252 Paris Cedex 05 France
[3]Laboratoire Traitement et Communication de l'Information, Ecole Nationale Supérieure des Télécommunications (ENST-Paristech), 46 rue Barrault, 75634 Paris Cedex 13 France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore, MD 21201 USA

hueber@ieee.org, gerard.chollet@tsi.enst.fr, denby@ieee.org, gerard.dreyfus@espci.fr, mstone@umaryland.edu

Abstract

The article describes a video-only speech recognition system for a "silent speech interface" application, using ultrasound and optical images of the voice organ. A one-hour audio-visual speech corpus was phonetically labeled using an automatic speech alignment procedure and robust visual feature extraction techniques. HMM-based stochastic models were estimated separately on the visual and acoustic corpus. The performance of the visual speech recognition system is compared to a traditional acoustic-based recognizer.

Index Terms: speech recognition, audio-visual speech description, silent speech interface, machine learning

1. Introduction

In recent years, several systems using articulatory data to synthesize speech in real time have been described in the literature. These data may be derived from EMG/EPG measures [1], from a "non audible murmur microphone" signal (NAM [2]) or, in our case, from ultrasound and optical images of the voice organ [3]. Such a synthesizer, driven only by articulatory data, may be qualified as a "silent speech interface" (SSI), in that it could be used as an alternative to tracheo-oesophageal speech or electrolarynx for laryngeal cancer patients, in situations where silence must be maintained, or for voice communication in noisy environments.

Laptop based high performance ultrasound medical imaging systems are available today, allowing to envisage, for example, an ultrasound-based SSI telephone for home use. Ultimately, a wearable real-time SSI system with an embedded ultrasound transducer and camera, piloted by a personal digital assistant (PDA), should be possible.

In [4], a static neural network was used to learn the "visuo-acoustic" mapping between the ultrasound tongue and optical lip images (called the "video" data) and a set of Line Spectrum Frequencies (LSF). That study demonstrated the relevance of visual features for describing the voice organ but permitted only LPC-based synthesis without an appropriate excitation signal.

Here, we propose visual speech recognition as a first step towards corpus-based silent speech synthesis, which furthermore allows the possibility of introducing linguistic constraints in our analysis. In the context of "silent" speech, the present article focuses on speech recognition from video-only data; the use of an ultrasound video stream for denoising a corrupted audio stream will be addressed in future work. Our approach is based on building a corpus associating video-extracted visual feature sequences to phoneme labels. HMM-based stochastic models trained on this database are then used to predict target phonetic sequences. An overview of the system is given in figure 1.

Section 2 of the article details data acquisition and ultrasound image preprocessing, while speech segmentation techniques appear in section 3. The visual feature extraction is discussed in section 4. Procedures involved in speech recognition from video-only data are presented in section 5 and our results detailed in section 6. A discussion of the results and some ideas for future improvements appear in section 7.

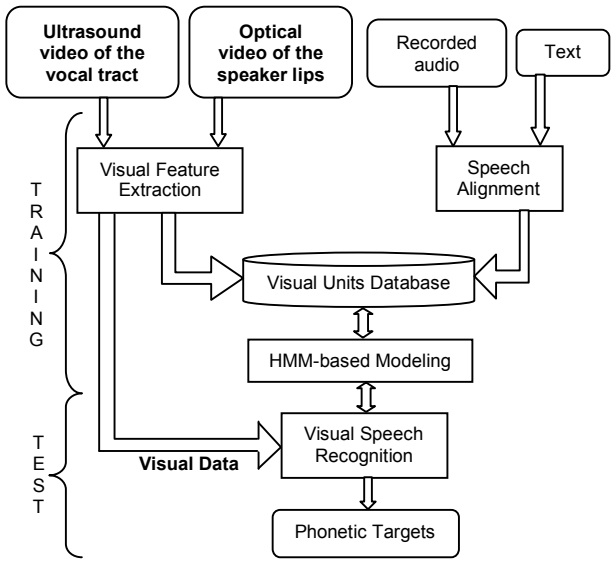


Figure 1: *An overview of the visual speech recognition system. Features derived from images, text and acoustic signals are combined together in an audio-visual speech corpus. HMM-based modeling method is used for speech recognition from video-only data.*

2. Data acquisition and preprocessing

An audio-visual database comprising video sequences of the voice organ together with the uttered speech signal was constructed using the Vocal Tract Visualization Lab HATS system [5]. This system is needed to fix the speaker's head and support the ultrasound transducer under the chin without disturbing speech. A lip profile image is embedded into the ultrasound image, as shown in figure 2.

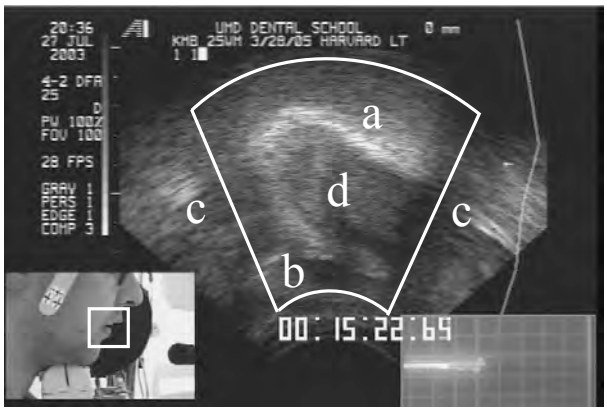


Figure 2: *Example of an ultrasound vocal tract image with embedded lip profile and regions of interest: (a) tongue surface; (b) hyoid bone; (c) hyoid & mandible acoustic shadows; (d) muscle, fat, connective tissue*

The recorded speech dataset consists of the 720 sentences (organized in 72 lists) of the IEEE/Harvard corpus [6] pronounced by a male native American English speaker. After cleaning the database, the resulting speech (43 minutes) was stored as 72473 JPEG frames and 720 WAV audio files sampled at 16000 Hz. Though acquisition of longer datasets would clearly be advantageous, speaker comfort issues render this impossible with our acquisition system in its present form. The synchronous acquisition of two different video streams with the audio signal is also a critical issue. An analog video mixer limits the frame rate of our acquisition chain to 30 Hz, which, compared to a standard speech analysis rate of 100 Hz, is insufficient to register all vocal tract configurations. The recording of larger databases with higher frame rate with an improved system is in the planning stages. The IEEE/Harvard base was initially chosen because all sentences have roughly equal intelligibility and approximately the same duration, grammatical structure and intonation across lists. They are furthermore constructed to preserve the mean frequencies of occurrence of segmental phonemes in the English language. The 1889 words of the IEEE/Harvard base were transcribed into phoneme sequences using the CMU[1] and British English[2] (BEEP) pronunciation dictionaries. Nonetheless, with mean and standard deviation of the number of phone occurrences of 393.5 and 358.5, respectively, the phonetic coverage of the sentences is rather sparse, and some of the phonetic models undoubtedly suffer from a lack of examples in the training database.

3. Segmental speech description

As the visual and audio streams are synchronized, the initial phonetic segmentation of the video sequences can be obtained from the temporal boundaries of the phonemes in the audio signal. The alignment procedure can thus be viewed as a simplified recognition task in which the phonetic sequence is already known. The HTK front-end [7] was used to accomplish this task. The speech acoustic signal is first parameterized using 12 Mel-frequency cepstral coefficients, with their normalized energies, deltas and accelerations (zero mean and unit standard deviation). The transcribed multi-speaker DARPA TIMIT speech database [8] is then used to build initial HMM acoustic models. These 5-state (with one non-emitting initial state, and one non-emitting terminating state), 16 mixture, left-to-right HMM models are finally applied to segment the audio stream of the corpus.

4. Visual feature extraction

The ultrasound images are first reduced to a polar region-of-interest grid delimited by the acoustic shadows (figure 2) of the hyoid bone and mandible. In [4], a PCA-based feature extraction approach, called "EigenTongues" in analogy to Turk and Pentland's "EigenFaces" for face recognition [9], interpreted the ultrasound image as a linear combination of standard vocal tract configurations, thus extracting more information from the images than a contour-based approach. In the present study, this method was improved by adapting the coding of the standard vocal tract configurations to a speech description context. Rather than using a random subset of frames to build the basis vectors of the "EigenTongue" decomposition, visual units from each phone class were picked to constitute the training database. This guarantees a better exploration of the possible vocal tract configurations, and tests showed that equivalent coding quality could be obtained with fewer input features than in the previous method. An analogous approach, called "EigenLips," was used to code the lip frames. The first three basis vectors of the "EigenTongue" and "EigenLip" decompositions are shown in figure 3. Finally, in order to improve the segmentation precision, visual feature sequences were oversampled from 30 Hz to 100 Hz, using linear interpolation.

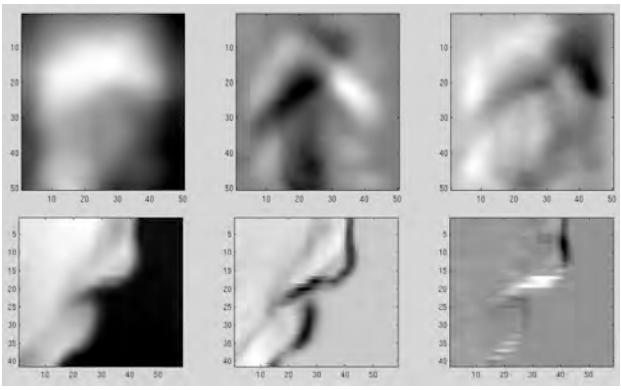


Figure 3: *The first three EigenTongues (top) / EigenLips (bottom), from left to right.*

5. Visual speech recognition

As our speech database is less than one hour long, and has rather sparse phonetic coverage, the use of context-dependant models cannot be envisioned in this study. Rather, a set of 45 left-to-right, 5-state (3 emitting states), continuous monophone HMM's is used to model the visual observation sequences of each phoneme class. Each visual observation is composed of 15 EigenTongues and 5 EigenLips with their

[1] www.speech.cs.cmu.edu/cgi-bin/cmudict
[2] svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

delta and acceleration coefficients, centered and normalized (unit variance). The number of projections onto the set of EigenTongues or EigenLips used for coding is obtained by empirically evaluating the quality of the image reconstruction from its first few components. Once this initial set of models has been created and initialized, embedded training is performed, and the HMM models are incrementally refined by increasing the number of Gaussians per state to 32.

Visual speech recognition is performed using a Viterbi algorithm which finds the optimal path through the word model network, where word models are obtained by concatenating phone HMM models. As the experiment is intended to show the quality of the HMM-based modeling, no statistical language model is used in this study. Thus, speech recognition is constrained only by the use of a pronunciation dictionary built from the IEEE/Harvard sentences, containing in our case 2390 items (some words of the IEEE/Harvard corpus are transcribed with several pronunciations).

In order to increase the statistical relevance of the speech recognizer performance, a jackknife (leave-one-out) technique [10], in which each list of ten sentences was used once as the test set, was employed. For each phone class, a representative measure P of the recognizer performance is defined as

$$P = 100\frac{X}{N} = 100\frac{N-D-S-I}{N} \quad (1)$$

where N is the total number of phones in the test set, S the number of substitution errors, D deletion errors, and I insertion errors. A 95% confidence interval Δ is computed from the Wilson formula [11]:

$$\Delta = 100\frac{\frac{t_\alpha}{N}\sqrt{1+4NX(1-X)}}{1+t_\alpha^2/N} \quad (2)$$

using $t_\alpha = 1.95$ and a normal approximation.

An identical procedure was used for traditional speech recognition based on acoustic features. As in visual speech recognition, a set of 45 context-independent, left-to-right, 5-state, 16-mixture, continuous monophone HMM's is estimated on each training pass. Because the goal of this study is not to achieve high accuracy on audio speech recognition, no more sophisticated modeling methods have been employed. In fact, adding or removing state transitions, tying parameters between models or forming some context-dependant models where possible could have refined these acoustic models. As such, the performance of our acoustic-based recognizer can be considered a target for this database.

6. Results

Figure 4 illustrates qualitatively the performance of the visual speech recognizer on an example in which the predicted phone sequence is time aligned with the reference phonetic transcription via a dynamic programming based string alignment procedure. Correct predictions as well as errors are apparent.

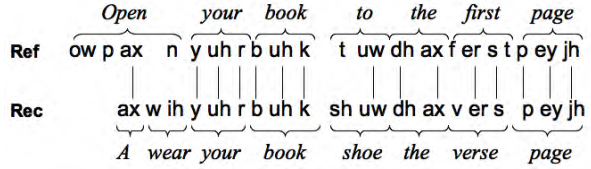


Figure 4: *Reference phonetic transcription (Ref) and predicted phonetic transcription (Rec) derived from visual features.*

The overall performance figures of video-only and audio-only speech recognition experiments are presented in table 1. As the visual (VSR) and audio (ASR) speech recognizers share the same decoding dictionary and do not use language models, the accuracy of the visual HMM models as compared to ASR can be directly deduced from the table. Though as yet inadequate for synthesis purposes, the results are nonetheless quite promising.

Table 1. *Performance comparison of the visual (VSR) and acoustic-based (ASR) speech recognizers.*

Criterion	ASR	VSR
P	71.0 %	54.5 %
Δ	1.3 %	1.4 %
D	874	2994
S	2485	4123
I	2101	1459
N	18874	

The high deletion error rate in visual speech recognition, defined as

$$d = 100 \times \frac{D}{N} \quad (3)$$

may be due to the original video sampling rate of 30 Hz. Indeed, this rate makes the visualization of the vocal tract configuration difficult for very short phones, as illustrated in table 2.

