Table des matières

Table des illustrations

Liste des acronymes utilisés dans le manuscrit

PSOLA	*Pitch Synchronous Overlap and Add*
MFCC	*Mel-frequency cepstral coefficients*
LF	Liljencrants-Fant
CALM	*Causal-Anticausal Linear Model*
ODG	Onde de Débit Glottique
ODGD	Onde de Débit Glottique Dérivée
LPC	*Linear Predictive Coding*
EGG	Électroglottographe
dEGG	Signal électroglottographique dérivé
HBB	*Human Beat Box*
DL	*Deep Learning*
DBN	*Deep Belief Network*
RBM	*Restricted Boltzmann Machine*
CD	*Contrastive Divergence*
DAE	*Deep Autoencoder*
HMM	*Hidden Markov Model*
HTS	*HMM-to-Speech*
TTS	*Text-to-speech*
tRBM	*Translational RBM*
MSD	*Mean Sum of Distances*
LSF	*Line Spectral Frequencies*
FFT	*Fast Fourier Transform*
OFR	*Orthogonal Forward Regression*
MLP	*Multilayer Perceptron*
PCA	*Principal Component Analysis*

Introduction générale

Les technologies de l'information et de la communication peuvent être utilisées afin de diffuser, valoriser et préserver le patrimoine culturel. En particulier, les techniques de chant mettent en œuvre des savoir-faire complexes ; certaines techniques de chant ont jusqu'à ce jour été transmises principalement oralement. Ces savoir-faire restent mal compris, et sont par conséquent fragiles et menacés dans certains cas (comme par exemple les polyphonies corses dont nous parlerons dans ce manuscrit). Ce manuscrit tente d'apporter des éléments de réponse à la question suivante : peut-on modéliser le savoir-faire des chanteurs ?

Dans ce but, nous étudierons l'application des techniques d'ingénierie biomédicale pour l'apprentissage et la sauvegarde des techniques de chant rares. Nous nous intéressons en particulier aux mouvements des articulateurs impliqués dans la production vocale. Afin de visualiser les mouvements de la langue de façon non invasive, nous utilisons une sonde échographique placée sous la mâchoire inférieure. Avec ce dispositif nous obtenons une vue sagittale et temps réel de la langue d'un chanteur. Par ailleurs, nous avons souhaité multiplier les modalités d'enregistrement de données sur ces techniques de chant et les combiner par diverses méthodes de traitement du signal et de modélisation. Cependant, les images échographiques sont des images sombres, bruitées et illisibles pour une personne inexpérimentée en lecture d'images échographiques. Nous souhaitons par conséquent fournir une version augmentée de l'image échographique de la langue, plus lisble que l'image brute à un élève chanteur, dans un contexte de protocole d'apprentissage par retour visuel articulatoire (*biofeedback*). En premier lieu, notre approche consiste à extraire ded façon automatique le contour de la surface supérieure de la langue observable dans des images échographiques, ce qui nous permet d'obtenir des informations plus lisibles sur la position de la langue. Ces informations sont utiles, et pourraient trouver des applications pédagogiques pour une personne désireuse d'apprendre à positionner sa langue correctement. Dans [1], les auteurs présentent en effet un programme de rééducation utilisant des images échographiques pour des enfants atteints d'apraxie verbale, c'est-à-dire un trouble de l'acquisition des gestes permettant l'articulation du langage et des difficultés à la planification des mouvements impliqués dans la parole. Cette étude montre une amélioration des performances articulatoires des enfants testés pendant et deux mois après les sessions d'entraînement. Ce type de méthodes pédagogiques d'apprentissage de la prononciation par retour visuel échographique a

également été présenté dans [2] et [3]. Pour des questions de temps de traitement, nous souhaitons déterminer le contour de la langue de façon automatique. Cependant, la plupart des méthodes d'extraction de contour automatique voient leurs performances se dégrader au cours du temps. Notre intuition est qu'une méthode utilisant un apprentissage statistique permettrait de s'affranchir de ce problème. En particulier, nous supposons qu'un apprentissage sur une base de données de grande taille fournirait une information suffisante pour modéliser la position de la langue et extraire les paramètres du conduit vocal. Nous avons donc utilisé des bases de données avec plusieurs milliers d'images. Compte tenu de la diversité des informations contenues dans les images échographiques, nous faisons l'hypothèse que l'apprentissage profond (*deep learning*) nous permettra d'extraire des descripteurs synthétisant la complexité de la base d'images pour effectuer cette modélisation. Nous avons par ailleurs émis l'hypothèse assez forte que la base d'apprentissage pourrait être automatiquement étiquetée par un outil paramétrable détectant les contours de la langue en utilisant le contraste dans l'image. Cet outil possède de bonnes qualités de détection de contour, à condition d'effectuer un seuillage adapté à la qualité des images. Une des perspectives de ce travail est d'appliquer ces méthodes de retour visuel échographique au domaine de l'apprentissage du chant. Une autre approche pédagogique en termes de biofeedback vocal consiste à piloter une tête parlante, comme décrit dans les travaux de [4], [5], [6] et [7].

En plus du contour de la langue, d'autres informations articulatoires peuvent permettre la construction d'un modèle acoustique du conduit vocal. En particulier, la combinaison des informations fournies par la langue et les lèvres permet de distinguer la plupart des sons. Les membres de mon équipe de recherche ont émis l'hypothèse qu'il était possible de reconstruire un modèle acoustique du conduit vocal en utilisant des informations extraites à partir des images de langue et de lèvres. Au cours de ma thèse, j'ai imaginé une méthode de synthèse vocale utilisant un modèle de source dont les paramètres sont extraits grâce à une mesure de l'activité glottique par un électroglottographe, et un modèle de filtre utilisant les coefficients de prédiction linéaire, déterminés par apprentissage statistique à partir des images des articulateurs. Ce type de synthèse nécessite l'extraction de descripteurs à partir des images des articulateurs. Puisque la relation entre ces deux types de données et le spectre de prédiction linéaire ne semble pas évidente, nous avons émis l'hypothèse qu'un modèle d'extraction de descripteurs non linéaire et cherchant des régularités dans les données sera plus performant

qu'un modèle linéaire. Par conséquent, nous avons comparé deux méthodes de prédiction des coefficients du filtre du conduit vocal : la première utilise un autoencodeur permettant d'extraire des descripteurs issus de la combinaison des deux modalités que sont les images de langue et de lèvres, tandis que la seconde utilise une représentation des données obtenues par projection dans un sous-espace en utilisant un modèle linéaire. Nous avons souhaité comparer les spectres reconstruits ainsi que l'audio synthétisé à l'aide des données articulatoires. Notre travail présente donc une méthode complète de synthèse de voix chantée à partir des données d'imagerie articulatoires et les signaux glottiques.

Ce manuscrit est organisé en trois parties. Nous présentons dans un premier temps le contexte de la thèse : les mécanismes de production vocale, des modèles de synthèse vocale, les méthodes d'enregistrements de données articulatoires, les techniques de chants rares étudiés dans le cadre du projet, des notions d'apprentissage statistique et nos outils d'acquisition de données. Ensuite, nous présentons une première méthode permettant d'utiliser les informations articulatoires à des fins pédagogiques en extrayant de façon automatique le contour de la langue à partir des images échographiques. Nous détaillons dans un premier temps quelques techniques semi-automatiques et automatiques d'extraction du contour de langue à partir des images échographiques. Ensuite, nous mettons en œuvre une méthode reposant sur des principes d'apprentissage statistique et comparons la qualité du contour extrait à ce qu'un étiquetage manuel ou automatique permettrait d'obtenir. Dans notre troisième partie, nous combinons des informations extraites des images de la langue à des informations extraites des images des lèvres afin de proposer un nouveau modèle de synthèse vocale en voix chantée. Dans cette partie, nous détaillons la nature et le calcul des variables à prédire, l'utilisation de l'apprentissage statistique pour l'extraction multimodale de descripteurs et la prédiction des paramètres articulatoires, puis les méthodes de synthèse vocale que nous utilisons et enfin les résultats obtenus sur différents types de données. Nous présentons les résultats et les comparaisons entre les différentes méthodes en combinant mesures objectives et résultats perceptifs sur des sujets.

1 Contexte

Ce travail de thèse s'inscrit dans un contexte pluridisciplinaire autour de la préservation du patrimoine culturel immatériel. Dans cette partie, nous exposons les mécanismes de production vocale, puis nous détaillons différents modèles pour l'analyse et la synthèse de la voix humaine. Nous détaillons par la suite les méthodes d'enregistrement de données articulatoires, en vue de leur application à l'étude des chants rares, puis nous décrirons les techniques de chants rares concernées par le projet. Nous présentons ensuite des méthodes d'apprentissage statistiques pouvant être appliquées dans notre contexte. Enfin, nous décrivons le matériel et les méthodologies nous permettant d'enregistrer des données illustrant les gestes vocaux des différents experts.

1.1 La production vocale

1.1.1 Description de l'appareil phonatoire

Pour modéliser les mécanismes du chant, il faut comprendre comment la voix humaine est engendrée par le corps. Nous allons par conséquent définir l'anatomie de l'appareil phonatoire. Celui-ci comporte deux étages : le premier est constitué des organes dits de soufflerie ; le second, appelé conduit vocal, est constitué du larynx et des résonateurs (voir Figure 1). Les organes de soufflerie émettent l'air impliqué dans la phonation. Le conduit vocal définit le trajet de l'air impliqué dans la phonation en sortant des poumons, du larynx jusqu'au nez en passant par le pharynx et les lèvres.

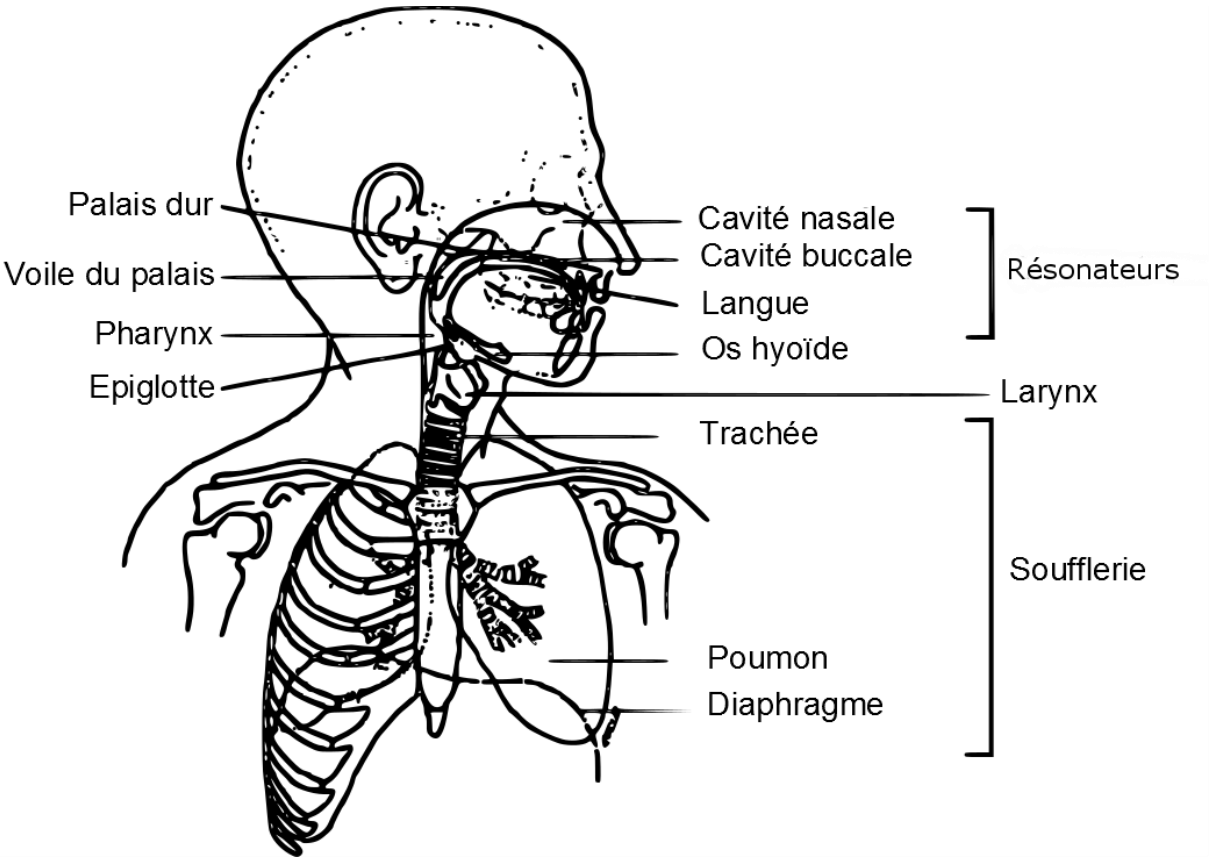


Figure 1 - Anatomie de l'appareil phonatoire, d'après [8].

Le larynx occupe une place centrale dans la production vocale. C'est au niveau du larynx que sont situés les plis vocaux (plus communément appelés cordes vocales, bien que cette image ne soit pas en accord avec la réalité anatomique, comme le montrent la Figure 2 et la Figure 3). Le larynx est constitué de cartilages recouverts de tissus mous. Parmi ces cartilages, le cartilage cricoïde, en forme d'anneau, se trouve dans le prolongement de la trachée. Les cartilages aryténoïdes et le cartilage thyroïde sont reliés au cartilage cricoïde. La fermeture du larynx est contrôlée par l'abaissement du cartilage épiglottique (aussi appelé épiglotte), lui-même relié au cartilage thyroïde. Les plis vocaux sont attachés à la fois à la base de l'épiglotte et aux pointes intérieures des cartilages aryténoïdes. L'air expulsé par les poumons provoque la vibration des plis vocaux, ce qui permet de produire les sons de la voix. Ces sons résonnent ensuite au niveau des cavités buccale et nasale. Les mouvements de ces résonateurs permettent de produire des modes d'articulation différents donnant accès à une grande variété de sons. Nous utiliserons le terme conduit vocal pour nous référer au larynx, ainsi qu'aux différents résonateurs et articulateurs. L'espace situé entre les plis vocaux se nomme glotte.

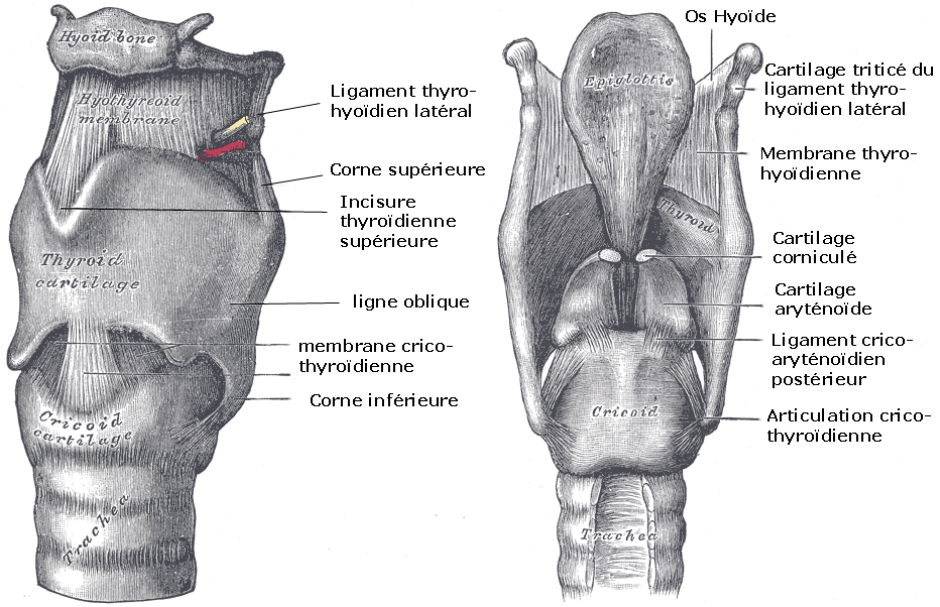


Figure 2 - Vues antérieures et postérieures du larynx, d'après [9].