Table 2. *Relation between deletion error rate and mean phone duration. Illustration for phones having the first three highest/lowest deletion error rates.*

Phoneme	d	Mean Duration
dh	37.3 %	0.05 s
t	19.3 %	0.09 s
ax	17.7 %	0.05 s
sh	4.0 %	0.17 s
uw	3.2 %	0.12 s
ey	1.6 %	0.16 s

A decomposition of our results into the different phoneme classes appears in table 3. The recognition scores of plosives (*p, b*), fricatives (*f, v*) and nasal (*m,n*) phonemes show that labial movements are relatively well detected. Velar sounds (*ng, k, g*), formed by the tongue body and articulated near the soft palate, are also well recognized. However, vocal tract configurations corresponding to dental sounds (*th, dh*) and alveolar sounds (*s, sh, t, d*) are more difficult to detect. This can be explained by the lack of information about the relative position of the apex (tip of the tongue) and the teeth. Indeed, in the ultrasound images, the apex is hidden by the acoustic shadow of the mandible. Finally, the performance on vowel detection, which can theoretically be classified by how far forward and how high the tongue is in the mouth, is more difficult to interpret, and for some phonemes (*ah, uh*), the performance of our VSR system is quite low. It seems likely that context-independent HMM models used are not efficient enough to cope with the variability of these phones caused by the co-articulation phenomena.

Table 3. *Visual speech recognizer performance P by phoneme, where Δ is the 95 % confidence interval and N the number of occurrences.*

Phone	Typical word	*P* (in %)	*N*	Δ (in %)
zh	a**z**ure	0	1	NA
hh	**h**ay	9.7	256	7.2
ah	b**u**t	19.6	322	8.5
ch	**ch**oke	27.4	142	14.3
sh	**sh**e	32.9	149	14.7
uh	b**oo**k	34.2	114	16.9
jh	**j**oke	35.3	99	18.1
er	b**ir**d	41.9	203	13.3
ih	b**i**t	43.4	934	6.3
ae	b**a**t	47.4	449	9.1
z	**z**one	52	713	7.2
th	**th**in	52	98	19.1
dh	**th**en	53	915	6.4
y	**y**acht	53.5	114	17.7
d	**d**ay	54	995	6.1
eh	b**e**t	58	379	9.8
ax	**a**bout	58.5	1767	4.6
t	**t**ea	58.9	1733	4.6
b	**b**ee	59.3	440	9.1
uw	b**oo**t	59.8	249	11.9
n	**n**oon	60.6	1453	5
v	**v**an	60.7	349	10
ao	b**ough**t	62	600	7.7
aa	b**o**tt	62	261	11.6
g	**g**ay	62.5	224	12.4
ey	b**ai**t	64	425	9
ow	b**oa**t	66.5	323	10.1
m	**m**om	68.5	524	7.9
ix	deb**i**t	68.7	32	29
f	**f**in	69.6	539	7.7
p	**p**ea	70.6	582	7.3
s	**s**ea	71.8	1131	5.2
ng	si**ng**	74.7	186	12.2
el	bott**le**	75	24	30.6
aw	b**ou**t	75.1	173	12.6
iy	b**ee**t	75.2	733	6.2
ay	b**i**te	76.2	425	8
r	**r**ay	82.5	1157	4.3
k	**k**ey	86.4	805	4.7
w	**w**ay	87.9	537	5.5
l	**l**ay	90	1121	3.5
oy	b**oy**	91.8	49	14.7

7. Conclusions and perspectives

The ability to extract discrete phones from continuous physiological data of the voice organ will be an important step in the design of a silent speech interface. In this article, promising, relevant performance measures have demonstrated the feasibility of phone recognition from ultrasound images of the tongue and optical images of the lips.

At present, the single target phonetic sequence derived from the visual features cannot directly be used to drive the research of acoustic segments in the corpus. The single target will have to be enlarged to a lattice of phonetic targets through which a data-driven unit search of the corpus can correct the stochastic model prediction errors. It would also be desirable to provide improved visual features. The use of optical flow based techniques [12], for example, is currently under study in order to model the movement of the visible articulators. The visual speech recognition will furthermore have to be validated on a larger dictionary with a robust language model, or as a limiting case, without any dictionary. Finally, the construction of a larger database, with a higher video sample rate and an additional front view of the speaker's face, is foreseen.

8. Acknowledgements

The authors would like to acknowledge useful discussions with Leila Zouari, Elsa Angelini, Yacine Oussar, and Pierre Roussel and the reviewers for their suggestions. This work is supported by the French Department of Defense (DGA) and the French National Research Agency (ANR).

9. References

[1] Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., "Session independent non-audible speech recognition using surface electromyography," IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 331-336, 2005.

[2] Nakajima, Y., Heracleous, P., Saruwatari, H., Shikano, K., "A Tissue-conductive Acoustic Sensor Applied in Speech Recognition for Privacy," Smart Objects & Ambient Intelligence Oc-EUSAI 2005, pp. 93-98, 2005.

[3] Denby, B., Oussar, Y., Dreyfus, G., Stone, M., "Prospects for a Silent Speech Interface Using Ultrasound Imaging," IEEE ICASSP, Toulouse, France, pp. I365- I368, 2006.

[4] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," IEEE ICASSP, Honolulu, pp. I1245-I1248, 2007.

[5] Stone, M., and Davis, E., "A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement," Journal of the Acoustical Society of America, vol. 98 (6), pp. 3107-3112, 1995.

[6] IEEE, "IEEE Recommended Practice for Speech Quality Measurements," IEEE Transactions on Audio and Electroacoustics, vol. 17, pp. 225-246, 1969.

[7] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., The HTK Book, Revised for HTK Version 3.3, September 2005, *http://htk.eng.cam.ac.uk/*.

[8] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-100354, 1993.

[9] Turk, M. A., Pentland, A. P., "Face Recognition Using Eigenfaces," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings CVPR, pp. 586-591, 1991.

[10] Efron, B., "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," Biometrika, vol. 68, pp. 589-599,1981.

[11] Hogg, R.V., Tanis, E.A., Probability and statistical inference (5th ed.), Upper Saddle River, NJ: Prentice Hall, 1996.

[12] Angelini, E., Gerard, O., "Review of myocardial motion estimation methods from optical flow tracking on ultrasound data," IEEE EMBS Annual International Conference, New York, NY, USA, pp.1537-1540, 2006.

OUISPER: CORPUS BASED SYNTHESIS DRIVEN BY ARTICULATORY DATA

Thomas Hueber [1,3], Gérard Chollet [3], Bruce Denby [1,2], Maureen Stone [4], Leila Zouari [3]

[1]Laboratoire d'Electronique, Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI-Paristech), 10 rue Vauquelin, 75231 Paris Cedex 05 France
[2]Université Pierre et Marie Curie – Paris VI, 4 place Jussieu, 75252 Paris Cedex 05 France
[3]Laboratoire Traitement et Communication de l'Information, Ecole Nationale Supérieure des Télécommunications (ENST-Paristech), 46 rue Barrault, 75634 Paris Cedex 13 France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore, MD 21201 USA
hueber@ieee.org, gerard.chollet@tsi.enst.fr, denby@ieee.org, mstone@umaryland.edu, leila.zouari@tsi.enst.fr

ABSTRACT

Certain applications require the production of intelligible speech from articulatory data. This paper outlines a research program (Ouisper : Oral Ultrasound synthetIc SPEech souRce) to synthesize speech from ultrasound acquisition of the tongue movement and video sequences of the lips. Video data is used to search in a multistream corpus associating images of the vocal tract and lips with the audio signal. The search is driven by the recognition of phone units using Hidden Markov Models trained on video sequences. Preliminary results support the feasibility of this approach.

Keywords: clinical phonetics, pathophonetics, speech synthesis, automatic speech recognition

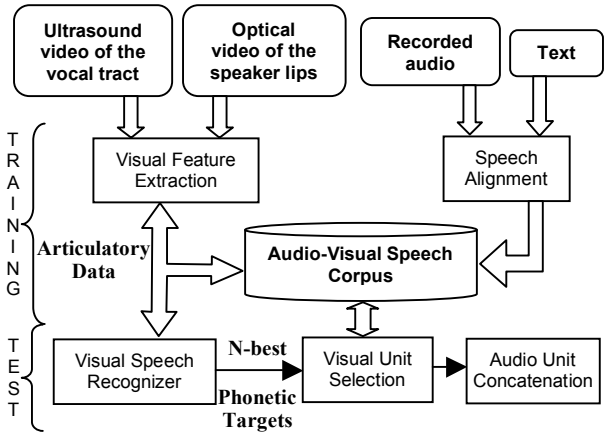
1. INTRODUCTION

The phonemes produced in a language contain a multiplicity of features that are used by the brain to understand and produce speech, but are difficult to reproduce in machine recognition or synthesis. Many articulatory models (CAIP [1], TractSyn [2], Maeda [3]) have focused on rule-based approaches to speech synthesis driven by articulatory parameters. At the same time, the state of the art in text to speech synthesis (for example, the Festival system [4]) uses a corpus-based approach which simply concatenates acoustic speech segments. The Ouisper project proposes to create a speech synthesizer driven by articulatory measurements computed from ultrasound images of the vocal tract and optical images of the speaker lips. It will thus extract discrete phonemes from a continuous data stream and use those as the basis of synthetic speech.

Such a speech synthesizer, driven only by articulatory data, could be used as an alternative to tracheo-oesophageal speech for laryngeal cancer patients, for situations where silence must be maintained, or for voice communication in noisy environments.

Our system is based on the building of an audiovisual corpus which associates articulatory measurements extracted from video to acoustic observations. HMM-based stochastic models, trained on this corpus and combined with a unit selection algorithm, are used to predict and find the optimal sequence of acoustic units, using video-only data. Figure 1 presents an overview of the Ouisper speech synthesis system.

Figure 1: Ouisper corpus-based synthesis system overview



Section 2 of the article details data acquisition and ultrasound image preprocessing, while section 3 describes the visual feature extraction

techniques. Speech segmentation is presented in section 4. Visual speech recognition and acoustic unit selection are introduced in section 5.

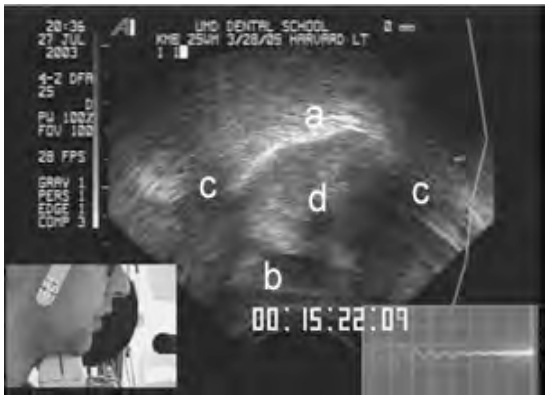
2. DATA ACQUISITION AND PREPROCESSING

The first task of an audiovisual corpus-based speech synthesis system is the construction of an articulatory database comprising video sequences of the voice organ together with the uttered speech signal.

2.1. Data Acquisition

Video sequences of the voice organ are taken using a 30 Hz ultrasound machine and the Vocal Tract Visualization Lab HATS system [5], which maintains acoustic contact between the throat and the ultrasound transducer during speech. A lip profile image is embedded into the ultrasound image, as shown in figure 2.

Figure 2: Example of an ultrasound vocal tract image with embedded lip profile: (a) tongue surface; (b) hyoid bone; (c) hyoid and mandible acoustic shadows; (d) muscle, fat and connective tissue.



The recorded speech dataset consists of 720 sentences organized in 72 lists from the IEEE/Harvard corpus [6], spoken by a male native American English speaker. The IEEE sentences were chosen because they are constructed to have roughly equal intelligibility across lists and all have approximately the same duration, number of syllables, grammatical structure and intonation. After cleaning the database, the resulting speech was stored as 72473 JPEG frames and 720 WAV audio files sampled at 16000 Hz (43 minutes of speech).

The corpus-based synthesis system currently developed in the Ouisper project provides a general methodology to deal with multimodal corpora. Because this approach is multi-stream based, other data streams can be added, such as dynamic electropalatography (EPG) and electromyography (EMG) [7], or a signal recorded from a "non-audible murmur microphone" [8].

2.2. Ultrasound image preprocessing

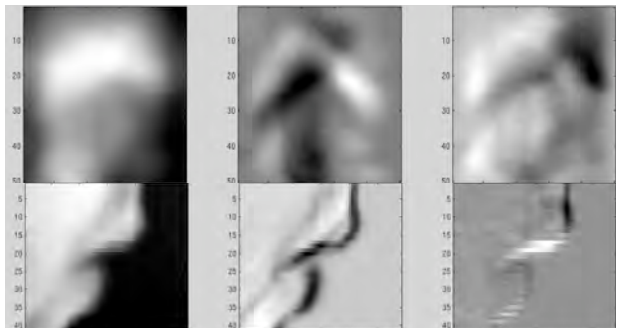
The ultrasound images are first reduced to a polar region of interest grid delimited by the acoustic shadows of the hyoid bone and mandible. The anisotropic diffusion filter of Yu [9] is then applied to remove speckle noise without destroying important image features.

3. VISUAL FEATURE EXTRACTION

3.1. Tongue feature extraction

In many studies (for example [10]), the position of the tongue surface in the image is considered to be the only relevant information in the ultrasound frame, and the extracted articulatory data are simply the parameterized tongue contour. The tongue surface is however poorly imaged when it is nearly parallel to the ultrasound beam, as in the case of the phoneme /i/ for example. Edge tracking algorithms are not enough efficient to cope with small gaps appearing in the tongue contour, and fail for such frames. A solution to this problem is the more global feature extraction approach introduced in [11], wherein Principal Component Analysis (PCA) is used to encode the maximum amount of relevant information in the images, mainly tongue position, of course, but also other structures such as the hyoid bone, muscles, etc. This approach is called "EigenTongues" in analogy to the "EigenFaces" method developed by Turk and Pentland for face recognition [12]. In this way, any vocal tract image is considered to be a linear combination of a set of standard articulatory configurations (*cf.* upper part of figure 3).

Figure 3: The first three EigenTongues (top) / EigenLips (bottom), from left to right.



3.2. Lip feature extraction

To characterize the lip information, a lip contour can of course be used to extract trajectories of the upper/lower lips and commissure from the video sequences. Accurate detection of the lip contour under varying rotations of the speaker face, however, is a difficult task. Hence, a statistical "EigenLip" method was also used to code the lip frames, as illustrated in the lower part of figure 3.

4. SEGMENTAL SPEECH DESCRIPTION

The availability of speech data transcribed at the phonetic level is useful for phonetic research and crucial in the field of corpus based-synthesis. Accurately transcribed speech is needed both for the training of audiovisual speech recognition systems and for the building of a segment database from many transcribed utterances from a single speaker.