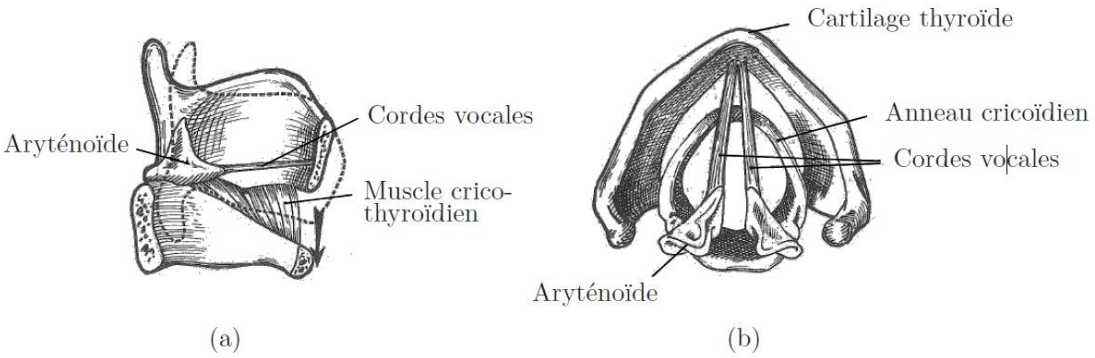


Figure 3 - Représentation des principaux cartilages et muscles du larynx. (a) vue de côté, (b) vue de dessus. D'après [10].

Les mouvements de translation et de rotation des cartilages aryténoïdes (voir Figure 4) permettent la mise en contact des plis vocaux. La pression de l'air provenant des poumons sur la glotte (pression sous-glottique) entraîne l'ouverture de la glotte. Ainsi, la pression sous-glottique diminue suite au passage de l'air, ce qui entraîne, grâce à l'élasticité des plis vocaux, une nouvelle fermeture de la glotte. Les ouvertures et fermetures glottiques se reproduisent ainsi de façon cyclique afin de produire des vibrations à l'origine de la phonation. La position de respiration correspond à un mouvement d'abduction des plis vocaux, au cours de laquelle l'air peut circuler librement.

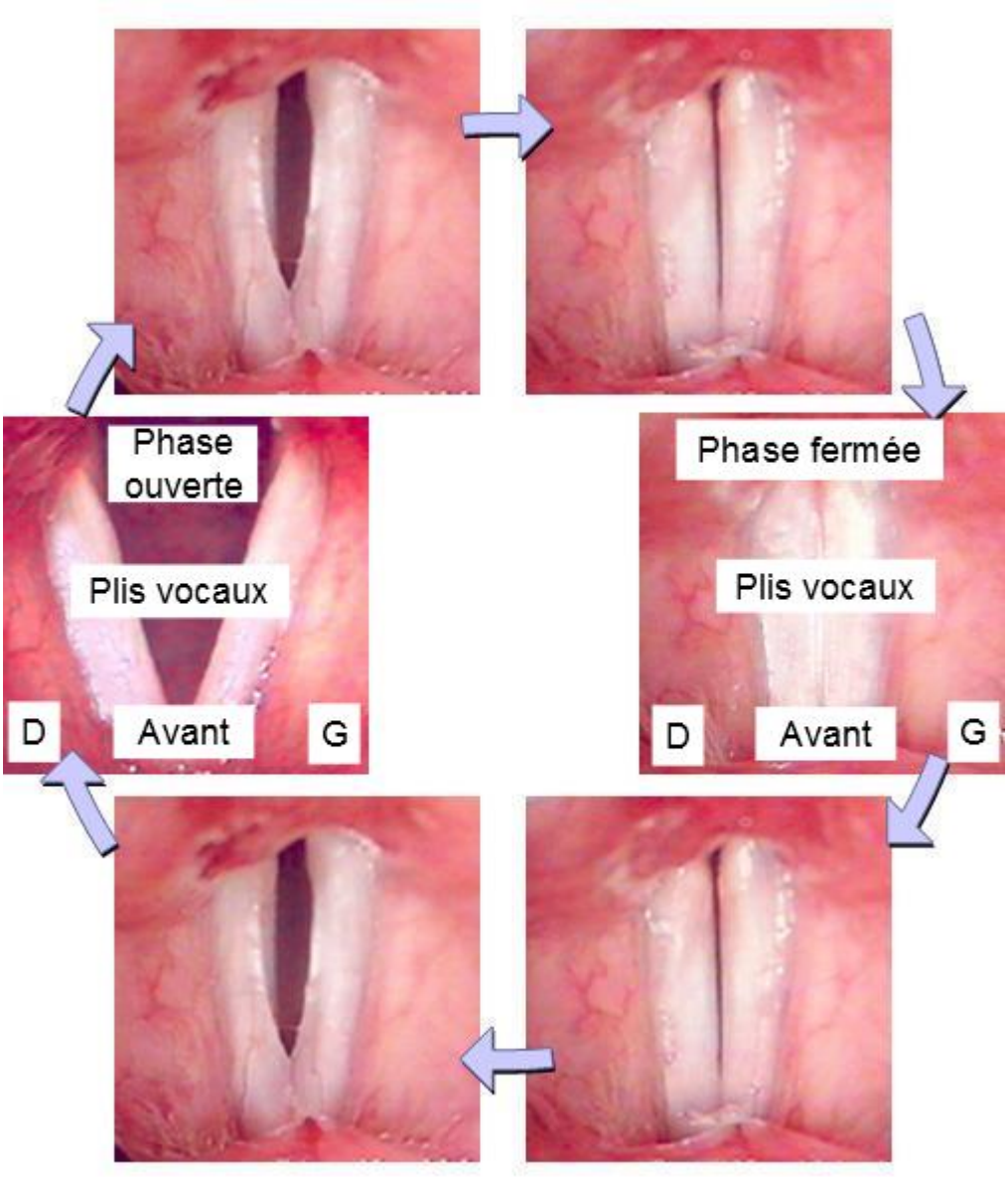


Figure 4 - Vue du dessus des plis vocaux au cours d'un cycle d'ouverture-fermeture glottique. D'après [11].

Au-dessus des plis vocaux se trouvent les bandes ventriculaires, bourrelets qui ont une apparence semblable aux plis vocaux mais ne possèdent pas de muscle interne. Les bandes ventriculaires interviennent dans certains types de chant, notamment le chant sarde mais aussi certains chants d'Asie (Tibet, Mongolie). La maîtrise du geste vocal implique un contrôle sur les différents muscles et cartilages présentés ci-dessus, afin d'adapter la géométrie du conduit vocal à la production vocale souhaitée.

1.1.2 Contrôle de la production de parole

L'être humain possède la capacité de modifier la géométrie de son larynx, ses cavités buccale et nasale et ses sinus para-nasaux (cavités emplies d'air situées autour du nez) afin de moduler

les sons produits au niveau de la glotte. La modification de la géométrie des résonateurs s'accompagne d'une modification des fréquences de résonance. La valeur de ces fréquences de résonance, appelées formants, dépend de la voyelle prononcée. Ainsi, la production de voyelles nasales est contrôlée par l'ouverture et la fermeture du voile du palais, ou velum. Par ailleurs, le spectre du conduit vocal comprend une antirésonance due à l'ouverture des fosses nasales et des sinus para-nasaux [12]. La langue, quant à elle, est très mobile et permet de faire varier la forme de la cavité buccale et influe donc elle aussi sur la nature de la voyelle prononcée. La position des lèvres ainsi que le degré d'ouverture de la mâchoire ont également une influence acoustique sur le son produit.

1.1.2.1 Production des voyelles

Parmi les phonèmes, les voyelles sont caractérisées par la vibration des plis vocaux ainsi que la stabilité de la géométrie des articulateurs au cours de la production du phonème. L'ouverture de la mâchoire, la position de la langue sur l'axe antéro-postérieur, l'utilisation de nasalité et la forme des lèvres permettent de décrire les voyelles du français. Ainsi, l'ouverture de la mâchoire distingue les voyelles ouvertes (comme la voyelle /a/[1]) des voyelles mi-ouvertes (comme la voyelle /O/) ou des voyelles fermées (comme la voyelle /i/). La position de la langue sur l'axe antéro-postérieur permet de distinguer les voyelles antérieures, comme la voyelle /e/ des voyelles centrales comme la voyelle /@/ et des voyelles postérieures, par exemple la voyelle /o/. On nomme voyelles nasales les voyelles dont la production utilise le conduit nasal et voyelles orales en l'absence de nasalité. La forme des lèvres distingue les voyelles arrondies des voyelles non arrondies. Les voyelles orales dépendent de l'ouverture de la mâchoire et de la position de la langue, qui influent respectivement sur le premier et le second formant. C'est ce qui explique la représentation des voyelles couramment utilisée qui est le triangle vocalique, présenté Figure 5. Dans cette représentation, les voyelles sont disposées spatialement en fonction de la valeur des deux premiers formants.

[1] Nous utilisons ici le jeu de caractères phonétiques SAMPA (Speech Assessment Methods Phonetic Alphabet) fondé sur l'alphabet phonétique international mais n'utilisant que des caractères ASCII.

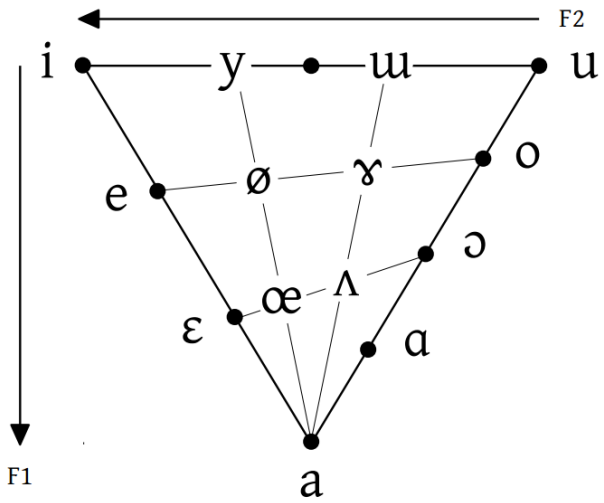


Figure 5 - Représentation des voyelles en fonction des fréquences des deux premiers formants. Cette représentation, de forme triangulaire, est nommée triangle vocalique. D'après [13].

1.1.2.2 Production des consonnes

L'articulation influence également la production des consonnes. Un son est dit voisé si les plis vocaux entrent en vibration. Lorsqu'il y a obstruction totale ou partielle du flux d'air par les articulateurs, il y a production d'une consonne. Il existe des consonnes voisées, comme les consonnes /b/ ou /m/, et des consonnes non voisées comme les consonnes /t/ et /k/. Selon si les plis vocaux sont en adduction (plis vocaux rapprochés permettant la vibration de la muqueuse) ou en abduction (plis vocaux écartés), il y aura production respectivement d'une consonne voisée ou non-voisée. En phonétique, les consonnes sont regroupées selon la localisation de la constriction principale, le mode d'articulation et la présence ou non de voisement. Les obstructions peuvent avoir lieu entre deux articulateurs ou bien entre une partie du palais et une partie de la langue. L'IPA (*International Phonetic Association*, voir [14]) distingue ainsi les consonnes labiales (/b/), dentales (/t/), alvéolaires (/n/), palatales (/J/), vélaires (/g/), uvulaires (/R/), pharyngales (comme l'aspiration /X\/ de l'arabe) et glottales (comme le /h/ de l'anglais anglais), selon le lieu d'articulation (constriction ou occlusion). Le type d'obstruction permet de définir les modes d'articulation. L'IPA distingue les consonnes occlusives (/p/), fricatives (/f/), spirantes (/j/), nasales (/n/). Ainsi, une obstruction totale donne une consonne occlusive, une obstruction partielle mais donnant lieu à une composante apériodique forte produit une consonne fricative alors qu'une obstruction faible donne lieu à une consonne sonnante. On distingue également le mode d'articulation oral (la bouche sert de cavité de résonance) du mode nasal (le nez sert de cavité de résonance) et le mode sourd (les plis vocaux n'entrent pas en vibration) du mode voisé, au cours duquel les plis vocaux entrent en vibration.

1.1.2.3 Articulation et coarticulation

L'articulation et la coarticulation jouent un rôle central dans la production d'une consonne. L'articulation d'une consonne peut être décomposée en trois phases [15] : la phase de déclenchement, la phase médiane et la phase de fin. Au cours de la phase de déclenchement, les articulateurs se déplacent vers la position de constriction dominante. Ce maximum de constriction est atteint au cours de la phase médiane. La phase de fin correspond à la phase au cours de laquelle les articulateurs s'éloignent les uns des autres. Suivant le type de consonne, la phase médiane correspondra à un arrêt du flux d'air (occlusive), un écoulement turbulent d'air (fricative) ou un écoulement du flux d'air plus libre (approximante). Selon la contrainte articulatoire, c'est-à-dire la durée minimale requise pour passer d'une configuration à une autre, et les contraintes expressives, la durée de chacune de ces phases est variable. Par ailleurs, il peut arriver qu'un segment de parole influence les segments suivants ou précédents. C'est ce que l'on appelle la coarticulation. Dans ce cas, une configuration articulatoire peut entraîner la modification des articulations pour les phonèmes adjacents.

1.1.3 Contrôle de la qualité vocale

1.1.3.1 Notion de qualité vocale

. Le contrôle du geste vocal implique le contrôle de l'articulation d'une part, et le contrôle de la qualité vocale d'autre part. Différents types de qualités vocales ont été définis dans [16] en tenant compte de la configuration du larynx et de la glotte dans les différents types de phonation. La voix chuchotée par exemple est produite par une fermeture incomplète des plis vocaux, ce qui permet un phénomène de turbulences au niveau du larynx sans vibration des plis vocaux. Ces turbulences peuvent être modélisées par des sources acoustiques à spectre large, qui sont, de même que la source produite par vibration des plis vocaux, modifiées par le filtre vocal. La fréquence de vibration des plis vocaux définit la fréquence fondamentale du son. Mais il existe d'autres paramètres laryngés qui peuvent être contrôlés lors de la phonation. L'étude de la qualité vocale fait le lien entre les caractéristiques physiologiques de la phonation et la qualité du son perçu. Les différentes qualités vocales communément admises sont la voix soufflée (*breathy voice*), la voix tendue (*tense voice*), la voix rauque (*creaky voice*) et la voix correspondant à une phonation normale (*modal voice*). La voix soufflée correspond à une fermeture incomplète de la glotte. Une voix rauque se traduit également par des fermetures glottiques incomplètes, tandis qu'une voix tendue correspondra

à une fermeture abrupte des plis vocaux. À cela vient s'ajouter la notion d'effort vocal, impliquée par exemple dans la voix criée. Des paramètres du signal de source, définis section 3.3, permettent de décrire de façon quantitative la qualité vocale et d'expliquer son lien avec les cycles d'ouverture et de fermeture glottique. Ces critères de qualité vocale s'appliquent aussi bien dans le domaine de la voix parlée que dans le domaine de la voix chantée. A la différence de la voix parlée, la voix chantée évolue dans une plage de fréquences plutôt large, ce qui nécessite une adaptation de la configuration du larynx.