4.1. Phonetic segmentation

Manual phonetic segmentation of the speech signal is a difficult and a time consuming task. Several methods have been proposed to speed up this process. The most successful methods have been borrowed from automatic speech recognition, such as Hidden Markov Models (HMM), or Dynamic Time Warping (DTW) techniques, because automatic alignment can be viewed as a simplified recognition task. In this study, an HMM recognizer is used to do forced alignment of speech, that is, a search of the phoneme time boundaries when the phonetic sequence is already known. The speech acoustic signal is parameterized using Mel frequency cepstral decomposition, with normalized energy, delta and acceleration coefficients. HMM acoustic models are initially trained on the transcribed multi-speaker DARPA TIMIT speech database [13].

4.2. Audiovisual database explorer

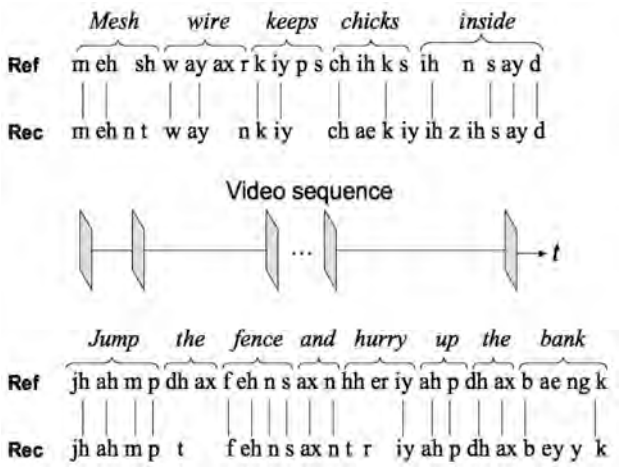
In order to check the speech alignment accuracy and the database coherence, an "audiovisual database explorer" was implemented in the real-time dedicated Max/MSP/Jitter[1] environment. This software allows audiovisual navigation among all of the occurrences of each phoneme classes. For example, a user can listen to all of the /i/ phones, alone or in their context, and simultaneously see the causal motion of the vocal tract and lips.

5. CORPUS BASED SYNTHESIS DRIVEN BY ARTICULATORY DATA

5.1. Speech recognition from video sequences of the vocal tract and lips

During the training phase, visual observation sequences of each phoneme class are modeled using a 5-state, 16 mixture, left-to-right HMM via an embedded re-estimation algorithm. For continuous speech recognition from visual articulatory data, a Viterbi algorithm is used to find the optimal path through the word model network, where word models are obtained by concatenating phone HMM models. In this study, no language model (word sequence probabilities codebook) is used. Thus, this system is driven only by the use of a pronunciation dictionary, which contains in our case 2390 items. Figure 4 illustrates the performance of this visual speech recognizer on two examples, where the predicted phone sequence is time-aligned with the reference phonetic transcription using a dynamic programming-based string alignment procedure.

Figure 4: Reference phonetic transcriptions (Ref) and predicted phonetic transcriptions (Rec) derived from articulatory visual features



Recognition errors are evident in the figure, but our HMM-based system is already able to perform phonetic transcription from *video-only* speech data with over 50% correct recognition. This figure is validated using a jackknife technique, in which each list of ten sentences is used once as the test set. This preliminary result is to be compared to a *best possible* of ≈70% obtained doing traditional

[1] http://www.cycling74.com

speech recognition directly on the *audio* signal, and as such is quite promising.

5.2. Unit selection and concatenation

Given the predicted phonetic sequence, speech synthesis can subsequently be envisioned as selecting phonetic units in the audiovisual corpus. This task is achieved by a Viterbi algorithm, which finds the optimal sequence of visual corpus units that best match the given predicted phonetic sequence. After having selected the visual units, the last step is the concatenation of their correspondents in the acoustical domain. The synthesized speech is of course of good quality for correctly predicted sequences; however, with the current system, the number of errors is still to high to produce a truly usable output signal. One approach will be to enlarge the single target to a lattice of n-best phone sequences, which could be used to drive the search for the optimal solution, a strategy which is an extension of the phonetic vocoder ALISP [14].

6. CONCLUSION AND PERSPECTIVES

The ability to extract discrete phonemes from physiological data is as yet unrealized in speech research. This work is the beginning of such a project and would provide a remarkable enhancement to the current state-of-the-art in speech recognition. The larger goal of the project, to synthesize high quality speech will also be useful to the many applications, commercial and medical, where synthetic speech can augment communication.

Future databases will incorporate a front view of the speaker's head. Optical flow based techniques will also be used to model the motion of the visible articulators. Speech recognition from video-only data will be validated using a larger dictionary, and may be improved by using a robust language model. Finally, the lack of energy, voicing, and rate information in the video sequence will necessitate the creation of a "virtual prosody" in order to obtain good quality speech synthesis. A data-driven approach, in which prosodic patterns are extracted from the corpus, is foreseen.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge useful discussions with Elsa Angelini and the reviewers for their suggestions. This work is supported by the French Department of Defense (DGA) and the French National Research Agency (ANR).

8. REFERENCES

[1] Sinder, D., Richard, G., Duncan, H., Flanagan, J., Krane, M., Levinson, S., Slimon, S., Davis, D. 1997. Flow Visualization in Stylized Vocal Tracts. *Proc. ASVA97* Tokyo.

[2] Birkholz, P., Jackèl, D. 2003. A three-dimensional model of the vocal tract for speech synthesis. *Proc. 15th ICPhS*, Barcelona, 2597-2600.

[3] Maeda, S. 1990. Compensatory articulation during speech : evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W., Marchal, A. (eds), *Speech production and speech modelling*, Dordrecht:Kluwer Academic Publishers, 131-149.

[4] Taylor, P., Black, A. and Caley, R. 1998. The architecture of the Festival Speech Synthesis System, *Proc. 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, 147-151.

[5] Stone, M., Davis, E. 1995. A head and transducer support (HATS) system for use in ultrasound imaging of the tongue during speech. *J. Acoust. Soc. Am.* 98, 3107-3112.

[6] IEEE, 1969. IEEE Recommended Practice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, vol. 17, 225-246.

[7] Jorgensen, C., Lee, D. D., Agabon, S., 2003. Sub Auditory Speech Recognition Based on EMG/EPG Signals. *Proc. Int. Joint Conf. on Neural Networks*, vol. 4, 3128-3133.

[8] Nakajima, Y., Heracleous, P., Saruwatari, H., Shikano, K., 2005. A Tissue-conductive Acoustic Sensor Applied in Speech Recognition for Privacy. *Proc Smart Objects and Ambient Intelligences Oc-EUSAI*, 93-98.

[9] Yu, Y., Acton, S. T., 2002. Speckle Reducing Anisotropic Diffusion. *IEEE Trans. on Image Processing*, vol. 11, 1260-1270.

[10] Denby, B., Oussar, Y., Dreyfus, G., Stone, M., 2006. Prospects for a Silent Speech Interface Using Ultrasound Imaging. *IEEE ICASSP*, Toulouse.

[11] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., 2007. Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface. *IEEE ICASSP*, Honolulu.

[12] Turk, M. A., Pentland, A. P., 1991. Face Recognition Using Eigenfaces. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 586-591.

[13] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-100354

[14] Chollet, G., Cernocky, J., Constantinescu, A., Deligne, S., Bimbot, F., 1998. Toward ALISP: Automatic Language Independent Speech Processing. In: Ponting, K., Moore, R. (eds), *NATO-ASI on Speech Pattern Processing*, Berlin: Springer Verlag, 375-388.

Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface

Thomas Hueber [1,3], *Gérard Chollet* [3], *Bruce Denby* [2,1], *Gérard Dreyfus* [1], *Maureen Stone* [4]

[1]Laboratoire d'Electronique, Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI ParisTech), 10 rue Vauquelin, 75231 Paris Cedex 05 France
[2]Université Pierre et Marie Curie – Paris VI, 4 place Jussieu, 75252 Paris Cedex 05 France
[3]Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, 46 rue Barrault, 75634 Paris Cedex 13 France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore, MD 21201 USA

hueber@ieee.org, gerard.chollet@tsi.enst.fr, denby@ieee.org, gerard.dreyfus@espci.fr, mstone@umaryland.edu

Abstract

Latest results on continuous speech phone recognition from video observations of the tongue and lips are described in the context of an ultrasound-based silent speech interface. The study is based on a new 61-minute audiovisual database containing ultrasound sequences of the tongue as well as both frontal and lateral view of the speaker's lips. Phonetically balanced and exhibiting good diphone coverage, this database is designed both for recognition and corpus-based synthesis purposes. Acoustic waveforms are phonetically labeled, and visual sequences coded using PCA-based robust feature extraction techniques. Visual and acoustic observations of each phonetic class are modeled by continuous HMMs, allowing the performance of the visual phone recognizer to be compared to a traditional acoustic-based phone recognition experiment. The phone recognition confusion matrix is also discussed in detail.

Index Terms: silent speech interface, visual speech recognition

1. Introduction

In recent years, the design of devices allowing silent verbal communication has emerged as a new field in speech processing research. Such "Silent Speech Interfaces" (SSI) could be useful for voice communication in situations where silence must be maintained, or, conversely, in very noisy environments. An SSI might also be an alternative to tracheo-oesophageal or electrolaryngeal speech for laryngeal cancer patients. To build an SSI, voice organ activity could be derived from EMG/EPG signals, as in [1], or, if whispered speech can be tolerated, using a "non-audible murmur microphone" (NAM [2]). In our work, an ultrasound transducer below the chin and a standard video camera (which would be integrated and miniaturized in a final application) are used to directly image the tongue and lips, respectively [3].

In [4], we addressed the problem of continuous-speech phone recognition from ultrasound and optical sequences of the vocal tract as a first step toward corpus-based synthesis. In that work, a visual speech recognizer (VSR) was evaluated on a 43 minute database containing ultrasound tongue images and the lips in profile. The goal of the present article is to evaluate the robustness of our VSR on a larger database with a different speaker. To that end, a new audiovisual database containing 61 minutes of ultrasound and optical sequences of the tongue and lips was recorded together with the uttered speech signal. Corpus text was chosen so that the recorded database would be appropriate for later corpus-based synthesis. The acquisition system was also modified to record both frontal and lateral lip views.

A schematic of the recognition/synthesis system is shown in figure 1. The visual phone recognizer predicts a target phonetic sequence from a continuous stream of visual features used to constrain a unit selection algorithm. This algorithm searches an audiovisual dictionary for the sequence of units which best matches input test data. This article focuses on the visual phone recognizer; the unit selection algorithm, which is an adaptation of the standard path search algorithm used in corpus-based speech synthesis, is described elsewhere [5].

Section 2 describes the acquisition of the new database, details its content and presents the visual feature extraction process. Section 3 details the implementation of the visual speech recognizer and evaluates its performance; a comparison between VSR using frontal and lateral lip images is also presented.

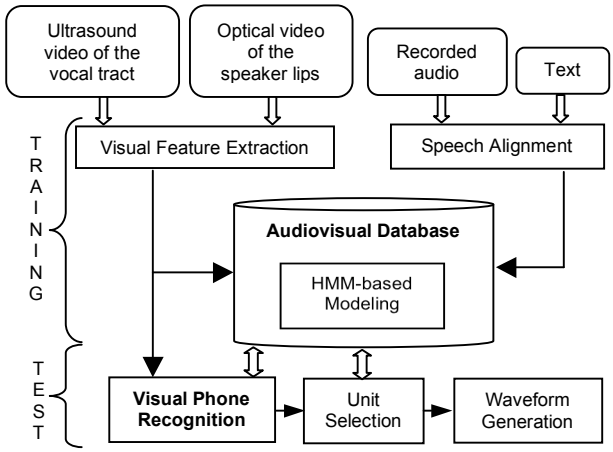


Figure 1: *Framework for corpus-based synthesis driven by visual observation of the tongue and lips.*

2. Building the Audiovisual Speech Database

2.1. Data acquisition protocol and evaluation

Data is recorded using the Vocal Tract Visualization Lab HATS system [6], which maintains the speaker's head immobile and supports the ultrasound transducer under the chin without disturbing speech. The acquisition setup was modified to include two cameras to provide synchronized lateral and frontal view of the speaker's lips together with the ultrasound images of the voice organ and the uttered speech acoustic signal, see figure 2. These three streams are mixed using an analog device, which unfortunately limits the frame rate of the acquisition chain to 30 Hz.

Figure 2: *Example of an ultrasound vocal tract image with embedded lip frontal and lateral view*

Because the recorded multimodal speech signal will serve both for phone-based visual speech recognition and as the basis of a diphone-based concatenative synthesizer, the textual material of the new database must be phonetically balanced and have good diphone coverage. For these two reasons, the CMU-Arctic corpus text [7], which is the basis of the Festvox Text-to-Speech system, was used for our new database. The Arctic database contains 1132 sentences divided into two sets (A and B) containing respectively 593 and 539 items. Both sets are in phonetically balanced American English. Furthermore, with a phoneme set of 41 elements (39 phonemes plus schwa and pause), diphone coverage in sets A and B is 78 % and 75.4 % respectively.

During acquisition, the speaker was instructed to read all sentences of sets A and B as neutrally as possible. Data is recorded in one session during which the speaker remains fixed in the HATS system. Because no re-calibration techniques are employed in our current system, recording data in multiple sessions is not feasible. Since ultrasound imaging of the tongue and its connective tissues (muscle, fat) is very sensitive to modifications of the transducer position, head movement within a session is monitored using palatal traces obtained from 10 cc water deglutitions executed during brief pauses every 90 sentences. During swallowing, the tongue contacts the roof of the mouth, and the ultrasound beam traverses soft tissue until it is reflected by the palate bone [8]. Palatal traces from 4 widely separated deglutitions are shown in figure 3. The proximity of these traces insures that the speaker's head remained stable during the acquisition.