1.1.3.2 Voix chantée et mécanismes laryngés

Un autre facteur de production vocale, très important dans le domaine de la voix chantée, est la notion de mécanisme laryngé (ou registre de voix). Dans [17], l'auteur rapporte une distinction entre les mécanismes laryngés suivants :

Le mécanisme 0 (*fry voice*) est employé dans la production de sons plutôt graves, il correspond à des plis vocaux courts, épais et peu tendus. La durée d'ouverture est faible en comparaison avec la durée d'une période de vibration.

Le mécanisme 1 correspond à des plis vocaux épais, qui vibrent sur toute leur longueur, avec une vibration très importante. Le rapport entre la durée d'ouverture glottique et la période est supérieur à celui du mécanisme 0 mais reste toujours inférieur à 0,5. Ce mécanisme est le mécanisme le plus utilisé en voix parlée pour les hommes, ainsi qu'en voix chantée pour les chanteurs basses, barytons, ténors et alti ainsi que des chanteurs de variété.

Le mécanisme 2 est caractérisé par une vibration des plis vocaux sur les deux tiers de leur longueur uniquement, car les cartilages aryténoïdes sont davantage comprimés. Le rapport entre ouverture glottique et période est plus élevé que dans le mécanisme 1, en général supérieur à 0,5. Ce mécanisme est utilisé par les femmes en voix parlée, les hommes lorsqu'ils souhaitent émettre un son aigu. Les chanteurs mezzo-soprano, soprano, altos et haute-contre utilisent presque exclusivement ce mécanisme. Le mécanisme 2 est plus utilisé en voix parlée par les hommes dans d'autres cultures (Asie, Afrique).

Le mécanisme 3 correspond à une voix dite de sifflet. Les plis vocaux sont très fins, allongés et tendus. L'amplitude de leur vibration est très faible. La durée de fermeture complète est presque nulle.

Le chant et la parole se distinguent par plusieurs aspects du contrôle des mécanismes laryngés. En effet, la fréquence fondamentale du son produit en voix chantée est contrôlé (hauteur de la mélodie, voir la comparaison entre les hauteurs de la voix parlée et de la voix chantée présentée Figure 6) et reste plus stable qu'en voie parlée où la fréquence fondamentale varie généralement plus rapidement. Par ailleurs, l'amplitude des variations de fréquence fondamentale (ambitus) est nettement plus importante en chant qu'en parole. Les variations d'intensité sont également plus marquées. Une autre distinction entre parole et chant est le rapport entre la durée totale des parties voisées, c'est-à-dire qui impliquent une vibration des plis vocaux, et la durée des parties non voisées. Ce rapport est nettement plus important en chant : les respirations sont modifiées par le chant, qui abrège l'inspiration et rallonge l'expiration.

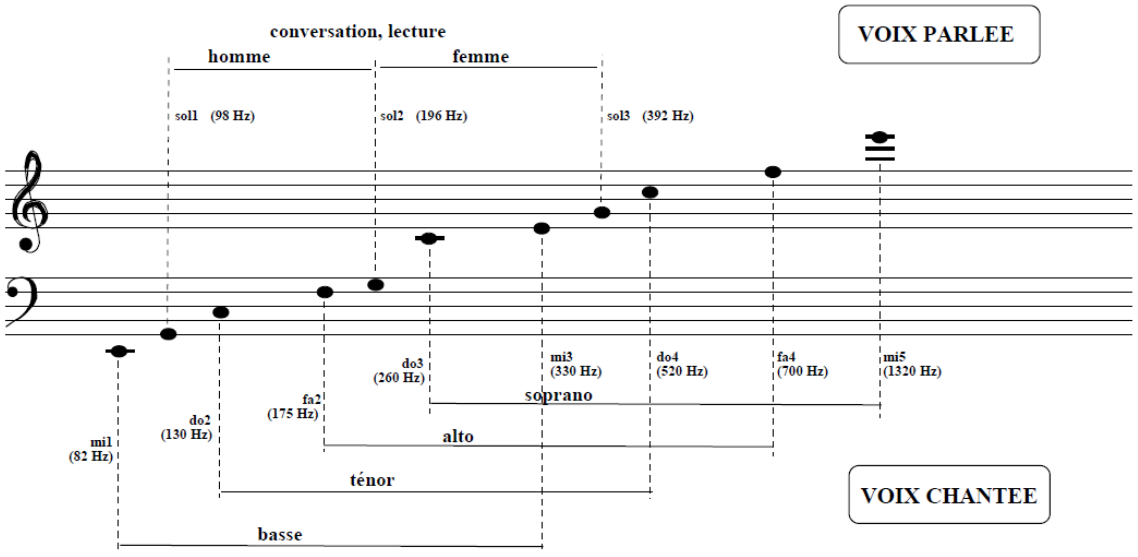


Figure 6 - Étendue vocale moyenne de la voix parlée et de la voix chantée pour les hommes et les femmes. D'après [18].

La Figure 7 illustre les différences du point de vue fréquentiel entre les différents mécanismes laryngés utilisés au cours d'un glissando[2]. Nous détaillerons en section 1.3 différents modèles permettant de représenter et analyser le signal acoustique.

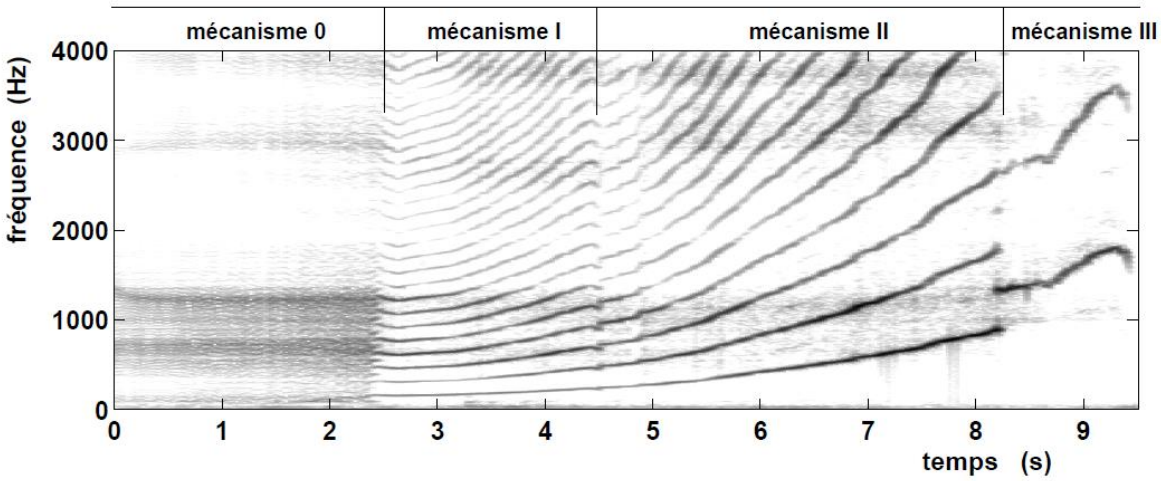


Figure 7 - Spectrogrammes (voir section 1.3.1.1) de glissandos effectués par une chanteuse (soprano légère) couvrant l'ensemble des mécanismes laryngés. D'après [17].

Chaque technique de chant possède ses spécificités en termes de rythmes mais également en termes d'articulation et de qualité vocale. Nous présentons en section 1.2 les différentes techniques de chant concernées par le projet.

1.2 Les techniques de chant rares

Dans le cadre de la sauvegarde du patrimoine culturel immatériel présenté dans le projet i-Treasures [19], plusieurs techniques de chant rares sont étudiées : le chant traditionnel corse, le chant traditionnel sarde, la musique byzantine et le *Human Beat Box*. Ces techniques de chant impliquent toutes des mouvements des articulateurs différents de ceux employés dans les techniques de chant les plus couramment étudiées comme le chant lyrique ou la variété. La section suivante propose donc un bref aperçu de ces techniques méconnues.

[2] En musique, un glissando désigne le passage d'une note à une autre, en général assez éloignées, de manière la plus continue possible, en faisant entendre rapidement les sons compris entre ces deux notes.

1.2.1 Le *Cantu in paghjella* (chant traditionnel de Corse)

Le *Cantu in paghjella* est une technique de chant polyphonique (voir Figure 8) comprenant trois voix d'hommes a capella. La voix principale, qui chante la mélodie, est appelée *a seconda*, la voix grave est nommée *bassu* et la voix aigüe est nommée *a terza*. Ce type de chant utilise des ornementations. Traditionnellement, la transmission des techniques de chant se fait de façon orale. Le répertoire du *Cantu in paghjella* comprend aussi bien de la musique profane que de la musique sacrée, mais le chant corse s'inspire traditionnellement de messes et psalmodies. Les textes de leurs chants sont soit en latin, soit en corse. Comme ces chanteurs n'utilisent ni partitions ni références de hauteur comme on peut le trouver en musique classique, les chanteurs utilisent principalement leurs yeux, leurs oreilles et leurs bouches pour communiquer entre eux. Le respect de la musicalité requiert ainsi une grande complicité et une forte interaction entre les chanteurs dont les interprétations s'influencent les unes les autres [20].

Figure 8 - Groupe de chanteurs de Cantu in paghjella*, d'après [19].*

1.2.2 Le *Canto a Tenore* (chant traditionnel sarde)

Le *Canto a tenore* de Sardaigne est, de même que le chant corse, un style de chant polyphonique composé de voix d'hommes uniquement (voir Figure 9), mais dont la tessiture est plus basse que dans le chant corse. La qualité vocale est également différente. Le *Canto a tenore* regroupe quatre voix d'hommes. Deux d'entre elles utilisent une phonation normale tandis que les deux autres utilisent davantage le larynx. La voix soliste est appelée *oche* ou

boche et utilise une phonation normale. L'autre voix utilisant ce mécanisme est appelée *mesu oche* ou *mesu boche*, ce qui signifie « demi-voix ». La voix grave est appelée *bassu* et l'autre voix utilisant le même mécanisme, chantant une quinte au-dessus du *bassu*, est appelée *contra*.

Figure 9 - Groupe de chanteurs de Canto a Tenore*, d'après [19].*

La technique de *Bassu* et de *Contra* nécessite une interaction entre les cordes vocales et les bandes ventriculaires [21], dont l'anatomie est détaillée Figure 10. Les bandes ventriculaires ne sont pas couramment utilisées dans le cadre d'une phonation normale. Cependant, leur utilisation a été observée dans certains chants gutturaux comme dans certaines cultures asiatiques. Cette technique de chant est associée avec un phénomène de doublement de période qui est à l'origine de la voix grave perçue [22]. De plus, un quatuor de chanteurs masculins produit perceptivement une cinquième voix. La hauteur résultante ressemble à une voix de femme [23].

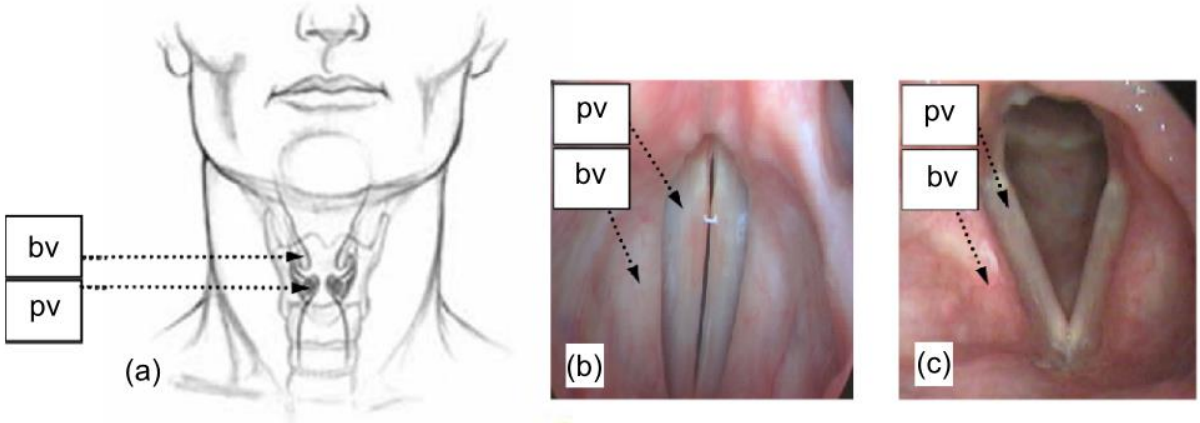


Figure 10 - Les positions des plis vocaux, désignés par les initiales « pv » et des bandes ventriculaires, désignées par les initiales « bv ». (a). Vue frontale du larynx. (b). Image du larynx au cours de la phonation obtenue par laryngoscopie. (c). Illustration du larynx au cours de la respiration obtenue par laryngoscopie, d'après [21].

1.2.3 La musique byzantine

La musique byzantine consiste en un chant choral monophonique et a cappella (voir Figure 11). Elle est parfois accompagnée par un son continu mais aucun instrument de musique n'est toléré dans les églises byzantines. La construction des huit modes (système musical permettant, au même titre que les gammes, l'organisation des hauteurs d'une échelle autour d'une note) utilisés chant byzantin repose sur une symbolique spirituelle. Les chanteurs byzantins utilisent une échelle musicale particulière mais prennent parfois la liberté d'improviser et de chanter des mélodies avec des ornementations. Le chanteur qui chante la mélodie conduit le chœur davantage avec sa voix qu'avec des gestes et les autres chanteurs imitent sa façon de chanter. Dans la musique byzantine, le texte est considéré comme moins important que la musique car la musique a pour vocation d'aider à atteindre un état spirituel. Dans les villages, la transmission des techniques était garantie par le chantre, qui avait également pour vocation d'enseigner une philosophie de vie. De nos jours, les apprentis ont tendance à utiliser des CD et internet pour apprendre et ils perdent ainsi le contact avec le chantre.

Figure 11 - Un ensemble vocal de musique byzantine, d'après [19].

1.2.4 *Human Beat Box*

Le *Human Beat Box* (HBB) est une technique de chant qui consiste à reproduire différents sons (rythmiques, instrumentaux, vocaux) avec la bouche [24]. Dans la musique contemporaine, le Beat Box, inspiré de la culture hip-hop, est pratiqué aussi bien dans le cadre de l'accompagnement de chant ou de rap que seul comme expression artistique à part entière. Les beat boxers adoptent des attitudes laryngo-pharyngées et posturales très complexes. Ils

pratiquent leur technique en utilisant des configurations articulatoires extrêmes et investissent tout leur corps dans le but d'imiter très précisément la géométrie de l'instrument qu'ils cherchent à reproduire [25]. Les beat boxers sont capables de produire une très grande variété de sons qui surpasse les combinaisons articulatoires de la plupart des langues. Cependant, l'ensemble de ces sons peuvent être décrits en utilisant l'Alphabet Phonétique International, alphabet dédié à la description des sons de parole. Ainsi, même si l'objectif du *Human Beat Box* est de produire des sons extralinguistiques, les Beat-boxers utilisent des combinaisons articulatoires qui existent parmi les langages humains [25].

Ces types de chant mettent en œuvre des techniques variées et complexes qui nécessitent une description multimodale. Nous détaillerons des méthodes d'analyse et synthèse adaptées au signal vocal en section 1.3, puis nous proposerons des méthodes permettant de traiter d'autres informations articulatoires en section 1.4.