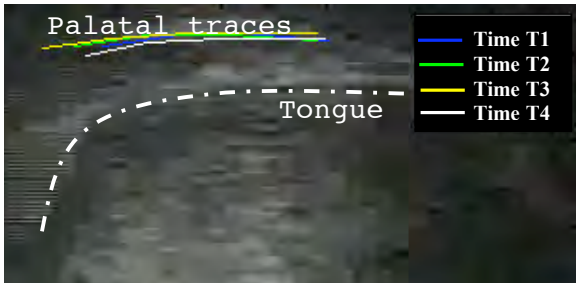


Figure 3: *Superposition of palatal traces extracted from 4 deglutitions recorded periodically during data acquisition.*

The full Arctic A set was acquired, but speaker fatigue, after more than 2 hours in HATS, allowed only 80 % of the B set to be recorded; the total number of sentences was thus 1020 rather than the expected 1132. After cleanup of the database, the resulting 61 minutes of speech was stored as 109553 Bitmap frames and 1020 WAV audio files sampled at 16000 Hz. The new database is thus 30 % larger than that used in our previous study.

2.2. Phonetic alignment of the speech waveform

The acoustic signal of each recorded sentence was first parameterized using 12 Mel-frequency cepstral coefficients, along with their energies and first and second derivatives. The phonetic forced-alignment procedure is a simplified recognition task in which the phonetic sequence is already known. This recognition task is achieved using an initial set of 40 HMM acoustic models trained on the transcribed multi-speaker DARPA TIMIT speech database [9]. These 5-state (with one non-emitting initial state, and one non-emitting terminating state), 16 mixture, left-to-right HMM models are refined on and then used to segment the audio stream of the corpus. All HMM work in our study was done using the HTK front-end [10]. With 33637 phonemes labeled, the actual diphone coverage obtained for our audiovisual speech database was 79.4 % (1271 different diphones of a possible 1599).

2.3. Visual feature extraction

Regions of interest for the tongue and the lips are first resized to 64x64 pixel images using cubic interpolation. In order to decrease the effects of speckle, each ultrasound frame is filtered using an anisotropic diffusion filter [11]. Then, the PCA-based "EigenTongues" decomposition described in [12] is used to encode each frame. An adaptation of the "EigenFaces" method [13], this technique projects each ultrasound image of the vocal tract into the representative space of "EigenTongues", which can be interpreted as the space of the "standard vocal tract configurations". A similar approach is used to code frontal and lateral images of the lips. Figure 4 illustrates how each ultrasound and optical image is coded by its coordinates β_T, β_F, β_L, in the "EigenTongues/EigenLips" space. The indices *n,m,p* which quantify the number of projections onto the set of EigenTongues/EigenLips used for coding are obtained empirically by evaluating the quality of the image reconstructed from its first few components. Typical values of the triplet *(n,m,p)* used for this database are *(20,15,15)*. Finally, visual feature sequences are oversampled from 30 Hz to 100 Hz using linear interpolation. The EigenTongues/EigenLips coefficients, with their first and

second derivative, are concatenated into the same "visual feature vector", in a *feature fusion* strategy.

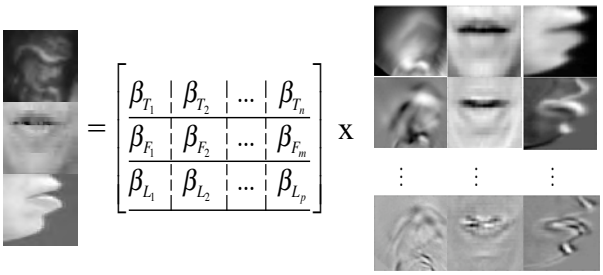


Figure 4: *Encoding ultrasound and optical images of the tongue and lips using the EigenTongue/EigenLips decomposition.*

3. Visual Phone Recognition

3.1. Protocol

The observed visual sequences of each phonetic class are modeled by a left-to-right, 5-state (3 emitting states), continuous HMM (monophone only, due to our dataset size). Models parameters are estimated and refined using incremental embedded training during which the number of Gaussians per state is increased up to 32. As our experiment is intended to show the quality of the HMM-based modeling, neither a statistical language model nor phonotactic constraints are used in this study.

The 1020 sentences of the database are divided into 34 lists of 30 sentences. During performance estimation, each list is used once as the test set while the other 33 lists compose the training set, using a jackknife strategy [14]. The recognizer performance P_{VSR} is defined as

$$P_{VSR} = 100 \cdot \frac{N - D - S - I}{N} \quad (1)$$

where N is the total number of phones in the test set, S the number of substitution errors, D deletion errors, and I insertion errors. Although frontal and lateral views of speaker's lips are available in the newly recorded database, the two streams are not used together in the visual phone recognition process, to simulate the conditions of a simple, wearable prototype. A comparative study of visual phone recognition using ultrasound and frontal or lateral lips is however made.

A traditional, acoustic-based phone recognizer is also evaluated on the same database using the HMM acoustic models estimated for the phonetic alignment of the audio-visual database in section 2.2. The performance of this acoustic-based phone recognizer is considered as a 'target' for VSR on this database.

3.2. Results and Interpretation

Table 1 presents the global performance of the visual-based and acoustic-based phone recognizers, with performance of the visual phone recognizer broken down into frontal and profile visual lip input features.

Results are significantly improved compared to [4], (54 % *vs* 60 %), but because the two databases have neither the same text material nor the same speaker, a strict comparison is not in order. The similar performances between [4] and this study may rather be interpreted as evidence of the method's robustness. In fact, as results using the lateral lip view are almost identical in the two studies (54 % [4] *vs* 56 %), much of the improvement (60 %) appears to be due to the use of the frontal lip view. This result agrees with the human lipreading experiments described in [15] as well as the digit recognition task detailed in [16].

	ASR	VSR = Tongue + Lips	
		Lateral	Frontal
P	83.9 %	56 %	**60 %**
Δ	0.7 %	0.9 %	0.9 %
D	1226	4776	4424
S	2389	7830	7036
I	1985	2695	2430
N	33637		

Table 1. *Visual and acoustic based phone recognizer performance with a 95% confidence interval Δ*

The performance of the visual phone recognizer can be analyzed using a confusion matrix as displayed at figure 5. As could be expected, it is phones with similar articulatory gestures (tongue and lips), {[p],[b],[m]}, {[k],[g],[ng]}, {[f],[v]}, {[s],[z]} and {[t],[d],[n]}, which are the most often confused by the system. Some of the vowel mismatches are quite "reasonable", such as [uh] (book) confused with [uw] (boot), and [iy] (beet) interpreted as [ih] (bit). The confusion of several phones with schwa [ah] can be explained by the well-known reduction phenomenon for vowels, or in other cases by the presence of a syllabic consonant, such as the [l] in "bottle". Diphthongs for which a tongue glide is involved are sometimes confused with one of their pure vowel components, for example [ey] (bait), [oy] (boy) and [ow] being matched with [ah], [iy] and [ao] (caught) respectively. The matrix also clearly shows an error occurring mainly on dental and alveolar sounds {[th],[dh]} (thin, then) and {[t],[d],[s],[sh]}. This is explained by the lack of information about the tongue tip (apex) in the ultrasound images, which is sometimes hidden by the acoustic shadow of the mandible. The relatively high number of insertions has a negative impact on the global performance, and the use of a statistical language model would certainly be helpful here. Finally, the predicted phonetic sequence is plagued by a large number of deletion errors. The phones which are most often deleted are very short ones such as the schwa [ah], as well phones corresponding to rapid articulatory gestures such as {[t]-[d]-[n]}. In fact, with a mean duration of 60 ms, the phone [t] is most often represented by fewer than two ultrasound frames with our current 30 Hz acquisition setup. A faster acquisition system is in the planning stages.

As the partitioning of phonetic space used is very fine (40 phonetic classes), our 60 % result is in fact pessimistic; it would no doubt be higher if some of the "reasonable" confusions mentioned, as well as mismatches due to incorrect phonetic labeling, were not considered "errors" in the performance computation. Too, such mismatches in the recognition stage need not necessarily lead to unintelligible synthesis. Some psychoacoustic effects and results provided by speech perception theory could potentially also be used to advantage. Thus, though as yet not perfect, our results are already promising enough to warrant investigating the feasibility of a phonetic vocoder driven by ultrasound and optical images of the tongue and lips.

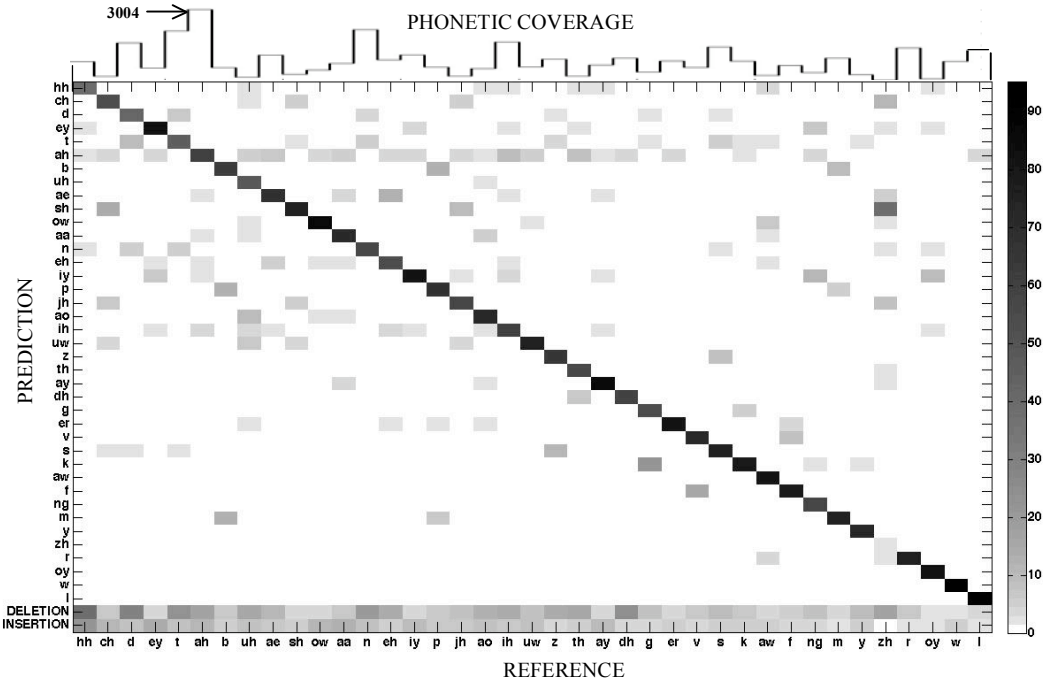


Figure 5: *Confusion matrix for phone recognition from ultrasound tongue sequences and frontal lip views. The color space map was chosen to emphasize the errors. Phone labels are in the TIMIT format. The histogram shows the number of occurrences of each phone in the database.*

4. Conclusions and perspectives

The visual phone recognizer is able to predict a 60 % correct phonetic target sequence from a continuous stream of video-only data. Applied to two different databases, with different textual materials and speakers (one male, one female), the proposed method appears robust, and could be a good starting point for phonetic vocoder driven only by visual observation of the voice organ. It is clear, however that the problem of phone insertion and deletion must be addressed more aggressively. The use of a language model and the acquisition of data at a higher rate are to be investigated in future work. We also intend to take into account possible asynchronies between articulators and compare the *feature fusion* strategy to a multistream HMM-based approach.

5. Acknowledgements

This work is supported by the French Department of Defense (DGA) and the French National Research Agency (ANR), under contract number ANR-06-BLAN-0166.

6. References

[1] Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., "Session independent non-audible speech recognition using surface electromyography," IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 331-336, 2005.

[2] Nakajima, Y., Heracleous, P., Saruwatari, H., Shikano, K., "A Tissue-conductive Acoustic Sensor Applied in Speech Recognition for Privacy," Smart Objects & Ambient Intelligence Oc-EUSAI 2005, pp. 93-98, 2005.

[3] Denby, B., Oussar, Y., Dreyfus, G., Stone, M., "Prospects for a Silent Speech Interface Using Ultrasound Imaging," IEEE ICASSP, Toulouse, France, pp. I365- I368, 2006.

[4] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips," Interspeech, pp. 658-661, Antwerp, Belgium, 2007.

[5] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Toward a Segmental Vocoder driven by Ultrasound and Optical Images of the Tongue and Lips," *submitted to these proceedings*.

[6] Stone, M., and Davis, E., "A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement," Journal of the Acoustical Society of America, vol. 98 (6), pp. 3107-3112, 1995.

[7] Black, A. W., Lenzo, K., Building voices in the Festival speech synthesis system, 2000, *http://festvox.org/bsv*.

[8] Epstein, M., Stone, M., Pouplier, M., Parthasarathy, V., "Obtaining a palatal trace for ultrasound images," Proc. of Meeting of Acoustical Society of America, 2004.

[9] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-100354, 1993.

[10] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., The HTK Book, September 2005, http://htk.eng.cam.ac.uk/.

[11] Y. Yu and S. T. Acton, "Speckle Reducing Anisotropic Diffusion," IEEE Trans. on Image Proc., vol. 11, pp. 1260-1270, 2002.

[12] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," IEEE ICASSP, Honolulu, pp. I1245-I1248, 2007.

[13] Turk, M. A., Pentland, A. P., "Face Recognition Using Eigenfaces," IEEE Computer Soc. Conf. on Comp. Vision and Pat. Reco., Proc. CVPR, pp. 586-591, 1991.

[14] Efron, B., "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," Biometrika, vol. 68, pp. 589-599, 1981.

[15] T. R. Jordan and S. M. Thomas, "Effects of horizontal viewing angle on visual and audiovisual speech recognition," in Journal of Experimental Psychology: Human Perception and Performance, vol. 27, no. 6, 2001, pp. 1386–1403.

[16] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," in Proceedings of the 8th IEEE Workshop on Multimedia Signal Processing (MMSP '06), pp. 24-28, Victoria, BC, Canada, October 2006.

Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips

Thomas Hueber [1,3], *Gérard Chollet* [3], *Bruce Denby* [2,1], *Gérard Dreyfus* [1], *Maureen Stone* [4]

[1]Laboratoire d'Electronique, ESPCI Paristech, 10 rue Vauquelin, 75231 Paris Cedex 05 France
[2]Université Pierre et Marie Curie - Paris VI, 4 place Jussieu, 75252 Paris Cedex 05 France
[3]Laboratoire Traitement et Communication de l'Information, Telecom Paristech, 46 rue Barrault, 75634 Paris Cedex 13 France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore, MD 21201 USA

hueber@ieee.org, gerard.chollet@tsi.enst.fr, denby@ieee.org, gerard.dreyfus@espci.fr, mstone@umaryland.edu

Abstract

This article presents a framework for a phonetic vocoder driven by ultrasound and optical images of the tongue and lips for a "silent speech interface" application. The system is built around an HMM-based visual phone recognition step which provides target phonetic sequences from a continuous visual observation stream. The phonetic target constrains the search for the optimal sequence of diphones that maximizes similarity to the input test data in visual space subject to a unit concatenation cost in the acoustic domain. The final speech waveform is generated using "Harmonic plus Noise Model" synthesis techniques. Experimental results are based on a one-hour continuous speech audiovisual database comprising ultrasound images of the tongue and both frontal and lateral view of the speaker's lips.

Index Terms: silent speech, corpus-based speech synthesis, visual speech recognition

1. Introduction

The objective of a "Silent Speech Interface" is to permit voice communication without the vocalisation of sound. Such a system primarily targets applications in which silence must be maintained, but could also be used to enable voice communication in situations where standard speech is masked by background noise. Since no glottal activity is required, it could furthermore have application as an alternative to tracheo-oesophageal and electrolaryngeal speech for laryngectomized patients. In the literature, silent communication has usually been envisioned as a speech recognition task driven by observation of the voice organ. The input articulator activity may be derived from EMG/EPG signals, as in [1], or, as in our case, from ultrasound and optical images of the vocal tract.

In [2] and [3], we addressed the problem of continuous-speech phone recognition from ultrasound and optical video sequences of the vocal tract. Here, we propose to use this visual phone recognition step (VSR) as the basis of a phonetic vocoder driven by video-only data. Our approach does not use a specific vocal tract model as in articulatory synthesis, but rather is based on the building of an audiovisual dictionary in which each visual unit has an equivalent in the acoustic domain. Given a test sequence of visual features and the phonetic target predicted by the VSR, a unit selection algorithm searches in this audiovisual dictionary the optimal sequence of units that best matches the input test data. The proposed unit selection algorithm is an adaptation of the standard path search algorithm used in corpus-based speech synthesis. The quality of the match is defined optimally as a compromise between a target cost evaluated in the visual space and a concatenation cost evaluated in the acoustic domain. The output speech waveform is generated by concatenating a "Harmonic plus Noise Model" (HNM) representation of acoustic segments for all selected units. An overview of the recognition/synthesis system is given in figure 1.

The system is evaluated on a 61 minute audiovisual database of ultrasound and optical sequences of the tongue and lips, recorded in synchrony with the uttered speech signal. Text material was chosen with corpus-based synthesis specifically in mind.

Section 2 of the article summarizes database content and acquisition, feature extraction procedures, and the visual phone recognition step (further details on these system blocks are given in [3]). The unit selection algorithm and speech waveform generation techniques used in the corpus-based synthesis, which are the main focus of this article, are detailed in section 3, along with preliminary experimental results of synthesis driven by video-only data.

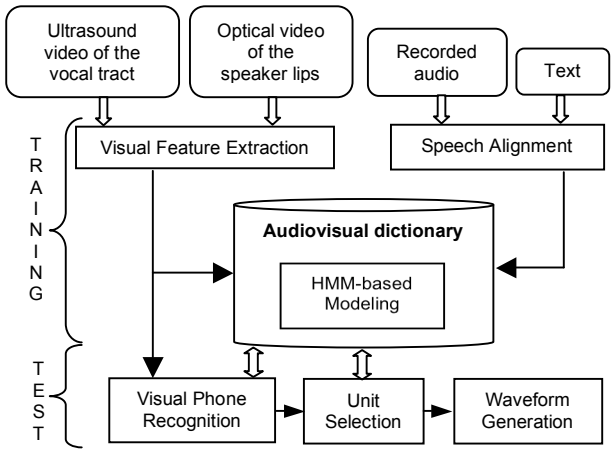


Figure 1: *Framework for a phonetic vocoder based on visual observation of the tongue and lips.*

2. Visual Phone Recognition Step

2.1. Database acquisition and phonetic content

The acquisition system fixes the speaker's head and supports the ultrasound transducer under the chin without disturbing articulator movement [4]. The protocol described in [2] was modified to include both lateral and frontal view of the speaker's lips along with the ultrasound tongue images and acoustic speech signal. The streams are mixed at a video frame rate of 30 Hz. A typical frame is shown in figure 2.

Figure 2: *Example ultrasound vocal tract image showing frontal and lateral lip views*

Because the recorded multimodal speech signal is used both for phone-based VSR and as the basis of a diphone-based concatenative synthesizer, the text of the database must be phonetically balanced and have good diphone coverage. The CMU-Arctic corpus text [5] was chosen for our acquisitions. This base consists of 1132 sentences divided into two phonetically balanced sets, A and B, of 593 and 539 items respectively. With a phoneme set of 41 elements (39 phonemes plus schwa and pause), the diphone coverage of sets A and B in the corpus is 78 % and 75.4 % respectively.

A native speaker of American English read sentences from sets A and B in a single session lasting over 2 hours. Speaker fatigue limited acquisitions to the first 1020 of the 1132 Arctic phrases (100 % of set A and 80 % of set B). Multiple sessions are not done at present in order to avoid compensating for imprecisions in the re-alignment of the transducer relative to the vocal tract. After cleanup, the resulting 61 minutes of speech was stored as 109553 bitmap frames and 1020 WAV audio files sampled at 16000 Hz.

2.2. Phonetic alignment of the speech waveform

The acoustic wave of each recorded sentence is parameterized by 12 Mel-frequency cepstral coefficients (MFCC) with their energies and first and second derivatives. The phonetic labeling is performed by an HMM-based forced alignment procedure with an initial set of 40 HMM acoustic models, trained on the transcribed multi-speaker DARPA TIMIT speech database [6]. These 5-state (with one non-emitting initial state, and one non-emitting terminating state), 16 mixture, left-to-right HMM models are used to segment the corpus audio stream. The HMM training and recognition procedure are done using the HTK front-end [7].

To facilitate the subsequent diphone speech synthesis, a segmentation of the database into diphones is deduced from the phone labeling by searching for spectral stability points at the boundaries of all phones. With 33637 phonemes labeled, the diphone coverage of our audiovisual speech database is found to be 79.4 % (1271 diphones out of a total of 1599).

2.3. Tongue and lip video feature extraction

Tongue and the lip regions of interest are resized to 64x64 pixels via cubic interpolation and ultrasound images are filtered using an anisotropic diffusion filter [8]. Then the PCA-based EigenTongues/EigenLips decomposition [9] is used to encode the tongue and lips by considering their positions as a linear combination of standard configurations. The number of useful EigenTongues and EigenLips coefficients to keep is fixed empirically at 20 and 15 respectively. Features are finally resampled at 100 Hz using linear interpolation before being concatenated with first and second derivatives into a single vector.

2.4. Recognition protocol and performance

Observed sequences of each phonetic class are modeled by a left-to-right, 5-state (3 emitting states, 32 gaussians per state), continuous monophone HMMs. No statistical language model is used at this stage so as to allow evaluation of the quality of the HMM-based modeling alone. The database is divided into 34 lists of 30 sentences. In the performance estimation, a jackknife strategy [10] uses each list once for validation while the other 33 make up the training set.

The recognizer performance P_{VSR} is defined as

$$P_{VSR} = 100 \cdot \frac{N - D - S - I}{N} \quad (1)$$

where N is the total number of phones in the test set, S the number of substitution errors, D deletion errors, and I insertion errors. To establish a performance target for the visual recognizer, a standard acoustic-based phone recognizer is evaluated on the same database. This uses 40 context-independent, left-to-right, 5-state, 16-mixture, continuous monophone HMMs estimated on each training pass.

Table 1 compares performances of the visual-based and acoustic-based phone recognizers. VSR performance is already almost 80 % of that obtained using ASR, indicating that initial synthesis experiments are indeed justified. A full discussion of the visual phone recognition step is given in [3].

	ASR	VSR = Tongue + Lips	
		Lateral	Frontal
$P \pm \Delta$	83.9 ± 0.7 %	56 ± 0.9 %	**60 ± 0.9 %**

Table 1. *Visual and acoustic based phone recognizer performance P for a 95% confidence interval Δ*

3. Corpus-Based Synthesis

3.1. Unit selection

The visual speech recognizer is able to identify a discrete sequence of phones in a continuous stream of visual features. The recorded database, automatically labeled at the phonetic level (section 2.2), can in turn be considered as an audiovisual dictionary of speech units in which each visual item has an equivalent in the acoustic domain. In our proposed phonetic vocoder, the visual phone recognizer drives corpus-based synthesis assisted by a unit selection procedure. Starting from the predicted phonetic target, the algorithm searches the optimal sequence of diphones that maximize similarity to input test data in visual space while limiting unit

concatenation cost in the acoustic domain. This algorithm is based on the standard path search algorithm used in concatenative speech synthesis described in [11]. The overall scheme is illustrated in figure 3.

Assuming a test sequence of visual features $v = v_1 \ldots v_N$ where N is the length of sequence, and $\tau = \tau_1 \ldots \tau_T$ the temporal segmentation of v given by the visual phone recognizer, the sequence t_τ of T target units is defined by:

$$t_\tau = [t_{\tau_1}],...,[t_{\tau_T}] = [v_{\tau_1},...,v_{\tau_2}],...,[v_{\tau_{T-1}},...,v_{\tau_T}] \quad (2)$$

The unit selection algorithm finds, among all appropriate units, the optimal sequence $\{u_k\}$ that best matches the target t_τ. The quality of the match is determined by two costs, C^t and C^c.

The target cost C^t expresses the visual similarity between target units and the units selected in the dictionary and is given by:

$$C^t(u_k, t_{\tau_i}) = DTW(u_k, t_{\tau_i}) \quad (3)$$

where *DTW(a,b)* is the cumulative distance obtained after a dynamic time warping between the two sequences of visual feature vectors. This non-linear alignment procedure takes naturally into account temporal stretching and compression of the motion of the articulators.

The concatenation cost C^c estimates the spectral discontinuity introduced by the concatenation of two units u_{LEFT} and u_{RIGHT} and is given by:

$$C^c(u_{LEFT}, u_{RIGHT}) = D\left(MFCC(u_{LEFT_{END}}), MFCC(u_{RIGHT_1})\right) \quad (4)$$

where D is the Euclidean distance and $MFCC(u_l)$ are MFCC coefficients of the unit u at frame l.

Because the audiovisual dictionary can be considered as a fully connected state transition network, the search for the least costly path that best matches the test sequence can be determined by a Viterbi algorithm [12]. In this network, each state is occupied by a unit. State occupancy is estimated using the visual-based target cost function and transition between states is evaluated by the acoustic-based concatenation cost.

3.2. HNM-based speech waveform generation

After the selection stage, speech can be synthesized by concatenating acoustic components of selected diphones. However, because no prosodic information such as pitch, energy and duration, is used during the unit selection stage, pitch and time-scale adaptations are necessary. Acoustic modifications are achieved using a "Harmonic Plus Noise" representation of the speech signal [13]. In the HNM framework, the spectrum of a speech frame *s(t)* is described as the sum of a harmonic part *H(t)* and a noise part *B(t)*:

$$s(t) = H(t) + B(t) = \left[\sum_{k=1}^{N} A_k \cos\left(2\pi k f_0 t\right) + \varphi_k\right] + \left[N_{gauss} * F(t)\right] \quad (5)$$

where N is the number of harmonics included in $H(t)$, f_0 is the estimated fundamental frequency, N_{gauss} a gaussian noise frame and $F(t)$ an autoregressive filter. Our implementation employs 12 harmonic components along with a 16th-order auto-regressive model for the noise part.

HNM is a pitch-synchronous scheme that is flexible enough to implement good-quality prosodic modifications. In our case, acoustic modifications consist of phone duration adaptation, and pitch and spectral smoothing. Phone durations are adapted according to the temporal segmentation provided by the HMM-based phone recognition step described in section 2. Because no information about the global evolution of pitch is directly available in a silent speech application (absence of glottal activity), a strong smoothing of the fundamental frequency over the sentence is applied. Such a basic treatment helps limit non-realistic prosodic variations but (empirically) can degrade voice naturalness. As a final step, the HNM parameters are smoothed near diphone boundaries using linear interpolation.

The example chosen for figure 3 illustrates the interplay between the two cost functions. The diphone [w-ih] is selected correctly for its similarity to the test sequence. However, the next diphone [ih-ah] does not match well with the input sequence (as at the end of phone [ah]); the selection of this unit is mainly due to its acoustic continuity with the previous unit. We note that in the present algorithm, the target and concatenation cost are weighted manually.

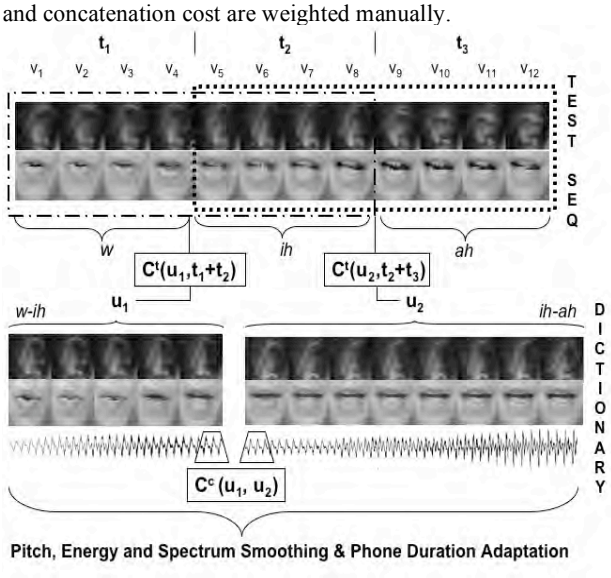


Figure 3: *Corpus-based synthesis procedure (T=3)*

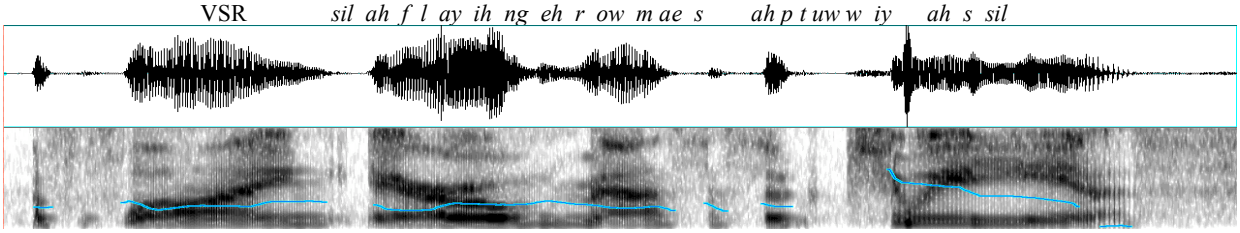
3.3. Experimental results

The quality of the synthesized waveform obviously depends strongly on the performance of the visual phone recognizer. In our current framework, the unit selection synthesis is driven exclusively by the predicted phonetic sequence, and thus an error during the recognition stage will necessarily corrupt the synthesis. With visual phone recognizer accuracy of only 60%, consistently intelligible synthesis is not yet possible.