1.3 Modèles pour l'analyse et la synthèse de la voix

1.3.1 Introduction à l'analyse et la synthèse vocale

Modéliser le conduit vocal permet de mieux comprendre ses propriétés, mais aussi d'en imiter le fonctionnement pour des applications comme la synthèse vocale. Dans le présent travail, nous avons besoin d'un modèle du conduit vocal à la fois réaliste et facile à implémenter pour modéliser les mécanismes de la voix chantée. Deux principaux modèles sont utilisés pour décrire la production de la voix humaine : les modèles issus de la théorie du signal et les modèles physiques. Les modèles des signaux utilisent souvent des méthodes d'inversion (estimation des paramètres du modèle à partir d'un son de référence) permettant des implémentations efficaces, mais dont le réalisme est limité. En particulier, si l'on connaît la fonction de transfert du filtre du conduit vocal, il est possible de construire l'inverse de ce filtre et ainsi de retrouver l'excitation glottique à l'origine de ce son. Les modèles physiques sont relativement faciles à interpréter mais ces modèles sont difficiles à inverser. La synthèse vocale, reposant sur l'un ou l'autre de ces modèles, peut être divisée en cinq grandes familles : la synthèse par concaténation, la synthèse additive, les modèles source-filtre, les modèles physiques simples et les modèles physiques complexes. La synthèse concaténative est fondée sur la concaténation d'unités préenregistrées de taille variable. Cette technique est à l'origine de nombreuses applications, en synthèse comme en modification de la parole,

comme par exemple la méthode PSOLA (*Pitch Synchronous Overlap and Add*) et ses variantes [26]. La synthèse additive a pour but de reconstruire le spectre d'un signal de parole en utilisant des informations sinusoïdales ou formantiques [27]. Les modèles source-filtre, que nous présenterons dans la section 1.3.1.3, reposent sur la modélisation des caractéristiques à la fois temporelles et fréquentielles du signal vocal. Les modèles source-filtre sont centrés sur le résultat de la phonation (le son produit), tandis que la particularité des modèles physiques est qu'ils sont centrés sur les causes de la phonation (mouvements laryngés et articulation). Les modèles physiques, qui sont détaillés en section 1.3.1.2, sont fondés sur l'activité du larynx, la configuration du conduit vocal et leur influence sur la production du son. Les modèles de traitement de la parole supposent qu'un segment de parole est stationnaire sur une fenêtre d'environ 15 à 20 ms, ce qui revient à considérer les mouvements des articulateurs négligeables au cours de cette durée. Ces hypothèses permettent des analyses spectrales trame (portion de signal d'une durée de l'ordre de la dizaine de millisecondes) par trame.

1.3.1.1 Représentations fréquentielles du signal de parole

La quasi-périodicité du signal acoustique justifie l'intérêt d'une analyse fréquentielle du signal de parole. L'analyse spectrale d'un signal de parole permet en effet d'identifier les différentes fréquences qui composent ce signal. Le suivi de la fréquence fondamentale au cours du temps (voir Figure 12) est une autre représentation d'une partie du contenu fréquentiel du signal de parole et permet d'étudier la prosodie. Il existe de nombreuses méthodes permettant le suivi de cette fréquence fondamentale, la plus intuitive étant la méthode dite de *zero-crossing* [28]. Il s'agit de repérer les passages par la valeur 0 du signal afin de repérer les périodicités du signal. Cependant, cette méthode n'est pas très efficace en présence de bruit ou de sources multiples [29] et nécessite des ajustements [30]. D'autres méthodes plus classiques utilisent l'autocorrélation du signal de parole [31] ou bien l'AMFD (*Average magnitude difference function pitch extractor*), décrit dans [32]. Il est également possible d'utiliser le signal électroglottographique afin de déterminer la fréquence fondamentale grâce à l'identification des instants d'ouverture et de fermeture glottique [33]. En dehors de la fréquence fondamentale du signal, un spectre (voir Figure 13) permet de déterminer harmoniques et formants [34].

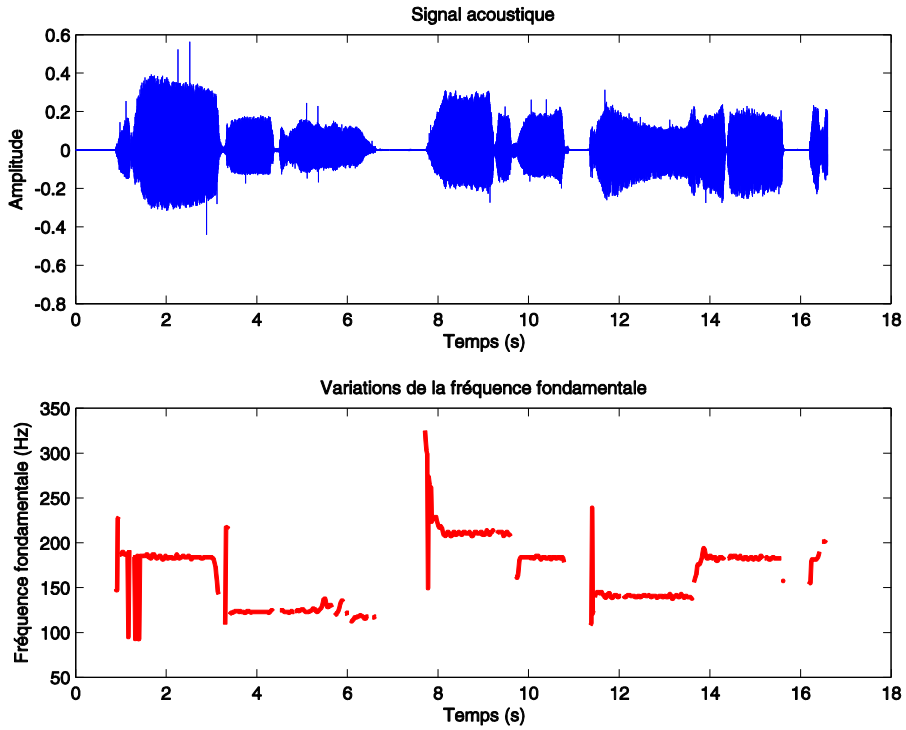


Figure 12 - Signal temporel et variations de la fréquence fondamentale au cours du temps. En haut, un exemple de signal temporel complet. En bas, les variations de la fréquence fondamentale de ce signal au cours du temps. Les silences ont été exclus du calcul.

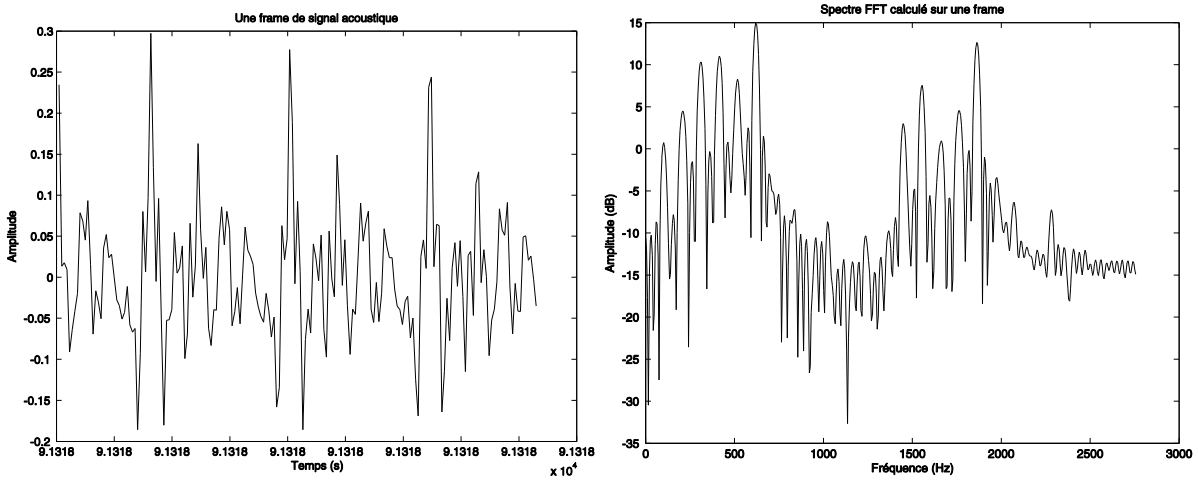


Figure 13 - Une trame de signal et le spectre calculé sur cette trame.

L'utilisation d'un spectrogramme permet de visualiser des variations de l'énergie du signal dans les différentes bandes de fréquence. Le spectrogramme est un diagramme permettant de représenter trois dimensions d'un signal acoustique que sont le temps, généralement en abscisse, la fréquence, en ordonnées et la puissance, représentée par la luminosité du point. Pour obtenir un spectrogramme, il faut découper le signal en fenêtres et calculer le spectre de chacune de ces fenêtres. Cette représentation est particulièrement utile pour observer des transitions (fricatives, voyelles, consonnes). Les spectrogrammes sont des outils très performants pour l'annotation phonétique et la segmentation phonémique. La détermination

automatique de la position des formants, en surimpression sur le spectrogramme (voir Figure 14), permet d'apporter des informations utiles pour l'analyse de la parole, en particulier pour les sons voisés.

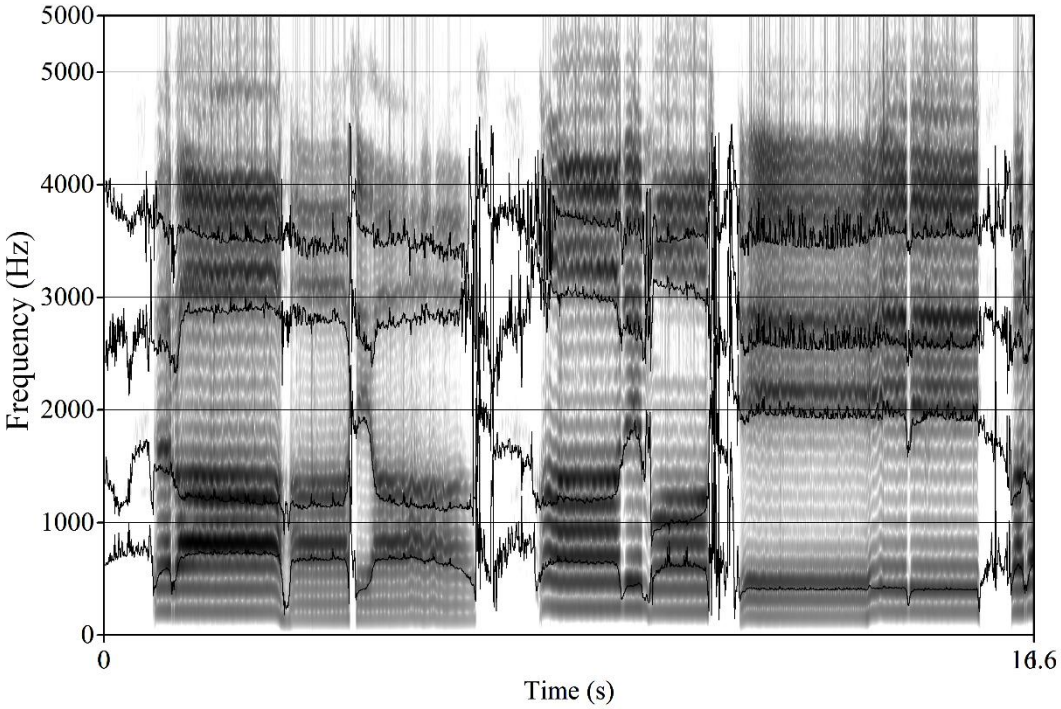


Figure 14 - Spectrogramme d'une portion de signal faisant apparaître les quatre premiers formants.

Nous souhaitons développer des modèles d'analyse du geste vocal et privilégions donc des descripteurs qui pourraient être extraits à partir d'informations visuelles. Il existe des modèles permettant de décrire le mécanisme de production de la parole en tenant compte de l'influence des différentes parties du conduit vocal.

1.3.1.2 Modèles physiques

Les modèles physiques, largement décrits dans [35], permettent une interprétation la plus facile de la production de parole, en traduisant l'interaction entre les différentes parties du conduit vocal. Parmi ces méthodes, on trouve les modèles mécaniques auto-oscillants et les modèles à géométrie forcée. Le modèle à deux masses est un exemple de modèle mécanique auto-oscillant dans lequel les plis vocaux sont représentés comme des masses reliées entre elles par un ressort et au reste du corps par l'association d'un ressort et d'un amortisseur. L'ensemble du système est décrit par des interactions mécaniques, aérodynamiques et acoustiques. Dans les modèles à géométrie forcée, la géométrie du conduit laryngée, la géométrie du conduit vocal ainsi que la pression sous-glottique sont des valeurs paramétrées

par le modèle. Ainsi, les grandeurs acoustiques comme la pression et le débit au niveau de la glotte, dans le conduit vocal ainsi qu'au niveau des lèvres et des narines sont calculées en utilisant des équations acoustiques (modèles de conduits) à partir des données géométriques du conduit vocal, qui peuvent varier au cours du temps. Cependant, il est délicat de modéliser comment les mouvements des articulateurs influencent les interactions entre les différentes parties du conduit vocal.

1.3.1.3 Modèle source-filtre

Contrairement aux modèles physiques, le modèle source-filtre repose sur le postulat d'une séparation entre la source idéale du son et l'effet des résonateurs et articulateurs qui « filtrent » cette source [36]. La production du son, modélisée par une source idéale, est considérée indépendante de sa modification au travers du conduit vocal, modélisé par un filtre linéaire auquel vient s'ajouter le rayonnement aux lèvres, modélisé par une dérivation du signal temporel. Puisque l'on suppose le système linéaire, l'ordre des filtres peut être interchangé, c'est pourquoi, dans la plupart des modèles, l'effet du rayonnement aux lèvres intervient directement sur le signal de source sous forme de dérivation (voir Figure 15). Le signal de source est donc dérivé afin de tenir compte de l'effet du rayonnement aux lèvres : l'onde de débit glottique dérivée sert de signal de source et ce signal est directement filtré par le filtre vocal.

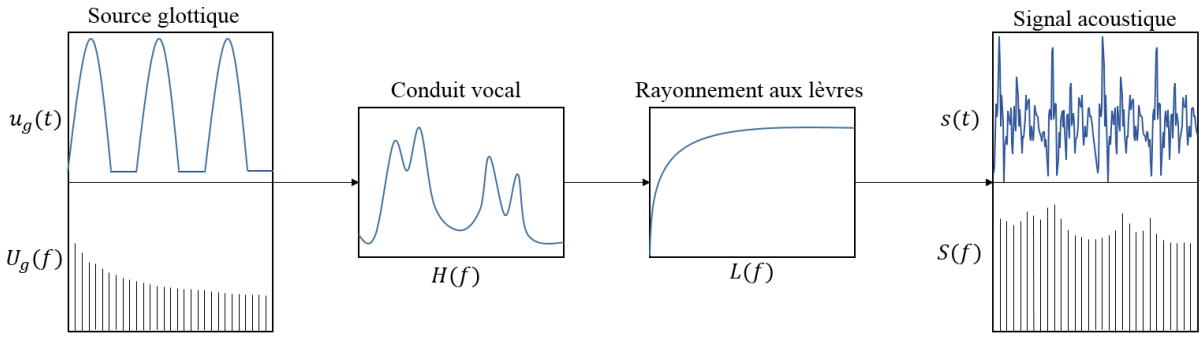


Figure 15 - Illustration du modèle source-filtre. Le produit des transformées de Fourier de la source glottique $U_g(f)$, de la transformée de Fourier du conduit vocal $H(f)$ et de la transformée de Fourier du rayonnement aux lèvres $L(f)$ donne un signal acoustique de représentation fréquentielle $S(f)$. D'après [17].