A first empirical evaluation of our "silent vocoder" is presented in figure 4. Example 1 represents a 'typical' performance of the system, with 69 % of the phones correctly identified, while the phrase of example 2 has 95 % of phones correctly matched. Two distinct types of errors are apparent (see also [3]): first, phones with similar articulatory gestures, such as {[p],[b],[m]} are sometimes confused; secondly, very short phones such as {[t],[b],[n]} can be missed due to the 30 Hz acquisition rate. The multimedia file provided for the second example illustrates the ability of the synthesis technique to produce an intelligible speech signal with "acceptable" prosody when the predicted phonetic target is correct. There are still difficulties identifying short pauses or within-sentence silences (anticipation phenomena), and better results are obtained on short sentences with no more than one or two prosodic groups. Clearly, a more detailed study of the impact of different types of error on synthesis quality will be necessary. Thus, although our system is still not fully functional, this approach for a segmental speech coder driven only by visual observation seems promising.

Example 1 - *A flying arrow passed between us* - P_{VSR} = 69 %

Reference *sil ah f l ay ih ng ae r ow p ae s t b ah t w iy n ah s sil*
VSR *sil ah f l ay ih ng eh r ow m ae s ah p t uw w iy ah s sil*



Example 2 - *They laughed like two happy children* - P_{VSR} = 95 %

Reference *sil dh ey l ae f t l ay k t uw hh ae p iy ch ih l d r ah n sil*
VSR *sil dh ey l ae f t l ay k t uw hh ae p iy ch ih l r ah n sil*

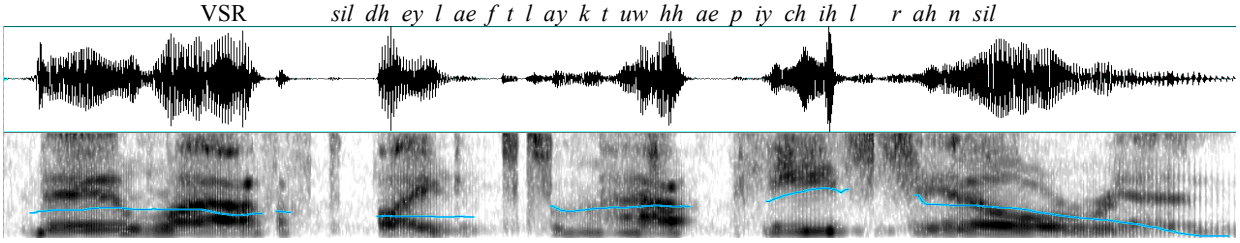


Figure 4: *Phone recognition and associated corpus-based synthesis from video-only data (fundamental frequency in light blue)*

Multimedia files submitted

"exampleX_synth.wav" with X=1,2 : Synthesis from video-only speech data
"exampleX_orig.wav" with X=1,2 : Original sentence (target)

4. Conclusions and Perspectives

The proposed segmental speech coder driven by video-only data combines an HMM-based visual phone recognition stage with an audiovisual unit selection algorithm and robust HNM-based synthesis techniques. To date, synthesis quality depends only on the performance of the visual phone recognizer, currently at 60 %. To improve the recognition stage, several solutions are envisioned. The use of a statistical language model or phonotactic linguistic constraints in the HMM decoding stage will be investigated. A new, higher rate acquisition system is also under development in order to reduce the number of phone deletion errors. As our modeling technique does not presently take into account possible asynchronies between articulators, the use of multistream HMMs [14] could prove useful. The unit selection algorithm furthermore is currently driven only by the output of the phone-based recognizer; it might be more fruitful to consider a combination of HMM-based stochastic modeling and data-driven techniques. A deeper dictionary search, also including longer units, such as polyphones, could capture more contextual effects and improve general performance. Finally, the system should be evaluated on more realistic test databases containing either whispered or totally silent speech. Such data will be very useful to learn about the particularities of tongue and lip movement in silent speech.

5. Acknowledgements

This work is supported by the French Department of Defense (DGA) and the French National Research Agency, under contract number ANR-06-BLAN-0166.

6. References

[1] Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., "Session independent non-audible speech recognition using surface electromyography," IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 331-336, 2005.

[2] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips," Interspeech, pp. 658-661, Antwerp, Belgium, 2007.

[3] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Visual Phone Recognition for an Ultrasound-Based Silent Speech Interface," *submitted to these proceedings*.

[4] Stone, M., and Davis, E., "A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement," Journal of the Acoustical Society of America, vol. 98 (6), pp. 3107-3112, 1995.

[5] Black, A. W., Lenzo, K., "Building voices in the Festival speech synthesis system," *http://festvox.org/bsv*, 2000.

[6] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NTIS order number PB91-100354, 1993.

[7] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., The HTK Book, September 2005, http://htk.eng.cam.ac.uk/.

[8] Y. Yu and S. T. Acton, "Speckle Reducing Anisotropic Diffusion," IEEE Trans. on Image Proc., vol. 11, pp. 1260-1270, 2002.

[9] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," ICASSP, Honolulu, pp. I1245-I1248, 2007.

[10] Efron, B., "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods," Biometrika, vol. 68, pp. 589-599, 1981.

[11] Hunt, A. J., Black, A. W., "Unit selection in a concatenative speech synthesis system using a large speech database," IEEE ICASSP, pp. 373–376, Atlanta, 1996.

[12] Forney, G. D., The Viterbi algorithm. Proceedings of the IEEE 61(3), pp. 268-278, 1973.

[13] Stylianou, Y., Dutoit, T., Schroeter, J., "Diphone Concatenation using a Harmonic plus Noise Model of Speech," Eurospeech, pp. 613-616, Rhodes, Greece, 1997.

[14] Gravier, G., Potamianos, G., and Neti, C., "Asynchrony modeling for audio-visual speech recognition," In Proceedings of the Second international Conference on Human Language Technology Research, San Diego, California, 2002.

Acquisition of Ultrasound, Video and Acoustic Speech Data for a Silent-Speech Interface Application

T. Hueber[1,3], G. Chollet[3], B. Denby[2,1], M. Stone[4]

[1]Laboratoire d'Electronique, ESPCI ParisTech, Paris, France
[2]Université Pierre et Marie Curie – Paris VI, Paris, France
[3]Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, Paris, France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, Baltimore, USA

hueber@ieee.org, gerard.chollet@tsi.enst.fr, denby@ieee.org, mstone@umaryland.edu

Abstract

This article addresses synchronous acquisition of high-speed multimodal speech data, composed of ultrasound and optical images of the vocal tract together with the acoustic speech signal, for a silent speech interface. Built around a laptop-based portable ultrasound machine (Terason T3000) and an industrial camera, an acquisition setup is described together with its acquisition software called Ultraspeech. The system is currently able to record ultrasound images at 70 fps and optical images at 60 fps, synchronously with the acoustic signal. An interactive inter-session re-calibration mechanism which allows recording of large audiovisual speech databases in multiple acquisition sessions is also described.

1 Introduction

Measuring the activity of the vocal tract during speech is critical in a variety of fields such as phonology, linguistics, speech pathology, anatomy and multimodal speech processing. In recent years and despite the success of MRI, the use of ultrasound for vocal tract imaging and analysis remains popular mainly because of its non-invasive property, its good time resolution, its clinical safety, and its ability to image the tongue in non-supine subjects.

In the "silent speech interface" developed in the *Ouisper* project, non-acoustic features, derived from ultrasound images of the tongue and optical images of the lips, are used to drive a speech synthesizer. A laryngectomyzed patient could use this kind of system as an alternative to oesophageal speech, tracheo-oesophageal speech or the electrolarynx. A silent-speech interface could also be useful in situations where silence must be maintained, or for voice communication in noisy environments. Described in [1] and [2], the *Ouisper* segmental speech coder is built around a large audiovisual corpus (more than one hour) which associates articulatory features extracted from a 30 Hz source video with acoustic observations. In the proposed approach, a sequence of phones is "recognized" from visible motion of the tongue and lips. In the visuo-phonetic decoding stage, phonetic prediction is hampered by a large number of deletion errors. Most often, deleted phones are very short phones for which correct visualization is difficult with a 30 Hz acquisition system. Thus, a faster acquisition system is needed.

Several systems able to acquire a sequence of ultrasound images of the tongue together with the uttered speech signal have been described in the literature. However, the coupling of an ultrasound imaging system with another imaging device, such as a high-speed camera, without decreasing the acquisition framerate, remains a difficult problem. In this article, a new acquisition system is introduced, which in addition to the acoustic signal, is able to synchronously record both ultrasound and optical streams at more than 60 fps on a single and "easy-to-transport" laptop-based machine.

Section 2 of the article presents briefly the state-of-the-art in ultrasound speech data acquisition via a

non-exhaustive list of existing solutions. Both hardware and software components of the proposed acquisition system, which is based on the *Terason T3000* portable ultrasound system coupled with an industrial camera and driven by the dedicated *Ultraspeech* application, are described in section 3. Finally, the capacity of the system to record synchronously multiple high-speed data streams is evaluated experimentally in section 4.

2 State of the Art in Multimodal Speech Data Acquisition System

Much of the progress that has been achieved in multimodal speech data acquisition systems concerns the positioning of the head and the transducer: These may be stabilized as in HATS [3] or related systems [4]; free to move using a helmet arrangement [5]; or tracked, using infrared emitting diodes (HOCUS [6]), visible markers (PALATRON [7]) or electromagnetic sensors [8].

The other crucial issue in such acquisition systems is the synchronisation of the ultrasound image stream with the uttered acoustic speech signal. In most systems ([3], [4], [6], [7]), this task is performed using an analog video mixer which downsamples the ultrasound data stream to 30 Hz. In the system proposed by Aron [8], ultrasound, electromagnetic and audio data are recorded synchronously with each modality keeping its own framerate. The synchronization of ultrasound data with audio is achieved in that system by time-aligning the ultrasound machine cineloop (a video buffer of the last 15 seconds recorded) with a timecode on an external PC.

3 Description of the Acquisition System

In the context of a silent speech interface based on tongue and lip imaging, the desired acquisition system should be able to record synchronously ultrasound data and video data at their respective maximum framerate together with the acoustic speech signal. In order to have a compact, transportable, and easy-to-use system, a PC-based hardware architecture coupled with a single control program has been adopted.

3.1 Hardware component of the system

As shown in figure 1, the hardware component of the system is based on:

- the *Terason T3000* ultrasound system which is based on a laptop running Microsoft Windows XP and provides 640x480 pixels resolution images
- a 140° microconvex transducer with 128 elements (8MC4)
- an industrial USB color camera able to provide 60 fps with a 640x480 pixels resolution (USB 2.0, WDM compliant)
- an external microphone connected to the built-in soundcard of the *T3000*

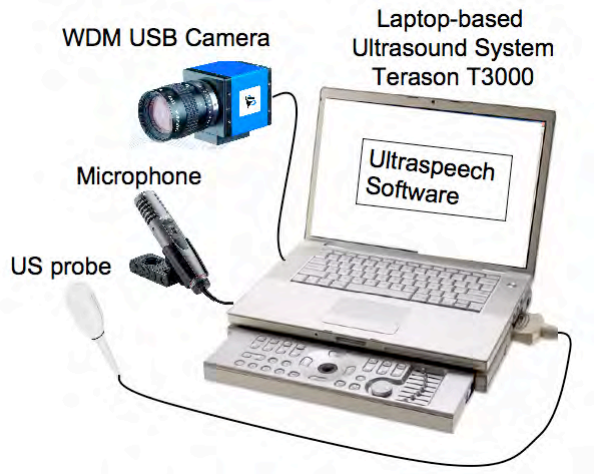


Figure 1: *Hardware component of the acquisition system*

In the described system, data streams are recorded, processed and stored digitally on a single PC using our stand-alone software *Ultraspeech*.

3.2 The *Ultraspeech* software

The open shared-memory client-server architecture of the *Terason T3000* system allows the development of stand-alone client applications with real-time access to the live stream of ultrasound images. The *Ultraspeech* MFC application (*Microsoft Foundation Classes*) is optimized for the *Terason T3000*, supports WDM compliant cameras (*Windows Driver Model*), and DirectX compatible soundcards. As shown in figure 2, *Ultraspeech* allows the real-time visualization of image streams and the automation of

the imaging devices. Internally, *Ultraspeech* uses multiple FIFO buffers to access image data.