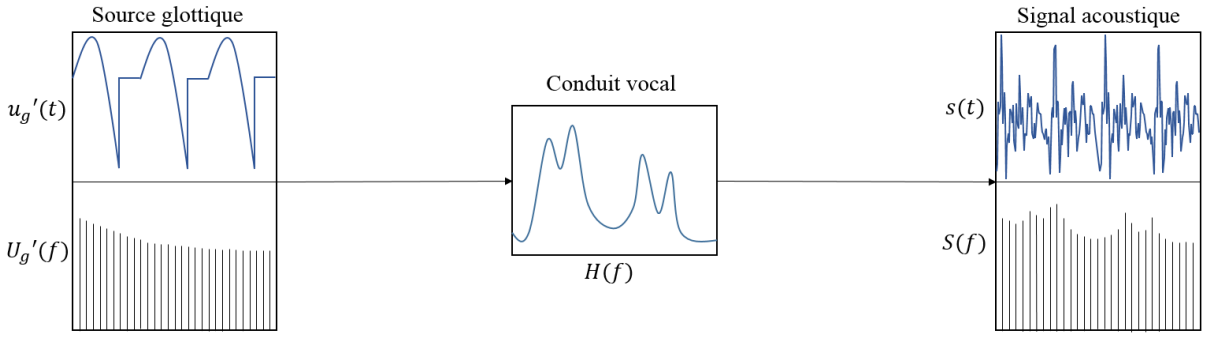


Figure 16 - Illustration du modèle source-filtre en tenant compte du rayonnement aux lèvres par dérivation du signal de source. Le produit de la transformées de Fourier de la source glottique dérivée $U_g'(f)$ par la transformée de Fourier du conduit vocal $H(f)$ produit le même signal acoustique de représentation fréquentielle $S(f)$ que celui montré. D'après [17].

1.3.1.4 Prédiction linéaire

Le codage prédictif linéaire de la parole par LPC (*Linear Predictive Coding*) est un outil permettant d'obtenir une représentation du signal de parole compatible avec un débit de transmission faible tout en limitant les dégradations du signal liées à la compression. Il s'agit d'exploiter l'hypothèse selon laquelle la parole peut être modélisée par un processus linéaire, en prédisant le signal à l'instant n à l'aide des p échantillons précédents. Les coefficients LPC sont les coefficients d'un filtre autorégressif permettant de modéliser le conduit vocal. Cependant, le processus de parole n'étant pas complètement linéaire, il est nécessaire de corriger ce modèle en introduisant un terme d'erreur de prédiction. Cette erreur est la différence entre le signal acoustique et celui prédit par codage linéaire. Les méthodes de détermination des coefficients de prédiction linéaire permettent de minimiser cette erreur de prédiction. La Figure 17 montre une trame de signal et son estimation par prédiction linéaire. La différence entre les deux constitue le signal de résidus et est présentée Figure 18.

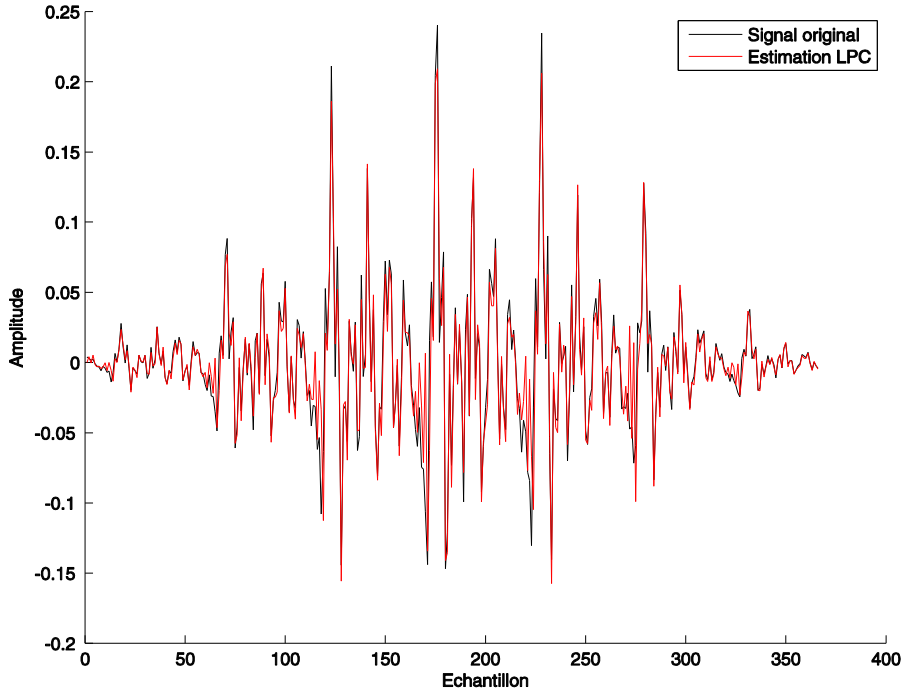


Figure 17 - Comparaison entre une trame de signal original (en noir) et l'estimation de cette trame par prédiction LPC (en rouge).

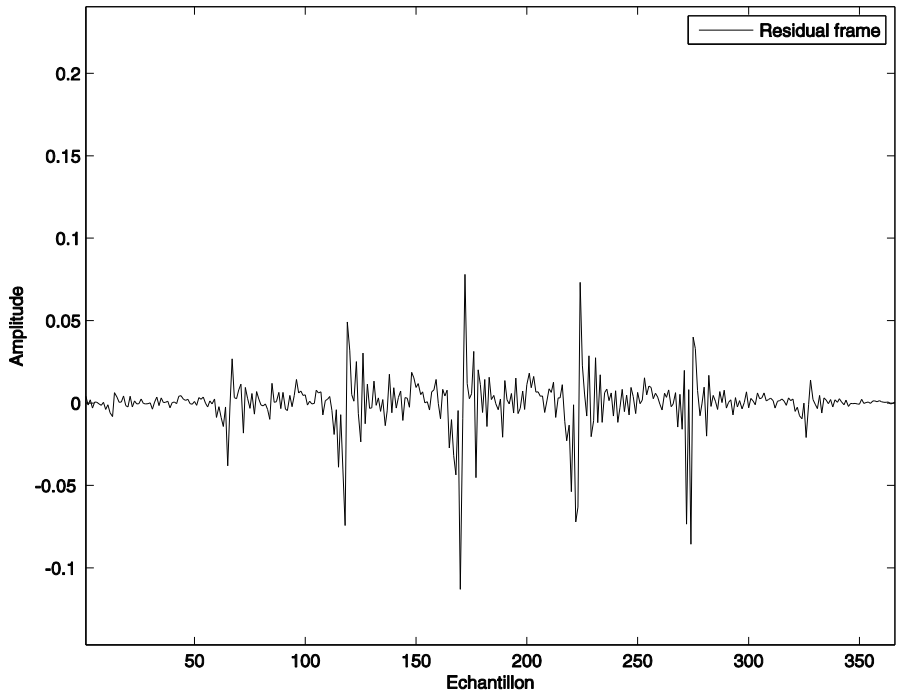


Figure 18 - Signal de résidu correspondant à la prédiction LPC montrée Figure 17.

La prédiction linéaire permet d'accéder à l'enveloppe spectrale d'une trame de parole. Cette enveloppe spectrale permet de situer les formants, dont la fréquence est donnée par les

maxima de l'enveloppe du spectre à un instant donné. La Figure 19 montre le spectre FFT d'une trame audio ainsi que son spectre obtenu par LPC.

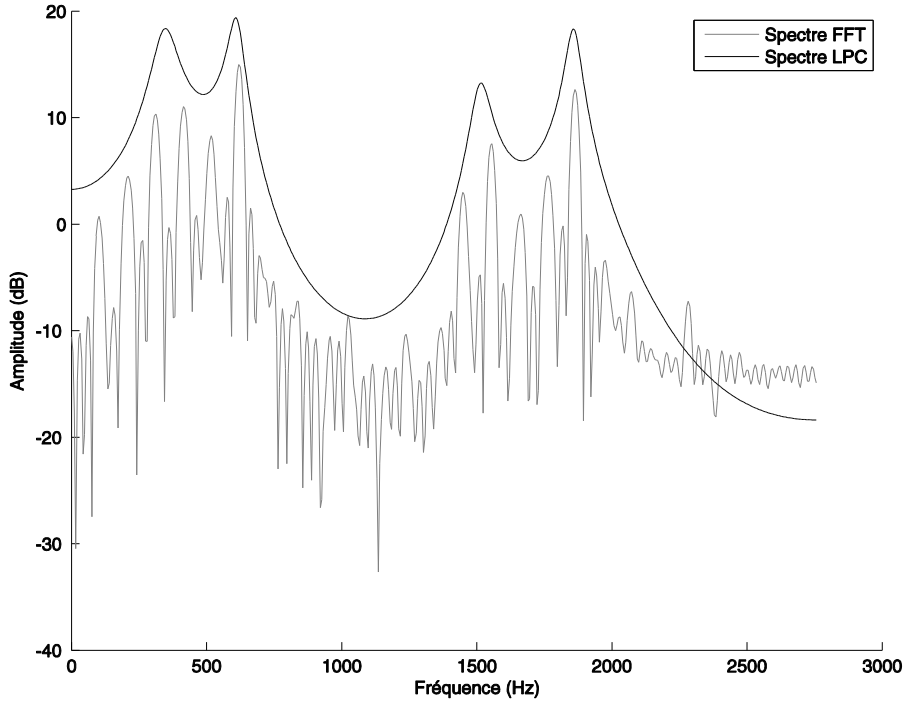


Figure 19 - Comparaison entre le spectre FFT (en gris) et le spectre LPC (en noir) d'une trame d'un extrait de chant. Tandis que le spectre FFT fait apparaître les harmoniques de la fréquence fondamentale, le spectre LPC donne accès aux valeurs des formants.

Les LSF (*Line Spectral Frequencies*, aussi connus sous le nom *Line Spectral Pairs*) sont déduits des coefficients LPC. Les LSF ont la propriété d'être plus robustes aux distorsions et en particulier aux erreurs de quantification que les LPC [37]. Soit $A(z)$ la fonction polynômiale de prédiction linéaire. $A(z)$ s'écrit sous la forme :

$$A(z) = 1 - \sum_{k=1}^{N} a_k \; z^{-k} \tag{1}$$

$A(z)$ peut également être décomposé comme suit :

$$A(z) = 0{,}5\,[P(z) + Q(z)], \tag{2}$$

où
$$P(z) = A(z) + z^{-(N+1)}A(z^{-1}) \tag{3}$$

et
$$Q(\mathrm{z}) = A(z) - z^{-(N+1)}A(z^{-1}) \tag{4}$$

Les LSF sont les arguments des racines des polynômes P et Q compris entre 0 et π. Les zéros des fonctions polynomiales $P(z)$ et $Q(z)$ associées à P et Q sont dessinés sur le cercle unité Figure 20.

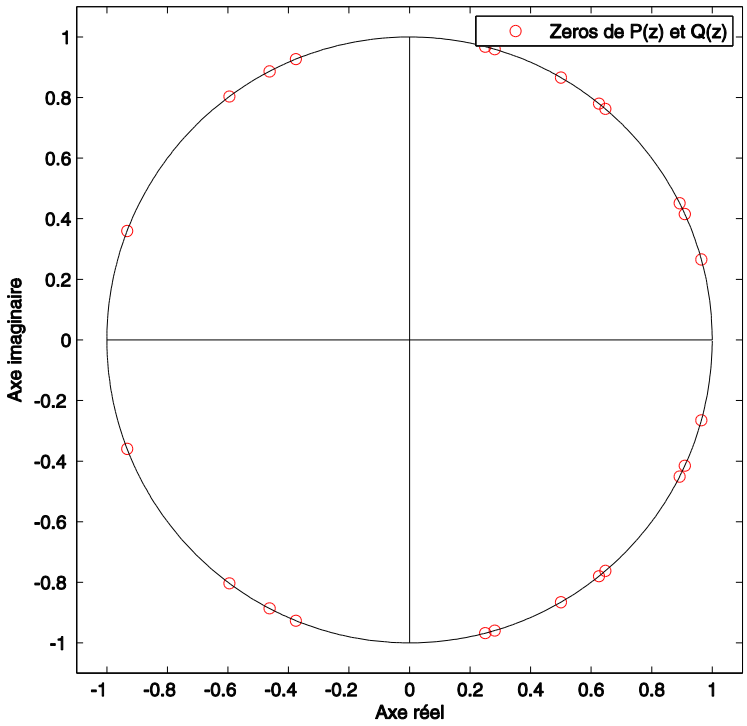


Figure 20 - Zéros des fonctions polynomiales $P(z)$ et $Q(z)$ sur le cercle unité calculées sur la portion de signal présentée Figure 17. Les LSF correspondent à l'argument des racines de P et Q. Il est à noter que puisque les coefficients de P et de Q sont réels, chaque racine est associée à une racine conjuguée et ainsi, seule la moitié des coefficients permet de coder l'information contenue dans les LPC. Par convention, les LSF correspondent aux coefficients compris entre 0 et π.

1.3.1.5 Modélisations de la source glottique

1.3.1.5.1 Onde de débit glottique

La glotte est responsable de la transformation de la pression provenant des poumons en une série d'impulsions liées aux cycles d'ouverture et de fermeture glottique. Ce signal issu de la source glottique est nommé onde de débit glottique (ODG). On peut la modéliser comme la réponse d'un filtre à un train d'impulsions. La période de l'ODG est égale à la période fondamentale du signal acoustique. Le signal permettant de tenir compte du rayonnement aux lèvres au niveau du signal de source sous forme de dérivation se nomme onde de débit glottique dérivée, ou ODGD. La Figure 21 représente une période de l'ODG et de sa dérivée. La plupart des modèles de l'ODG sont des modèles temporels mais leurs paramètres varient selon les modèles.

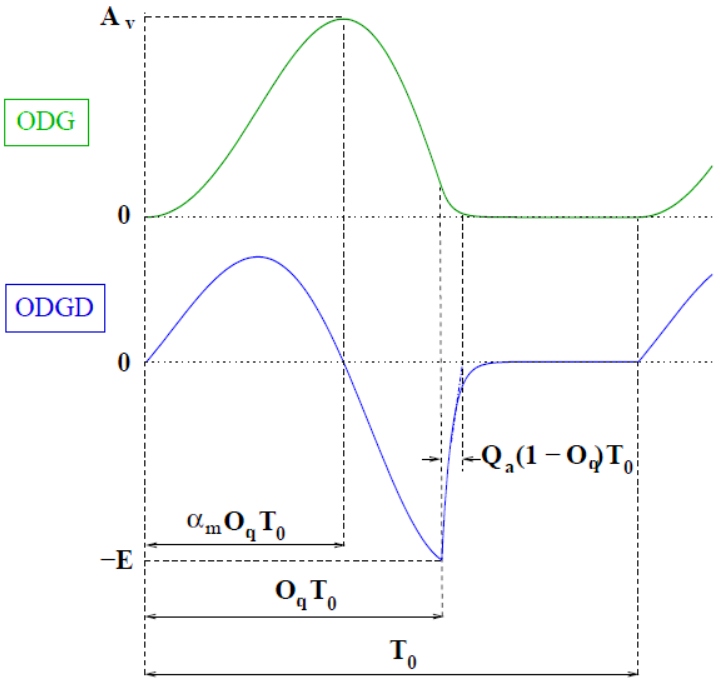


Figure 21 - Représentation d'une période de l'ODG et de sa dérivée. Le paramètre de source A_v désigne l'amplitude de voisement, T_0 la période fondamentale, O_q le quotient ouvert, α_m le coefficient d'asymétrie, Q_a le quotient de phase de retour et E la vitesse de fermeture. Leur interprétation est donnée dans le Tableau 1. D'après [17].