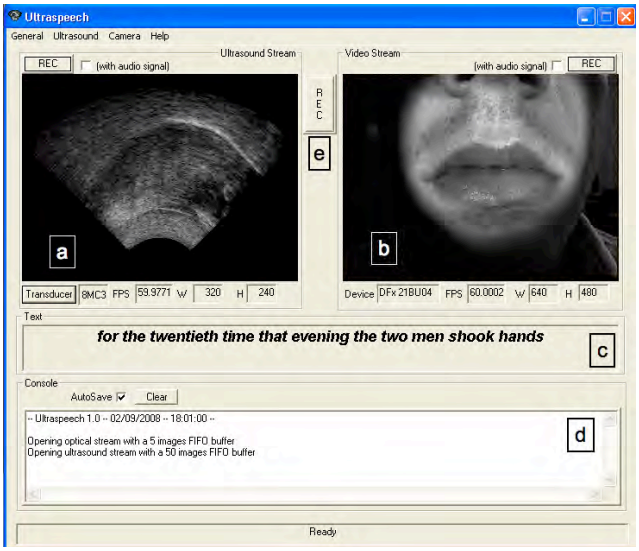


Figure 2: *Ultraspeech software (main window) with ultrasound and camera visualization area (a and b), text stimuli display area (c), console output (d) and start/stop recording button (e)*

The main feature of the *Ultraspeech* software is the synchronous recording of both image streams with the audio signal. Data recording is triggered by simply clicking on a start/stop button. Thanks to multithreading programming techniques, all streams are processed in parallel. Streams share the same multimedia timer so that each frame and each audio buffer can be tagged with the timer value during the recording. Any initial asynchrony between streams is captured during the acquisition, and synchrony is restored automatically in a post-processing stage. The entire recording procedure is fully automatic and no *a posteriori* human check is needed. After each acquisition, data are directly available as series of bitmaps for both image streams and WAV files for the audio stream, in the specified directory (local or remote). Furthermore, *Ultraspeech* provides convenient tools for large database recording, such as an automatic file naming system and the automatic display of the text stimuli for each item to record (*i.e* the word or sentence to pronounce).

3.3 Inter-session re-calibration procedure

Techniques involved in the silent vocoder described in [1] require the recording of a large amount of multimodal speech data. In our earlier work, data was recorded in a single long session during which the subject remained fixed in the HATS system. Data acquisition in multiple sessions (spaced in time) requires an inter-session re-calibration mechanism to position the speaker's head at a reference position (the probe remains fixed). The procedure shown in figure 3 is based on real-time averaging of a live image with a target reference image. During this interactive re-calibration procedure, the subject adjusts the position of his/her head in order to fit to the target reference position. A similar procedure is used for ultrasound, where the live tongue image is super-imposed on a target reference. When coupled with a head stabilization system, this procedure is convenient, rapid and effective.

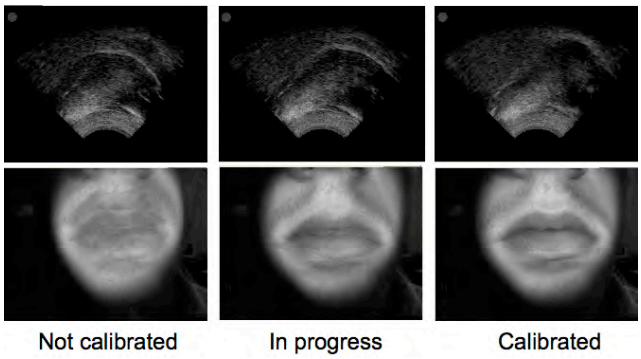


Figure 3: *Interactive speaker inter-session re-calibration mechanism at different stages of the procedure*

4 Experimental results

In order to check the synchronization of the different streams, the following experiment, illustrated in figure 4, is done. A hammer is used to tap a bottle of ultrasound gel (fig 4, a and b), ejecting a droplet onto the probe (c). The droplet shows up immediately on the probe (e) and should be synchronized with video of the hammer (d) hitting the bottle and the sound of this contact (f). Stream synchrony can be observed in figure 4 where a 71 fps ultrasound stream is displayed with a 60 fps video stream and the audio signal on the same time scale.

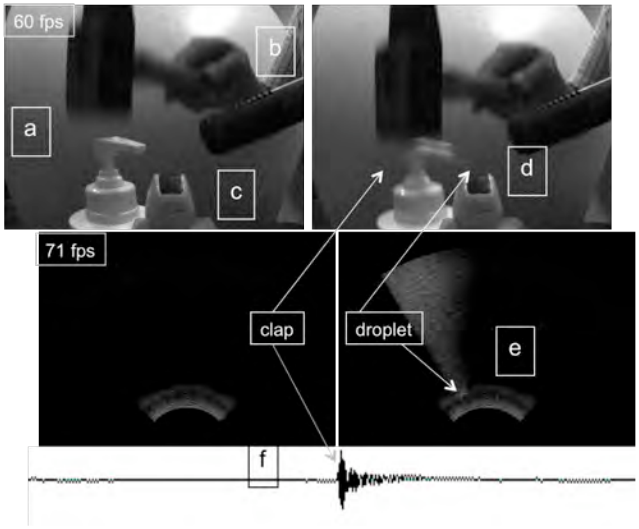


Figure 4: Interactive *speaker inter-session re-calibration mechanism (ultrasound is clearly synchronized with video and audio)*

This synchronization check procedure has been shown to be highly reproducible and has been tested on a variety of different ultrasound probes (*Terason 5MC2, 8MC3, 8MC4*). The ultrasound and video streams are found always to be synchronized. A residual delay occasionally observed between visual (ultrasound and video) and audio is always less than the inter-frame gap (*i.e* 15 ms at 60 fps). To summarize, the system is currently able to record synchronously:

- the ultrasound stream at **71 fps** (7cm depth, 320x240 pixels resolution, bitmap format)
- the video stream at **60 fps** (640x480 pixels resolution, bitmap format)
- the audio signal (44100 Hz, 16 bits, mono, PCM)

5 Conclusion and Perspectives

The conception of a silent speech interface based on tongue and lip imaging requires a high-speed acquisition system. The flexible PC-based architecture of the *Terason T3000* ultrasound system has allowed the development of *Ultraspeech*. This software interfaces ultrasound with a high-speed USB camera and an audio device. The system is able to record synchronously these different data streams which retain their respective framerates. For the recording of large databases in multiple acquisition sessions, an inter-session re-calibration procedure has also been introduced. The system has been used for the recording of small databases (100 words) and is now ready to be validated on a large dataset recording task.

6 Acknowledgements

This work is supported by the French Department of Defense (DGA) and the French National Research Agency, under contract number ANR-06-BLAN-0166.

7 References

[1] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, M. Stone, “Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips”, Interspeech, to appear, Brisbane, Australia, 2008.

[2] T. Hueber, G. Chollet, B. Denby, M. Stone, L. Zouari, “Ouisper: Corpus Based Synthesis Driven by Articulatory Data”, International Congress of Phonetic Sciences, Saarbrücken, pp. 2193-2196, Germany, 2007.

[3] M. Stone, “A guide to analyzing tongue motion from ultrasound images”, Clinical Linguistics and Phonetics, 19(6-7): pp 455–502, 2005.

[4] L. Davidson, “Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance”, Journal of the Acoustical Society of America 120:1, 407-415, 2005.

[5] A. Wrench, J. Scobbie, M. Linden, “Evaluation of a helmet to hold an ultrasound probe”, Ultrafest IV, NY, USA, 2007.

[6] D. Whalen, K. Iskarous, M. Tiede, D. Ostry, H. Lehnert-Lehouillier, E. Vatikiotis-Bateson, D. Hailey, “The Haskins optically corrected ultrasound system (HOCUS)”, Journal of Speech, Language, and Hearing Research, 48(3): pp 543–553, 2005.

[7] J. Mielke, A. Baker, D. Archangeli, S. Racy, “Palatron: A Technique for Aligning Ultrasound Images of the Tongue and Palate”, in D. Siddiqi, and B. V. Tucker, Eds., Coyote Papers. vol. 14. 97-108, 2005.

[8] M. Aron, N. Ferveur, E. Kerrien, M.O. Berger, Y. Laprie, “Acquisition and synchronization of multimodal articulatory data”, Interspeech, Antwerp, Belgium, 2007.

Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface

Thomas Hueber [1,3], *Elie-Laurent Benaroya* [1], *Gérard Chollet* [3], *Bruce Denby* [2,1], *Gérard Dreyfus* [1], *Maureen Stone* [4]

[1]Laboratoire d'Electronique, Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI ParisTech), 10 rue Vauquelin, 75231 Paris Cedex 05 France
[2]Université Pierre et Marie Curie – Paris VI, 4 place Jussieu, 75252 Paris Cedex 05 France
[3]Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, 46 rue Barrault, 75634 Paris Cedex 13 France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, 666 W. Baltimore Street, Baltimore, MD 21201 USA

hueber@ieee.org, laurent.benaroya@espci.fr, gerard.chollet@tsi.enst.fr, denby@ieee.org, gerard.dreyfus@espci.fr, mstone@umaryland.edu

Abstract

Recent improvements are presented for phonetic decoding of continuous-speech from ultrasound and optical observations of the tongue and lips in a silent speech interface application. In a new approach to this critical step, the visual streams are modeled by context-dependent multi-stream Hidden Markov Models (CD-MSHMM). Results are compared to a baseline system using context-independent modeling and a visual feature fusion strategy, with both systems evaluated on a one-hour, phonetically balanced English speech database. Tongue and lip images are coded using PCA-based feature extraction techniques. The uttered speech signal, also recorded, is used to initialize the training of the visual HMMs. Visual phonetic decoding performance is evaluated successively with and without the help of linguistic constraints introduced via a 2.5k-word decoding dictionary.

Index Terms: silent speech interface, visual speech recognition, multi-stream modeling

1. Introduction

Designing a device to allow speech communication without the necessity of vocalizing has become a challenge in its own right in the speech research community. This "Silent Speech Interface", or SSI, could be used to preserve the privacy of conversations, communicate in silence-restricted or high background noise environments, or for silent, hands-free data transmission during a security operation. Further applications are possible in the medical field, for example to assist laryngectomized patients, where the SSI would be used as an alternative to the electrolarynx; to oesophageal speech, which is difficult to master; or to tracheo-oesoephageal speech, which requires additional surgery. Different types of sensors can be envisaged in order to build an SSI. A speaker may for example produce airflow in his vocal tract and capture the resulting "murmur" with a stethoscopic microphone as in [1] and [2]. Other approaches, based on completely non-acoustic features have also been proposed, as for example in [3] where electromyographic electrodes placed on the speaker's face record muscular activity, or in [4] where magnets glued to the tongue and lips are tracked by sensors incorporated in a pair of eyeglasses. In our approach, articulator activity, mainly of the tongue and lips, is captured by a non-invasive multimodal imaging system composed of an ultrasound transducer placed beneath the chin and an optical camera in front of the lips [5].

In [6], we presented a framework for a phonetic vocoder driven exclusively by streams of visual observations, using an audio-visual unit dictionary that associates acoustic utterances with their visual phone equivalents. In the first stage of the system, the visuo-phonetic decoder finds the most likely phonetic targets for a given test sequence of visual data. These targets then constrain the selection in the dictionary of the sequence of units that best matches the input test data. In such a corpus-based approach, the quality of the synthesis depends strongly on the performance of the phonetic decoding stage, whose robustness must therefore be maximized. To that end, more sophisticated HMM-based modeling techniques have been recently tested. The two improvements presented and evaluated in the present paper are: the introduction of context-dependency in the modeling of the visual phones; and the use of a multi-stream approach to model jointly the ultrasound and the optical data streams. Systems derived from this approach will be compared to a baseline decoder, similar to that used in [6], which uses context-independent phonetic models and a feature fusion strategy. Because it is not *a priori* feasible to disambiguate all phonetic configurations only from tongue and lip observations, linguistic constraints can also be introduced to help the phonetic decoding, *via* for instance, a restriction on the allowed vocabulary. We therefore also evaluate our systems on both an unconstrained phonetic decoding task and on a more restricted one.

The development of the visuo-phonetic decoding baseline system is detailed in Section 2, where data acquisition and pre-processing, visual feature extraction techniques and evaluation protocols are described. Section 3 addresses the implementation and evaluation of the context-dependent visual phonetic decoder. In section 4, the multi-stream modeling approach is introduced and evaluated. Also in that section, the performance of the final system including both context dependent modeling and the multi-stream approach is discussed.

2. Baseline Visuo-Phonetic Decoder

2.1. Data acquisition and pre-processing

Ultrasound data is recorded using the Vocal Tract Visualization Lab HATS system [7]. In this setup, the transducer is locked in a fixed position beneath the chin, and

the head immobilized. An acoustic standoff is used to allow mandible motion so that speech production is relatively undisturbed. Two standard video cameras record both profile and frontal views of the speaker's lips, and a microphone captures the uttered speech signal. The three video streams (two cameras plus ultrasound) and the audio signal are merged into the same video sequence using an analog video mixer, which limits the frame rate of the video data to 29.97 Hz (NTSC format). A typical image recorded by this acquisition system is shown in figure 1.

Figure 1: *An ultrasound vocal tract image in the mid-sagittal plan with embedded lip frontal and lateral view. Dashed white lines represent tongue and lip regions of interest.*

The text material, chosen for the purposes of diphone-based concatenative synthesis, is based on the first 1020 sentences of the CMU Arctic corpus [8], read by a native speaker of American English instructed to speak as neutrally as possible. After cleanup of the recordings, the database contains 61 minutes of speech contained in 109553 bitmap frames. Audio files are sampled at 16 kHz.

2.2. Visual feature extraction

Regions of interest (ROI) selected in ultrasound and optical images, as shown in figure 1, are first resized to 64x64 pixels. Speckle noise typical of ultrasound images is reduced using the anisotropic diffusion filter described in [9]. It is suggested in [10] that a frontal view of the lips provides more articulatory information than a profile; thus, although both are present in our database, we chose to use only the frontal view in this study. The "EigenTongues" [11] decomposition technique is subsequently used to encode each ultrasound frame. In this method, the vocal tract configuration is interpreted as a linear combination of standard configurations, the "EigenTongues", obtained by performing a Principal Component Analysis (PCA) on a phonetically balanced subset of frames. A similar technique is used to encode frontal images of the lips ("EigenLips"). The numbers of projections onto the set of EigenTongues/EigenLips used for coding are obtained empirically by evaluating the quality of the image reconstructed from its first few components; typical values used on this database are 30 coefficients for each of the two streams. In order to be compatible with a more standard frame rate for speech analysis, the EigenTongues/EigenLips coefficient sequences are oversampled from 30 Hz to 100 Hz using linear interpolation. In this baseline system, EigenTongues/EigenLips coefficients, together with their first and second derivatives, are concatenated into a single "visual feature vector" in a feature fusion strategy. Dimensionality reduction techniques were not used in this analysis; their application is under study and will be addressed in a future work.