Les modèles d'ODG vérifient les critères suivants, décrits dans [17] : en premier lieu, l'onde de débit glottique est toujours positive ou nulle, avec une croissance pendant l'ouverture et une décroissance pendant la fermeture. De plus, l'ODG sera constante ou nulle si la glotte est fermée. En outre, la vitesse de la fermeture glottique étant généralement supérieure à la vitesse d'ouverture glottique, la forme de l'ODG est asymétrique. Les différents paramètres de source glottique que nous utilisons proviennent du modèle décrit dans [38] et sont les suivants : l'amplitude de voisement A_v, la période fondamentale de l'ODG T_0, le quotient ouvert O_q, le coefficient d'asymétrie α_m, la durée de phase retour T_a et la vitesse de fermeture E. La durée de la phase ouverte est $T_e = O_q T_0$ et la durée de la phase d'écartement des plis vocaux est $T_p = \alpha_m O_q T_0$. La durée de phase retour T_a correspond à la durée entre l'instant de fermeture glottique (noté GCI pour *Glottal Closure Instant*) et la fermeture effective. L'amplitude de voisement correspond à l'amplitude entre la valeur minimale et la valeur maximale du débit glottique. Le quotient ouvert correspond au rapport entre la durée de la phase ouverte T_e et la période fondamentale T_0. Le coefficient d'asymétrie correspond au rapport entre la durée de la phase d'écartement des plis vocaux et la durée de la phase ouverte. Le quotient de phase retour correspond au rapport entre la durée effective de retour et la durée de la phase fermée. La vitesse de fermeture décrit la vitesse du débit glottique à l'instant de

fermeture. Sa valeur est déterminée par l'amplitude du minimum de la dérivée du débit glottique. Le tableau ci-dessous rassemble ces différents paramètres ainsi que leur définition, détermination et valeurs typiques.

Tableau 1 - Les différents paramètres permettant de définir un modèle d'onde de débit glottique.

Paramètre	Nom	Interprétation	Valeurs typiques
A_v	Amplitude de voisement	Amplitude entre la valeur minimale et la valeur maximale du débit glottique	Exprimé en l/s
T_0	Période fondamentale	$T_0 = \frac{1}{f_0}$	
O_q	Quotient ouvert	Rapport entre la durée de la phase ouverte et la période fondamentale	$0{,}3 < O_q < 1$
α_m	Coefficient d'asymétrie	Rapport entre la durée de la phase d'écartement des plis vocaux et la durée de la phase ouverte $\alpha_m = \frac{T_p}{T_e}$	$0{,}5 < \alpha_m < 1$
Q_a	Coefficient de phase retour	Rapport entre la durée de phase de retour et la phase fermée $Q_a = \frac{T_a}{T_0 - T_e} = \frac{T_a}{(1 - O_q)T_0}$	
E	Vitesse de fermeture	Amplitude du minimum de la dérivée du débit glottique	Exprimé en l/s²

Ces paramètres de source peuvent avoir une influence sur la qualité vocale, décrite section 1.1.3.1. Il existe des modèles paramétriques de la source glottique, qui permettent de générer une forme d'onde. C'est le cas par exemple du modèle LF [39] (voir section 1.3.1.5.3). Ces paramètres de la source peuvent également être obtenus par filtrage inverse. Il est aussi possible de concevoir la source glottique elle aussi comme un filtre appliqué à un train d'impulsions de Dirac, comme dans le modèle CALM (*Causal-Anticausal Linear Model*, voir

section 3.5.2). Ainsi, un train d'impulsion pseudo-périodique est filtré par un filtre glottique, ce qui permet d'obtenir un modèle de l'onde de débit glottique.

1.3.1.5.2 Filtrage inverse

Afin d'obtenir l'onde de débit glottique à partir du signal acoustique rayonné, une méthode couramment utilisée est le filtrage inverse [40]. Pour ce faire, il faut d'abord estimer le filtre du conduit vocal, par exemple en utilisant les coefficients de prédiction LPC (*Linear Predictive Coding*). Une fois le filtre du conduit vocal estimé, son inverse est appliqué au signal rayonné pour obtenir l'ODG dérivée. Cependant, ce modèle d'analyse, bien que simple d'un point de vue conceptuel, ne semble pas adapté à la voix chantée. En effet, la fréquence fondamentale dans le cas du chant est assez élevée, en particulier pour les voix de femmes, ce qui nécessite d'adapter les fréquences d'analyse pour la prédiction LPC. Par ailleurs, ce modèle ne tient pas compte des interactions source-filtre, qui semblent avoir un impact dans le cadre du chant. En effet, la fréquence fondamentale n'est pas forcément décorrélée des fréquences formantiques ; des études [41] ont montré l'existence du phénomène de *formant tuning*, selon lequel le chanteur, par l'ouverture de la bouche, a tendance à rapprocher les formants des premiers harmoniques. Cet ajustement permet un gain d'énergie acoustique, notamment pour les voix aigues.

1.3.1.5.3 Modèles de l'onde de débit glottique

Des modèles de source glottique, temporels ou fréquentiels, peuvent être envisagés dans le cadre de l'analyse de la voix chantée. Ce paragraphe est inspiré de [42]. Le modèle LF, du nom de ses développeurs Liljencrants et Fant, est le modèle le plus utilisé pour l'estimation ou la modélisation du signal de source glottique. Il est contrôlé par 5 paramètres, à savoir l'amplitude de l'ODG au maximum d'excitation E_e, la période fondamentale T_0, la durée de la phase ouverte T_e, la durée d'écartement des plis vocaux T_p et la constante de temps de la phase retour notée T_a. Le modèle LF, illustré Figure 22, définit la dérivée de l'onde de débit glottique (ODGD) en ajustant deux signaux : à gauche de l'instant de fermeture glottique, l'ODGD est modélisée par une sinusoïde modulée par une exponentielle croissante et à droite de l'instant de fermeture glottique, la phase de retour est modélisée par une exponentielle décroissante.

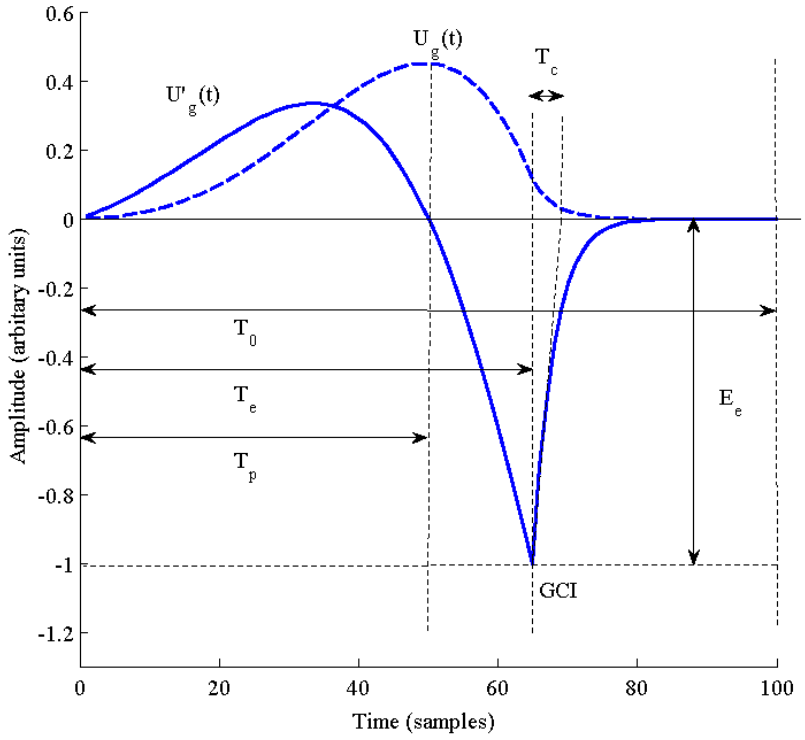


Figure 22 - Représentations de l'ODG et de sa dérivée décrites par le modèle LF et ses paramètres. D'après [42].

$$U'_g(t) = -E_e e^{a(t-T_e)} \frac{sin\left(\frac{\pi t}{T_p}\right)}{sin\left(\frac{\pi T_e}{T_p}\right)} pour\ 0 \leq t \leq T_e \tag{5}$$

$$U'_g(\mathrm{t}) = -\frac{E_e}{\varepsilon T_c}\left(e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_0-T_e)}\right) \text{ pour } T_e \leq t \leq T_0 \tag{6}$$

Les paramètres a et ε sont déterminés par la résolution de deux équations :

$$\varepsilon T_c = 1 - e^{\varepsilon(T_0-T_e)} \tag{7}$$

$$\frac{1}{a^2 + \left(\frac{\pi}{T_p}\right)^2}\left(e^{-aT_e}\left(\frac{\frac{\pi}{T_p}}{\sin\frac{\pi T_e}{T_p}}\right) + a - \frac{\pi}{T_p}\tan^{-1}\frac{\pi T_e}{T_p}\right) = \frac{T_0 - T_e}{e^{\varepsilon(T_0-T_e)} - 1} - \frac{1}{\varepsilon} \tag{8}$$

Ces équations sont obtenues par continuité de l'onde de débit glottique dérivée au point de fermeture glottique en utilisant la nullité de l'intégrale de l'onde de débit glottique sur un cycle.

En dehors du modèle LF, d'autres modèles ayant chacun leur ensemble de paramètres ont été proposés. Parmi ces modèles, les plus répandus sont le modèle KLGLOTT88 [43], R++ [44]

ou Rosenberg-B [45]. Une comparaison temporelle entre ces modèles est proposée Figure 23. Ces différents modèles ont un nombre variable de paramètres de forme, allant de 4 paramètres pour le modèle LF à 2 paramètres pour le modèles KLGLOTT88 et Rosenberg-B, en passant par 3 paramètres pour le modèle R++. Le modèle LF ainsi que sa réduction LF-Rd à un modèle à un paramètre sont encore très largement utilisés. Le succès du modèle LF s'explique par la grande diversité des formes d'onde qu'il permet d'obtenir grâce à son grand nombre de degrés de liberté. D'autres méthodes permettent de concevoir l'onde de débit glottique comme la réponse impulsionnelle d'un filtre linéaire.

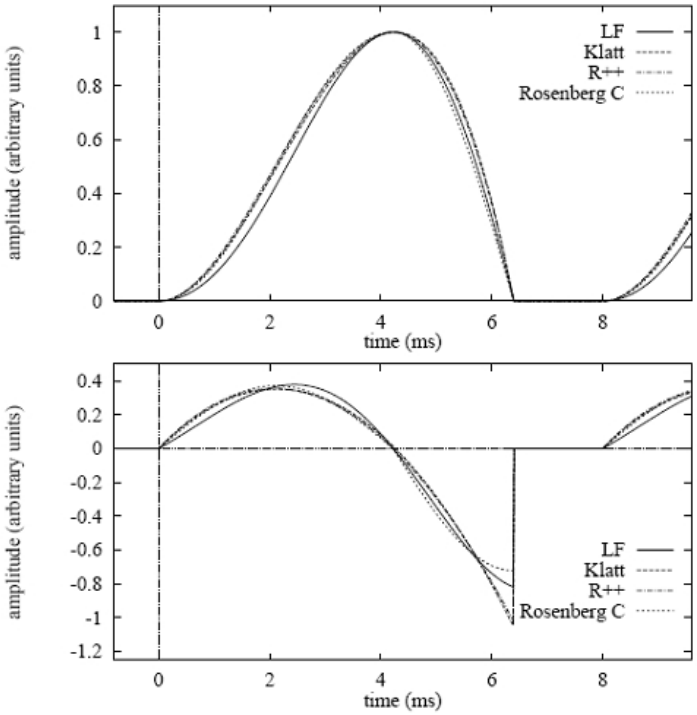


Figure 23 - Comparaison entre les modèles LF, Klatt, R++ et Rosenberg C, d'après [42].

1.3.1.5.4 Modèle CALM

Le modèle CALM (*Causal-Anticausal Linear Model)*, introduit dans [46], propose une synthèse entre les modèles temporels et les modèles fréquentiels de la source glottique. La source glottique est modélisée par la réponse impulsionnelle de deux filtres en cascade, un filtre anticausal et un filtre causal [46]. Le premier filtre, anticausal, est un filtre passe-bas du second ordre qui modélise la phase d'ouverture dans l'onde de débit glottique dérivée. Le second filtre est un filtre passe-bas du premier ordre, causal, modélisant la phase de retour de l'onde de débit glottique dérivée. Le deuxième filtre permet d'agir sur la pente spectrale Tl du modèle, qui est le paramètre fréquentiel lié à la vitesse de fermeture glottique.

Soit H_1 le filtre anticausal et H_2 le filtre causal. Le filtre passe-bas d'ordre 2 H_1 est défini comme suit :

$$H_1(\mathrm{z}) = \frac{b_{1z}}{1 + a_1 z + a_2 z^2} \tag{9}$$

avec

$$a_1 = -2e^{-a_p T_e} \cos\left(b_p T_e\right), \tag{10}$$

$$a_2 = e^{-2a_p T_{e,}} \tag{11}$$

$$a_p = \frac{\pi}{O_q T_0 \tan(\pi \alpha_m)}, \tag{12}$$

$$b_1 = E T_{e,} \tag{13}$$

$$b_p = \frac{\pi}{O_q T_0}, \tag{14}$$

Le filtre du premier ordre H_2 est défini comme :

$$H_2 = \frac{b_{Tl}}{1 - a_{Tl} z^{-1}} \tag{15}$$

avec :

$$a_{Tl} = \nu - \sqrt{\nu^2 - 1}, \tag{16}$$

$$b_{Tl} = 1 - a_{Tl}, \tag{17}$$

$$\text{avec } \nu = 1 - \frac{1}{\mu}, \tag{18}$$

$$\text{et } \mu = \frac{\frac{1}{e^{-Tl/10 \ln 10}} - 1}{\cos\left(2\pi \frac{3000}{f_e}\right) - 1}. \tag{19}$$

Ce modèle permet de faire la synthèse entre des modèles temporels et des modèles fréquentiels, par l'intermédiaire d'un modèle alliant des filtres causaux et anti-causaux. L'association de ce modèle de source glottique avec un modèle du conduit vocal permet de développer des méthodes de synthèse vocale. Différentes méthodes de synthèse vocale, fondée ou non sur le modèle source-filtre, sont détaillées en section 1.3.2.

1.3.2 La synthèse vocale

1.3.2.1 Généralités

La synthèse vocale permet de produire, à partir de moyens électro-acoustiques, des sons imitant la voix parlée ou chantée. On distingue deux catégories de méthodes de synthèse vocale : les méthodes utilisant des connaissances explicites et les méthodes par connaissances implicites [47]. Les méthodes par connaissances explicites, que nous décrivons en section 1.3.2, utilisent des modèles de l'appareil vocal (modèles physiques ou signaux). Les méthodes par connaissances implicites utilisent des segments de voix préenregistrés et n'utilisent pas d'*a priori* sur le conduit vocal. Il s'agit de méthodes de synthèse par concaténation, que nous abordons en section 1.3.2.3. Il existe également des méthodes intermédiaires, utilisant des bases de données de voix naturelle. La reconstruction du signal est dans ce cas pilotée en utilisant des paramètres issus de ces bases. Cette méthode est utilisée dans la synthèse HTS (HMM-To-Speech, voir section 1.3.2.4). La demande en ce qui concerne les applications commerciales a motivé de nettes améliorations en termes de qualité (naturalité et intelligibilité). Ces progrès sont notamment dus à l'utilisation de modèles statistiques, que nous détaillerons en section 1.4.

1.3.2.2 La synthèse par formants

La synthèse par formants repose sur un modèle linéaire source-filtre de production de la parole qui utilise des paramètres acoustiques comme entrée du synthétiseur. Le signal de source est convolué par des filtres, en parallèle ou en cascade, dont la fréquence centrale, l'amplitude et la bande passante correspondent à chacun des formants. Les valeurs des fréquences de ces formants sont déterminées à partir de l'analyse de signaux réels de voix. Ce type de synthèse permet des applications temps-réel. Les premiers synthétiseurs à formants sont apparus dans les années 50 avec les synthétiseurs PAT (*Parametric Artificial Talker*), synthétiseur avec des résonateurs en parallèle et le synthétiseur OVE I (*Orator Verbis Electris*), utilisant des filtres en cascade [48]. A cette époque, Fant [36] introduit les bases du modèle source-filtre. La synthèse par formants peut également être effectuée d'un point de vue temporel, en utilisant des formes d'onde formantiques (FOF). Il s'agit d'impulsions espacées d'une période fondamentale. Ainsi, chaque FOF représente une période d'un signal correspondant à un formant. La voix de synthèse est obtenue par sommation de chacune de ces contributions. Le synthétiseur de voyelles chantées CHANT [49] est fondé sur cette

technique. Cependant, c'est un synthétiseur destiné à la composition qui ne permet pas un contrôle du synthétiseur en temps réel.

1.3.2.3 La synthèse par concaténation

La synthèse par concaténation utilise de courts segments de parole préenregistrée, généralement des diphones[4]. En pratique, la taille des unités sélectionnées n'est pas uniforme (diphones, triphones…) afin d'obtenir plus de réalisme. A moins de disposer de bases de données conséquentes, cela implique d'utiliser un jeu de paramètres spécifique par locuteur. La prosodie peut être contrôlée par des méthodes de modification de hauteur comme PSOLA [50]. Les systèmes commerciaux utilisent généralement la synthèse par concaténation parce qu'elle fournit, au prix de bases de données suffisantes, un degré de réalisme important. Traditionnellement, la synthèse par concaténation engendre de nouveaux segments de parole en réarrangeant des segments de parole préenregistrée en tenant compte des contraintes acoustiques et linguistiques. Si cette méthode est bien paramétrée, cela permet de produire de la parole synthétique de bonne qualité en termes de naturalité et intelligibilité. Cependant, ce type de synthèse nécessite des bases de données équilibrées et étiquetées avec précision.

1.3.2.4 La méthode HTS

La méthode HTS [51] utilise une méthode de reconstruction de séquence de phonèmes fondée sur les HMMs (*Hidden Markov Models*). Sur une base de données de parole, les segments de parole sont étiquetés et des descripteurs spectraux sont extraits sur chacun de ces segments. Le modèle de Markov caché est ensuite entraîné sur cette base. La forme d'onde est générée en utilisant les valeurs de la fréquence fondamentale F_0 et des descripteurs spectraux des séquences les plus proches de la cible, à l'aide de calculs probabilistes basés sur les HMMs. Ainsi, aucune forme d'onde n'est stockée. En revanche, la base de données de parole est utilisée pour entraîner un ensemble de modèles de phonèmes dépendant du contexte. Ces

[4] En phonétique, un phone désigne un son d'une langue. Un diphone correspond à une paire de phones adjacents. Le phonème désigne la plus petite unité distinctive que l'on peut isoler dans un segment de parole. L'identification des phonèmes d'une langue est obtenue par construction de paires minimales, c'est-à-dire de paires de mots de sens différents et dont un seul son diffère de l'un à l'autre. Les phones sont donc les différentes réalisations d'un phonème.

phonèmes sont ensuite utilisés pour piloter un vocodeur au moment de la synthèse. Il existe également des méthode hybrides de synthèse par concaténation dans lesquelles un modèle statistiue paramétrique guide la sélection des segments [52].

1.3.2.5 La synthèse par modèle physique

La synthèse par modèles physiques repose sur des données articulatoires et l'analyse de l'évolution dynamique des articulateurs impliqués dans la phonation. Il s'agit de modéliser de façon explicite le mécanisme de production de la parole. Il est possible soit de s'intéresser au comportement d'un articulateur en particulier, soit de considérer l'appareil vocal dans sa globalité. Dans un premier temps, on simule les mouvements des articulateurs, ce qui nécessite un modèle de contrôle des articulateurs. Ensuite, on convertit ces informations de mouvement en succession continue de géométries du conduit vocal, qui s'appuie sur un modèle de celui-ci. Ensuite, à partir de ces informations géométriques et d'un modèle acoustique, on produit le signal acoustique. Les plis vocaux sont modélisés comme un système mécanique oscillant composé de deux masses reliées entre elles par un ressort de raideur linéaire et chacune maintenue à un support fixe par un ressort amorti linéairement et de tension non linéaire [53]. Le larynx est modélisé par un système multi-masses ou un modèle à poutres. Ces modèles mettent en évidence la physique du conduit vocal mais sont plus complexes que les simples modèles à deux masses. Un autre modèle physique [54] propose 7 paramètres de contrôle de la géométrie du conduit vocal : la position du corps de la langue, l'arrondissement et la protrusion des lèvres, les lieux et degré de constriction de la pointe de la langue, le degré de couplage avec la cavité nasale. Le calcul de l'onde acoustique résultante utilise souvent des bases de données élaborées à partir de mesures radiographiques qui décrivent l'aire des coupes sagittales à travers le conduit vocal [55]. Un modèle articulatoire fondé sur le geste phonétique et les contraintes de coordination entre les articulateurs a été développé en 1986 [56]. Il utilise l'activation temporelle de chaque geste ainsi que la coordination entre les articulateurs dont les mouvements évoluent selon les gestes. Les modèles plus récents incluent des modèles de glotte et de source de bruit et produisent des voix de bonne qualité, en particulier pour les voyelles statiques et les consonnes comme les fricatives, latérales et nasales y compris en voix chantée [43]. Dans ces travaux, un ensemble de règles permet de transformer les données de partition musicale en partition gestuelle puis en signal acoustique associé. Toutes les méthodes de synthèse présentées dans la section 1.3.2 ne produisent pas des signaux acoustiques d'égale qualité. Cette qualité peut en effet

concerner le réalisme des sons (la naturalité) ou leur intelligibilité.

1.3.3 Naturalité et intelligibilité

Dans [52], qui a inspiré la discussion présentée dans ce paragraphe, l'auteur décrit que les systèmes paramétriques statistiques sont les systèmes qui produisent la parole synthétique la plus intelligible, mais elle n'apparaît pas très naturelle à entendre. A l'inverse, la synthèse par concaténation, qui est décrite comme la solution permettant la voix la plus naturelle, produit des paroles bien moins intelligibles que les modèles paramétriques. Ainsi les systèmes paramétriques permettent d'atteindre une naturalité et une intelligibilité presque satisfaisante.

1.3.3.1 Les vocodeurs

La synthèse paramétrique repose sur l'utilisation d'un vocodeur afin de convertir les formes d'onde de façon paramétrique, puis de convertir les paramètres générés par le modèle en signaux acoustiques au cours de la synthèse. Il existe de nombreux types de vocodeurs. Le plus utilisé d'entre eux est le vocodeur STRAIGHT [57]. Ce synthétiseur a pour but de réaliser la séparation source-filtre, bien qu'il ne soit pas à proprement parler un modèle source-filtre. Il permet de modéliser l'enveloppe spectrale sans modélisation explicite du conduit vocal. Durant la phase d'analyse, les signaux de parole sont convertis en paramètres du modèle. Au lieu d'adopter un modèle particulier du conduit vocal, le modèle STRAIGHT part de l'hypothèse que l'enveloppe spectrale est lissée à la fois en temps et en fréquence. Ce modèle utilise une fenêtre adaptée en fonction de la fréquence du son afin de réduire les interférences harmoniques lors de l'estimation de cette enveloppe spectrale. Afin de faire la synthèse du signal de parole, un filtre doit être conçu à partir de l'enveloppe spectrale. Ce filtre est excité par un signal de source qui mélange un train d'impulsion à phase modifiée avec du bruit mis en forme.

Il existe d'autres types de vocodeurs comme les vocodeurs sinusoïdaux ou les vocodeurs harmoniques plus un bruit (*harmonic-plus-noise vocoders*) [58]. Ces vocodeurs se différencient des vocodeurs de type STRAIGHT par le fait qu'ils n'utilisent pas de modèle source-filtre. Ces vocodeurs tentent de modéliser le signal acoustique directement, sans référence explicite à aucun modèle de production de parole. Le signal de parole est modélisé comme la somme d'une partie déterministe (la structure harmonique, modélisée comme un

ensemble de sinusoïdes) et une partie stochastique (du bruit). Cette idée a donné naissance à des vocodeurs produisant moins d'artefacts que STRAIGHT. Cependant, comme indiqué dans [52], le nombre de paramètres nécessaires pour représenter le signal acoustique en utilisant un modèle *harmonic-plus-noise* est important et variable, ce qui le rend peu adapté pour une utilisation avec un *text-to-speech* (TTS) paramétrique. La qualité de la synthèse possible avec un vocodeur harmonique plus bruit en fait une solution malgré tout intéressante pour des implémentations hors ligne.

1.3.3.2 Modification des paramètres

1.3.3.2.1 Adaptation de modèle

La capacité à modifier les paramètres d'un système statistique paramétrique explique pourquoi ils sont si largement utilisés. Ainsi, transformer la fréquence fondamentale ou la vitesse de parole sont des modifications aisées lorsque l'on utilise des modèles paramétriques statistiques car il suffit de modifier la valeur de ces paramètres (valeur moyenne, écart-type). Mais il est également possible de faire des modifications plus sophistiquées, par exemple en appliquant des transformations différentes sur certains paramètres du modèle. Le modèle STRAIGHT possède l'avantage d'interpoler entre deux échantillons naturels et donc d'interpoler entre deux modèles statistiques. Il est ainsi possible de faire varier l'émotion, le style ou l'identité du locuteur. Ceci permet de créer des styles de voix en dehors des limites humaines.

1.3.3.2.2 L'édition automatique de signaux

Les méthodes paramétriques permettent un bon contrôle du signal de parole au cours du temps. Cependant, la forme d'onde est limitée par le vocodeur, qui impacte plus ou moins la naturalité du son produit. Seule la concaténation de signaux permet d'éviter cet écueil. Cependant, comme nous l'avons déjà indiqué, la concaténation est difficile à paramétrer et est très sensible à la finesse de l'étiquetage de la base de données. Ainsi, lors de la synthèse, on ne parvient pas toujours à choisir les sons qui semblent les plus naturels à cause de ces limites. Dans les systèmes commerciaux à temps différé, les segments sélectionnés pour la synthèse ne sont pas toujours les segments qui obtiennent le meilleur score en termes de coût mais un ajustement manuel d'après des critères perceptifs sélectionne parfois le deuxième ou troisième meilleur segment.

1.3.3.3 Synthèse de voix chantée

La synthèse de voix chantée a la particularité de nécessiter une grande expressivité, ce qui constitue un défi supplémentaire par rapport à la synthèse de voix parlée. La qualité de la source glottique, la précision de l'articulation et l'expressivité sont donc des critères déterminants pour l'évaluation de la qualité d'un extrait de voix chantée synthétique. Parmi les méthodes développées en synthèse vocale, des approches de synthèse par concaténation d'unités, l'utilisation de vocodeurs ainsi que des approches de synthèse articulatoire ont été proposées pour l'application en voix chantée. Une des méthodes consiste à utiliser un vocodeur pour produire de la voix chantée à partir d'un extrait de voix parlée et un codage de la musique, comme avec le vocodeur STRAIGHT [59] ou un vocodeur de phase [60]. La synthèse de voix chantée par concaténation d'unités a connu un grand succès avec le développement du système commercial VOCALOID [61]. La méthode de synthèse par formant [62] a l'avantage d'être très modulaire et de permettre de tester les différences perceptives entre différentes sources glottiques ou différentes configurations du conduit vocal. Un contrôle gestuel de la synthèse de voix chantée a été proposé dans [63] puis [64]. Dans le projet CantorDigitalis comme dans le projet Calliphony, une tablette graphique sert d'interface de contrôle. Une méthode de synthèse articulatoire a été présentée dans [65]. Une méthode de synthèse de voix chantée permettant une synthèse expressive de bonne qualité a été proposée dans [66]. De façon générale, la synthèse par concaténation d'unités semble permettre une meilleure naturalité du son [67].

Afin de modéliser au mieux la voix chantée, nous souhaitons compléter les informations obtenues à partir du signal acoustique par des informations multimodales sur le geste vocal. Les types de chant étudiés mettent en œuvre des techniques variées et complexes, c'est pourquoi souhaitons utiliser des modèles d'apprentissage statistique afin d'extraire des informations permettant le développement d'outils pédagogiques adaptés à l'apprentissage de ces techniques de chant.

1.4 Méthodes d'apprentissage statistique

1.4.1 Introduction à l'apprentissage statistique

Nous identifions en section 1.5 des appareils de mesure, qui nous permettent de collecter des données complexes corrélées au fonctionnement des articulateurs du conduit vocal. Nous voudrions construire des modèles permettant de transformer ces données brutes en des indicateurs ayant un sens ou une utilité. L'apprentissage statistique rassemble des méthodes qui permettent de construire un modèle à partir de données, en contrôlant la qualité du modèle ainsi que sa capacité de généralisation face à de nouvelles situations [68]. En apprentissage statistique, l'algorithme est capable d'estimer les paramètres d'un modèle depuis des données d'entrée. L'apprentissage statistique trouve sa motivation dans le fait de résoudre certaines tâches de prédiction lorsque l'interprétation directe est délicate. L'apprentissage peut se faire de façon supervisée ou non supervisée.

Dans le cas d'un apprentissage supervisé, les exemples d'apprentissage sont fournis sous la forme de couples entrée/sortie désirée (x_i, y_i). L'objectif est de déterminer une sortie $\boldsymbol{y}$ pour chaque nouvelle entrée $\boldsymbol{x}$ qui soit le plus proche possible de la sortie attendue – on mesure alors la distance entre la sortie obtenue et la sortie désirée avec une fonction de coût. Si la sortie représente un nombre fini de classes, on parle alors de tâche de classification. Si la sortie représente des valeurs continues, il s'agit alors d'une tâche de régression. Ainsi, l'apprentissage supervisé consiste à inférer une sortie pour une entrée donnée, connaissant une base d'exemples formée de couples entrée-sortie différents. Parmi les méthodes d'apprentissage supervisé, nous pouvons citer les régressions linéaires, les perceptrons multicouches ou encore les machines à vecteur support.