2.3. HMM-based modeling

The modeling of visual feature sequences by continuous HMMs requires their initial temporal decomposition at the phonetic level. As visual and audio modalities have been recorded synchronously, this initial segmentation can be derived from the labeling of the acoustic signal. This task is performed using a forced alignment procedure with an initial set of 40 acoustic HMMs trained on the acoustic component of the recorded database. The acoustic wave of each recorded sentence is parameterized by 12 Mel-frequency cepstral coefficients (MFCC) with their energies and first and second derivatives. In this study, all the procedures involving HMM manipulations are done using the HTK front-end [12]. After initialization, 40 left-to-right (monophones), 5-state (3 emitting states), continuous visual HMMs (with diagonal covariance matrices) are first trained separately using the standard Baum-Welch re-estimation algorithm. Then, embedded training, during which the number of Gaussians per state is incrementally increased, is used to refine the models and the temporal segmentation of the visual stream. In the testing stage, phonetic decoding is performed using the standard "Token Passing" algorithm, which finds the optimal path through an HMM network. Because some very important sources of information are missing in the visual data, such as nasality and the voiced/unvoiced flag, linguistic constraints can be introduced to help the phonetic decoding. With that in mind, we introduce two decoding scenarios. In the first, considered "unconstrained", the structure of the decoding network is a simple loop in which all phones loop back to each other. In the second, or "constrained" scenario, the phonetic decoder is forced to recognize words contained in the CMU Arctic sentences. In that case, the decoding network allows all possible word combinations which can be built from a 2.5k word dictionary. No statistical language model is used in the present study.

The 1020 sentences of the recorded database are divided into 34 lists of 30 sentences. In order to increase the statistical relevance of the speech recognizer performance, a jackknife (leave-one-out) technique [13] was employed, in which each list was used once as the test set while the other 33 lists composed the training set. Two test lists were however excluded from this jackknife procedure to be used as a validation set for the optimization of two "hyper" parameters: the model insertion penalty; and the number of Gaussians per state of the visual HMMs. For the baseline system, the optimal number of Gaussian per state was found to be 32.

For each phone class, a representative measure P of the recognizer performance is defined as:

$$P = \frac{N - D - S - I}{N} \quad (1)$$

where N is the total number of phones in the test set, S the number of substitution errors, D deletion errors, and I insertion errors. Section A of table 1 presents the performance of the baseline visuo-phonetic decoder in the two decoding scenarios.

3. Context-Dependent Modeling

Articulatory features such as those derived from the recorded images of the tongue and lips are naturally sensitive to context effects such as co-articulation and anticipation. Introducing

context-dependency in the modeling of visual features sequences should therefore increase the robustness of the visuo-phonetic decoding. In this study, we propose to model visual triphones by adding information about left and right contexts to the phone models. Traditionally, triphone modeling presents several practical issues. Since many triphones have only a few occurrences in the training data, the accurate estimation of their corresponding HMM parameters is difficult. Also, many triphones may be missing in the training corpus, especially in a relatively small dataset such as the one used in this study. To overcome these issues and make visual triphone training viable, a tree-based state-tying strategy is adopted. Using the procedure in [14], a binary decision tree is constructed for each state of each phone, in order to cluster together all of the corresponding states of all of the associated triphones. The decision tree recursively partitions this pool of states by querying left/right contexts. States reaching the same leaf node are considered similar enough to be tied together.

The "yes - no" questions associated with the tree nodes are usually based on phonetic knowledge such as backness and height for vowels, place and manner of articulation for pulmonic consonants, etc. A typical question attached to a node of the decision tree might be, "Is the previous phone (left context) a bilabial consonant ?". However, as tongue and lip configurations are explicitly represented here, we propose to use a gesture-based approach to build the contextual questions. A feature set in which tongue body, tongue tip, and lip configurations are described explicitly, using the articulatory phonology theory introduced in [15], is used. With this description, the articulatory configuration corresponding to the phone [sh], for instance, would be characterized as a configuration where the lips are in a default "labial" position, the tongue tip in the "palato-alveolar" region, "tongue body" in the "palatal" region, etc. A typical contextual question on the decision tree built from this feature set would be, "Does the next phone (right context) require the tongue body moving to the palatal region?" However, although present in the feature set, no questions based on the glottal activity (which is meaningless in a silent speech context) or on the velum (which is not visible in an ultrasound image) was built.

To build the context-dependent visual phonetic decoder, a set of 40 visual HMMs (monophones) is first trained using the same procedure as for the baseline system. These monophone models are then cloned to initialize their corresponding untied triphones. As each training set of the jackknife procedure contains approximately 8500 distinct triphones with, on average, only 4 occurrences apiece, state tying appears to be essential. After the tree-based clustering procedure, the total of 25500 states (8500 x 3 emitting states) is reduced roughly to 1800 clusters (~7% of the original number of states). Tied-state models are then refined by incrementally increasing the number of Gaussian mixture components up to an optimal number which was found to be 4. Finally, models for unseen triphones are generated; decision trees are asked to find which combination of already trained state models is the most adapted to represent the context of a given unseen triphone. At the end of the training stage, a set of 67200 visual triphone models (all possible triphones and biphones built from a 40 element phone set) is available for decoding. As for the baseline system, decoder performance is evaluated on both unconstrained (free phonetic decoding) and constrained (using a 2.5k word dictionary) scenarios, as shown in sections A and B of Table 1. Compared to the baseline system, performance of the context-dependent system is significantly improved in both decoding scenarios, with improvements of 4.9% (unconstrained) and 3.2% (constrained) respectively. A more detailed analysis of the remaining decoding errors is given at the end of the next section.

4. Multi-Stream *vs.* Feature Fusion

As in audio-visual speech recognition, two approaches can be envisioned to integrate tongue and lip data streams in an HMM-based phonetic decoder: "feature fusion" which was adopted in the baseline system; and "(classifier) decision fusion". As described in [16], different strategies can be used to combine modalities at the classifier level. In this work, an "early integration" strategy based on state-synchronous Multi-Stream Hidden Markov Models (MSHMM, [17]) is used to model tongue and lip feature sequences. In a MSHMM, each stream has, for each state, its own Gaussian mixture and thus its own emission probability density function. Given a "tongue (T) and lips (L)" visual observation vector $o_t^{TL} = [o_t^T; o_t^L]$, the resulting emission likelihood b_j for state j is expressed as:

$$b_j(o_t^{TL}) = \prod_{S \in \{T,L\}} \left[\sum_{m_S=1}^{M_S} c_{jSm_S} N(o_t^S; \mu_{jSm_S}; \Sigma_{jSm_S}) \right]^{\lambda_S} \quad (2)$$

where $N(o;\mu;\Sigma)$ is the value at o of a Gaussian mixture with mean μ and covariance Σ, M_S the number of mixture components and $\lambda_S = \{\lambda_T, \lambda_L\}$ are the weight parameters discussed below. In this equation, stream components are forced to be state-synchronous and thus asynchrony between tongue and lips movements, well described in [18], is not taken into account. However, since asynchrony is often correlated with phonetic context, the use of context-dependent models could potentially compensate this phenomenon. The combination of the stream likelihoods also requires the definition of the weight parameters λ_T and λ_L. Widely discussed in the context of audiovisual speech recognition (AVSR) [16], the estimation of stream exponents can be achieved either by measuring stream reliabilities using an SNR or a "degree of voicing" criterion for the audio modality, which is not possible here, or by maximizing system performance on a validation data set. In this initial test of multi-stream modeling of tongue and lip data, a very simple optimization procedure is adopted: only class-independent weights are used, and system performance is evaluated on a validation set for different pairs of weights, which we constrain to sum to one. As expected, the tongue carries the most important part of the accessible articulatory information, and the optimal values found for tongue and lip feature streams are $\lambda_T = 0.7$ and $\lambda_L = 0.3$.

In our procedure, a multi-stream phonetic decoder using context-independent models is first trained using the same procedure as for the baseline system. Its performance is shown in section C of Table 1. Compared to the baseline system, the multi-stream approach brings a 2% improvement in the unconstrained decoding scenario (with fewer substitution and insertion errors but more deletion errors), and a 3,7% improvement in the constrained one. Also, when the multi-stream approach is combined with context-dependent modeling, as in the "final" system whose performance is shown in section D of Table 1, the performance improvement is about 8% higher than that of the baseline system. While still not ideal, these results are nonetheless promising and demonstrate the relevance of the two new adopted strategies.

Quite naturally, most of the substitution errors are made on phones with similar tongue and lip gestures, such as

Table 1. *Performance of the different visuo-phonetic decoders. Δ is the 95% confidence interval, "CI", "CD", "Unconst.", and "Const." stand for context-independent and context dependent models, unconstrained and constrained decoding scenarios, respectively.*

	A		**B**		**C**		**D**	
	Baseline Decoder *CI – Feature Fusion*		**Context-dependent Decoder** *CD – Feature Fusion*		**Multi-Stream Decoder** *CI – MSHMM*		**Context-dependent & Multi-Stream Decoder** *CD - MSHMM*	
	Unconst.	**Const.**	**Unconst.**	**Const.**	**Unconst.**	**Const.**	**Unconst.**	**Const.**
P	57,7%	67,4%	62,6%	70,6%	59,5%	71,1%	65,6%	74,7%
Δ	1.0%	1.0%	1,0%	1,0%	1.0%	1,0%	1.0%	1.0%
D	6043	4666	3452	3196	7531	5174	4294	3964
S	7077	5398	7080	4799	5897	4157	6279	3613
I	1568	1270	2451	2210	658	696	1397	1232
N	34693	34693	34693	34693	34693	34693	34693	34693

{[p],[b],[m]}, {[t],[d],[n]}, {[f],[v]}, {[k],[g],[ng]}, {[ch],[jh]}, {[sh],[zh]} and {[th],[dh]}. In fact, if we consider these phone groups as equivalence classes, in which within-group confusions are not counted as errors, the performance of the final system in the unconstrained scenario can be further increased to 73,2% (78% for the constrained decoding scenario). Most of the remaining substitution errors are due to: vowels confused with the phone [ah], which is, in continuous speech, certainly a consequence of the vowel reduction effect; diphthongs matched sometimes with one of their vowel components; and dental and alveolar consonants, which are difficult to image with ultrasound because the apex (tongue tip) may be hidden by the acoustic shadow of mandible. Some of these mismatches in the phonetic decoding would not necessarily lead to unintelligible synthesis; some psychoacoustic effects could potentially also be used to advantage. The relatively high number of deletion and insertion errors, however, remains problematic, and will continue to be addressed in future work.

5. Conclusions

In order to improve the visuo-phonetic decoding stage of a planned ultrasound-based silent speech interface, the modeling of tongue and lips feature sequences using multi-stream and context-dependent HMMs has been proposed. On an open-vocabulary continuous speech decoding task, the system is able to correctly identify 65,6% of the phones from visual information only. When the vocabulary is limited to 2.5k words, the performance increases to 74,7%. Compared to a baseline system based on context-independent models and a feature fusion strategy, this new approach has led to a 8% absolute performance improvement. To reduce the remaining decoding errors (mainly deletions and insertions), the recording of visual data at a higher frame rate and the use of a statistical language model are currently under study.

6. Acknowledgements

This work is supported by the French Department of Defense (DGA) and the French National Research Agency (ANR), under contract number ANR-06-BLAN-0166.

7. References

[1] Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., "Non-audible murmur recognition", Proc. of Eurospeech, pp. 2601-2604, 2003.

[2] Tran V.-A., Bailly G., Lœvenbruck H., Jutten C., "Improvement to a NAM captured whisper-to-speech system", Interspeech, Brisbane, Australia, pp. 1465-1498, 2008.

[3] Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., "Session independent non-audible speech recognition using surface electromyography", IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 331-336, 2005.

[4] Fagan, M.J., Ell, S.R, Gilbert, J.M., Sarrazin, E., Chapman, P.M., 2008. Development of a (silent) speech recognition system for patients following laryngectomy. Medical Engineering & Physics, vol. 30, issue 4, pp. 419-425.

[5] Denby, B., Oussar, Y., Dreyfus, G., Stone, M., "Prospects for a Silent Speech Interface Using Ultrasound Imaging", IEEE ICASSP, Toulouse, France, pp. I365- I368, 2006.

[6] Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., "Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips", Interspeech, pp. 2028-2031, Brisbane, Australia, 2008.

[7] Stone, M., and Davis, E., "A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement", Journal of the Acoustical Society of America, vol. 98 (6), pp. 3107-3112, 1995.

[8] Black, A. W., Lenzo, K., "Building voices in the Festival speech synthesis system", http://festvox.org/bsv, 2000.

[9] Y. Yu and S. T. Acton, "Speckle Reducing Anisotropic Diffusion", IEEE Trans. on Image Proc., vol. 11, pp. 1260-1270, 2002.

[10] Lucey, P., Potamianos, G., "Lipreading using profile versus frontal views", in Proc. IEEE MMSP'06, Canada, pp. 24-28, 2006.

[11] Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M., "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface", IEEE ICASSP, Honolulu, pp. I1245-I1248, 2007.

[12] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., The HTK Book, September 2005, http://htk.eng.cam.ac.uk/.

[13] Efron, B., "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods", Biometrika, vol. 68, pp. 589-599, 1981.

[14] Young, S. J., Odell, J. J., Woodland, P. C. "Tree-based state tying for high accuracy acoustic modeling", In Proc. of the Workshop on Human Language Technology, pp. 307-312, 1994.

[15] Browman, C. P., Goldstein, L., "Gestural specification using dynamically-defined articulatory structures", Journal of Phonetics, 18, pp. 299-320, 1990.

[16] Potamianos, G., Neti, C., Luettin, J., Matthews, I., "Audio-Visual Automatic Speech Recognition: An Overview". In: Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press, 2004.

[17] Bourlard, H., Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands", In Proc. ICSLP, pp. 426-429, 1996.

[18] Livescu, K., Glass, J., "Feature-based pronunciation modeling for speech recognition", in Proc. HLT/NAACL, 2004.