Dans le cas d'un apprentissage non supervisé, les exemples d'apprentissage fournis au système se résument aux entrées x_i. L'objectif est alors de trouver, sans apriori sur les données, des subdivisions des entrées en sous-groupes homogènes. Parmi les méthodes d'apprentissage non-supervisé, nous pouvons citer les méthodes de *clustering*. L'apprentissage statistique est couramment utilisé en traitement du signal, lorsque l'on souhaite disposer d'algorithmes adaptatifs, pour des tâches trop complexes pour être décrites de façon déterministe. Les étapes d'un algorithme d'apprentissage sont les suivantes : tout d'abord, il s'agit d'identifier le problème d'apprentissage et de construire la base de données. Il convient ensuite de choisir une représentation numérique pertinente des données. Ensuite,

on entraîne le modèle sur un jeu de données nommé base d'apprentissage en ajustant les paramètres du modèle. Une base dite de validation est utilisée afin de valider l'apprentissage et d'ajuster les hyperparamètres du modèle. Enfin, les données de la base de test permettent de tester les performances de généralisation de l'algorithme. Chacune de ces bases doit être statistiquement représentative des données et disjointe des deux autres bases.

Les méthodes d'apprentissage statistique ont de nombreuses applications. En traitement de la parole, les algorithmes de type réseaux de neurones sont des candidats sérieux pour certaines tâches comme par exemple la reconnaissance de locuteur [69] ou de phonème [70].

1.4.2 Notions de *Shallow learning* et *Deep Learning*

Dans un algorithme classique d'apprentissage statistique, la première difficulté, une fois les données collectées, est de trouver des descripteurs pertinents permettant de représenter les données et de contenir de l'information utile pour la tâche souhaitée. Ainsi, pour chaque modèle considéré, plusieurs types de descripteurs peuvent être étudiés avant de trouver une description satisfaisante des données. L'utilisation de descripteurs géométriques ou de moments statistiques sont des méthodes couramment utilisées pour obtenir des descripteurs. Les risques sont que l'ensemble des descripteurs soit incomplet ou bien au contraire redondant. Un autre problème concerne la collecte de données, qui peuvent être de qualités variables. En outre, les échantillons de la base d'apprentissage doivent être représentatifs des données à partir desquelles le modèle est construit.

Un réseau de neurones artificiels correspond à une association en un graphe d'objets élémentaires appelés neurones formels. L'architecture de ce graphe (par exemple en couches), son niveau de complexité (par exemple la présence ou non de boucles de rétroaction), les fonctions d'activation des neurones (par exemple sigmoïde) sont des exemples de critères permettant de distinguer les réseaux de neurones. L'analogie avec un réseau de neurones biologique peut se faire en considérant les entrées d'un neurone comme des dendrites, les connexions avec les autres neurones comme des synapses, la fonction d'activation comme un noyau qui active la sortie en fonction des stimulations en entrée et la sortie du neurone comme un axone. L'apprentissage profond ou *Deep Learning* (par opposition au *shallow learning*, apprentissage peu profond) est un apprentissage réalisé sur un réseau de neurones avec

plusieurs couches cachées. Le principe du *Deep Learning* repose sur un apprentissage hiérarchique couche par couche. Entre chaque couche interviennent des transformations non linéaires et chaque couche reçoit en entrée la sortie de la couche précédente. Dans le *Deep Learning*, l'extraction de descripteurs est pilotée directement à partir les données. Autrement dit, le *Deep Learning* repose donc sur un paradigme d'apprentissage que l'on pourrait qualifier de « supervisé par les entrées » – où les sorties attendues du modèle sont les entrées elles-mêmes. Dans ce paradigme, l'apprentissage dépend d'une fonction de coût (comme dans les apprentissages supervisés), sans avoir pour autant à fournir de données de sortie au modèle (comme dans les apprentissages non-supervisés).

L'information contenue dans des données peut être représentée de différentes manières. Par exemple, une image peut être codée comme un vecteur de valeurs d'intensité par pixel, ou bien un ensemble de contours, de régions avec une forme particulière. Certaines représentations permettent un meilleur apprentissage de certaines tâches à partir d'exemples [71]. Un des atouts du *Deep Learning* est de remplacer la détermination manuelle de descripteurs par des algorithmes d'extraction de descripteurs hiérarchiques. Il existe plusieurs manières de construire un réseau de neurones profond, notamment le DBN (*Deep Belief Network*). La méthode la plus répandue afin d'entraîner efficacement un réseau de neurones profond est d'utiliser un algorithme glouton (algorithme qui recherche, étape par étape, un minimum local) d'apprentissage couche par couche par le biais de machines de Boltzmann Restreintes. Plus précisément, il s'agit d'entraîner de façon non supervisée chaque couche afin d'extraire les descripteurs principaux à partir de la distribution des données d'entrée. La première couche cachée correspond donc à une représentation de ces entrées. Cette représentation est ensuite utilisée comme entrée pour la couche suivante. La méthode de *Deep Learning* peut être utilisée comme initialisation des poids et biais avant l'utilisation d'un algorithme supervisé comme la rétro-propagation du gradient (cette méthode permet de calculer le gradient de l'erreur pour chaque neurone d'un réseau de neurones, de la dernière couche vers la première. Dans l'apprentissage d'un réseau profond, la rétro-propagation joue alors de rôle de *fine-tuning*). L'utilisation d'une telle stratégie d'apprentissage de réseaux profonds est plutôt efficace. Il a été montré [72] qu'initialiser les poids d'un perceptron multicouche avec un réseau profond (type *Deep Belief Network*, ou DBN) donnait de meilleurs résultats qu'une initialisation aléatoire.

Utiliser un DBN a donc plusieurs avantages, notamment le fait que les unités cachées les plus profondes peuvent être calculées efficacement ; l'apprentissage glouton par empilement de RBM permet une réduction de la complexité de l'apprentissage liée à la profondeur du réseau [73]. Ceci explique pourquoi les DBN ont été utilisés dans de nombreuses applications de traitement du signal, comme détaillé dans [74]. Les applications du *Deep Learning* dans le domaine de l'acoustique et du traitement de la parole sont largement discutées dans [70]. Une stratégie d'apprentissage dite gloutonne (voir [75]) de représentations sur un réseau profond utilise les machines de Boltzmann restreintes.

1.4.3 Les machines de Boltzmann restreintes (RBM)

1.4.3.1 Machines de Boltzmann et restrictions

Les machines de Boltzmann, décrites dans [76] et [77], sont des réseaux utilisés pour apprendre des représentations internes dans des problèmes à la combinatoire élevée (voir Figure 24). Leur nom provient de la distribution de Boltzmann, modèle physique utile pour prédire la distribution des particules d'un gaz entre différents niveaux d'énergie.

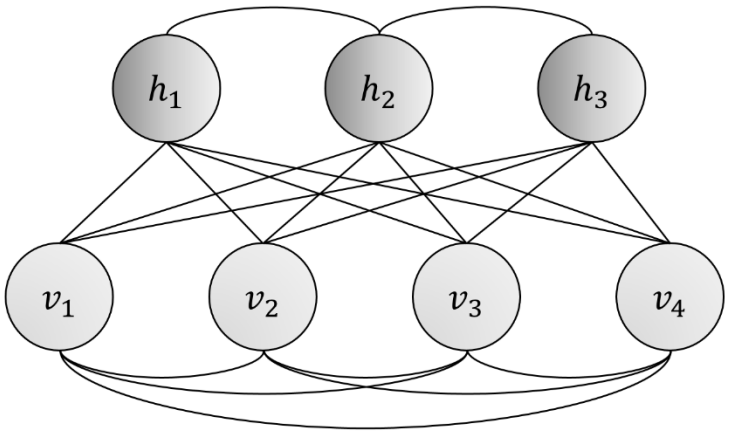


Figure 24 - Illustration d'une machine de Boltzmann. Dans une machine de Boltzmann, des connexions existent entre les différentes unités cachées et les différentes unités visibles, d'après [78]. Les unités visibles sont les unités dont l'état peut être observé. L'état des unités cachées n'est pas spécifié par les données observables.

Dans une machine de Boltzmann, de même que dans un réseau de Hopfield, des unités binaires sont connectées de façon à former un réseau dont l'énergie globale est définie comme une combinaison des états de ces unités plus un biais. Une machine de Boltzmann restreinte est un réseau de neurones stochastique capable d'apprendre une distribution de probabilités à partir d'unités d'entrée. Les unités des différentes couches peuvent être activées (*on*) ou

désactivées (*off*). Ces unités sont connectées les unes aux autres par des liens bidirectionnels. Les poids affectés à ces connexions sont symétriques, c'est-à-dire que le poids du neurone N_i au neurone N_j est égal au poids du neurone N_j au neurone N_i. Les poids peuvent prendre des valeurs positives ou négatives. La probabilité qu'une une unité se trouve dans un état on dépend de la distribution des unités voisines ainsi que des connexions entre ces unités. Dans une machine de Boltzmann, les seules restrictions sont qu'aucune unité n'a de connexion avec elle-même et que toutes les connexions sont symétriques. Cependant, en raison de leur grande complexité, ces réseaux sont bien moins utilisés que les Machines de Boltzmann Restreintes (RBM), qui sont des Machines de Boltzmann dans lesquelles les connexions entre les unités sont limitées, formant ainsi un graphe biparti [78]. Les Machines de Boltzmann Restreintes se sont largement répandues depuis 2006 grâce aux progrès des capacités de calcul [72] et au développement d'algorithmes rapides. Les applications les plus courantes des RBM sont la réduction de dimension, la classification et la modélisation et peuvent être utilisées de façon supervisée ou non.

Les RBM vérifient les propriétés suivantes :

- Elles ont une seule couche d'unités binaires stochastiques cachées ;
- Il n'y a pas de connexion entre les unités visibles, de même entre les unités cachées, les seules interactions possibles sont les connexions entre une unité cachée et une unité visible (voir Figure 25) ;
- Les unités cachées sont conditionnellement indépendantes connaissant les unités visibles.

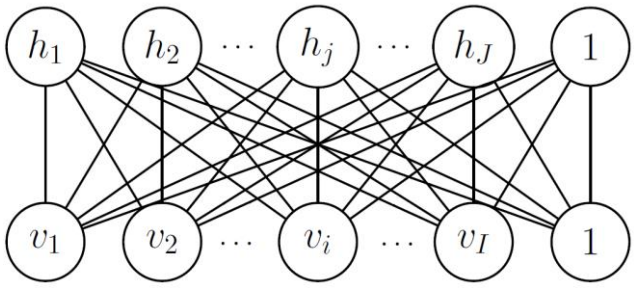


Figure 25 – Un RBM avec I unités visibles et J unités cachées, I et J pouvant prendre des valeurs distinctes, d'après [18].

1.4.3.2 Etapes d'apprentissage des Machines de Boltzmann Restreintes

L'apprentissage consiste à modifier l'intensité de la connexion entre les unités de façon à ce que tout le réseau développe un modèle interne qui capture la structure sous-jacente des

données. Supposons notre réseau composé d'un ensemble d'unités visibles $\boldsymbol{v} \in \{0,1\}^I$ et d'un ensemble d'unités cachées $\boldsymbol{h} \in \{0,1\}^J$. L'énergie de l'état $\{\boldsymbol{v}, \boldsymbol{h}\}$ est donnée par :

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\frac{1}{2}\boldsymbol{v}^T L\, \boldsymbol{v} - \frac{1}{2}\boldsymbol{h}^T J\, \boldsymbol{h} - \boldsymbol{v}^T W\, \boldsymbol{h}, \tag{20}$$

où θ représente les paramètres du modèle, W, J et L, représentant respectivement les termes d'interaction visible-cachée, visible-visible et cachée-cachée. Dans le cas d'une machine de Boltzmann restreinte, les interactions visible-visible et cachée-cachée sont supposés inexistantes, donc les matrices J et L sont nulles. L'énergie de l'état $\{\boldsymbol{v}, \boldsymbol{h}\}$ devient :

$$E(\boldsymbol{v}, \boldsymbol{h};\ \theta) = -\boldsymbol{v}^T W \boldsymbol{h} \tag{21}$$

La probabilité qu'un modèle affecte au vecteur $\boldsymbol{v}$ des états des unités visibles est :

$$p(\boldsymbol{v}, \boldsymbol{h}; \theta) = \frac{1}{Z(\theta)} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h}; \theta)}, \tag{22}$$

$$Z(\theta) = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h}; \theta)}, \tag{23}$$

où $Z(\theta)$ la fonction de partition. Dans le cas d'une RBM, la probabilité que l'unité cachée j soit dans l'état 1 connaissant les états des unités visibles est :

$$p(h_j = 1 | \boldsymbol{v}) = \sigma\left(\sum_{i=1}^{I} w_{ji} v_i + b_j\right), \tag{24}$$

La probabilité que l'unité visible i soit dans l'état 1 connaissant les états des unités cachées est :

$$p(v_i = 1 | \boldsymbol{h}) = \sigma\left(\sum_{j=1}^{J} w_{ij} h_j + a_i\right). \tag{25}$$

où $\sigma(x)$ est la fonction sigmoïde $\sigma(x) = \frac{1}{1+e^{-x}}$, a_i et b_j sont des biais.

Par souci de simplicité, nous considérons une unité de biais toujours active (les unités constamment actives sur la droite de la Figure 25), présente dans la couche visible comme dans la couche cachée. Nous réécrivons alors les probabilités conditionnelles de l'équation

(24) et (25) comme suit :

$$p(h_j = 1|\boldsymbol{v}) = \sigma\left(\sum_{i=1}^{I} w_{ji} v_i\right) \tag{26}$$

$$p(v_i = 1|\boldsymbol{h}) = \sigma\left(\sum_{j=1}^{J} w_{ij} h_j\right) \tag{27}$$

Ces équations sont utilisées pour mettre à jour les valeurs des unités au cours de l'apprentissage. Le principe consiste à alterner entre la mise à jour des unités cachées et la mise à jour des unités visibles [79]. La mise à jour des poids se fait par un algorithme de descente de gradient :

$$w_{ij}(t+1) = w_{ij}(t) + \varepsilon \frac{\partial \log\big(p(\boldsymbol{v})\big)}{\partial w_{ij}} \tag{28}$$

où

$$p(\boldsymbol{v}) = \frac{1}{Z}\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h};\theta)} \tag{29}$$

$$\Delta w_{ij} = \varepsilon \frac{\partial \log\big(p(\boldsymbol{v})\big)}{\partial w_{ij}} = \frac{\varepsilon}{Z}\left[\frac{\partial E_{\text{données}}}{\partial w_{ij}} - \frac{\partial E_{\text{modèle}}}{\partial w_{ij}}\right] \tag{30}$$

Le gradient est de la forme $\langle \boldsymbol{v}\,\boldsymbol{h}\rangle_{\text{données}} - \langle \boldsymbol{v}\,\boldsymbol{h}\rangle_{\text{modèle}}$, où $\langle \ldots \rangle_d$ représente les moyennes relativement à la distribution d. La mise à jour des poids du réseau se fait ainsi par minimisation de la divergence contrastive (CD) entre deux distributions, qui sont la distribution "objectif" $\langle \boldsymbol{v}\,\boldsymbol{h}\rangle_{données}$ de la base d'apprentissage, à savoir l'entrée, et la distribution $\langle \boldsymbol{v}\,\boldsymbol{h}\,\rangle_{modèle}$ modélisée par la machine de Boltzmann. Il s'agit d'une différence entre deux divergences de Kullback-Liebler :

$$CD = \langle \boldsymbol{v}\boldsymbol{h}\rangle_{\text{données}} - \langle \boldsymbol{v}\boldsymbol{h}\rangle_{\text{modèle}} \tag{31}$$

La règle d'apprentissage utilisée est la suivante :

$$\Delta w_{ij} = \varepsilon \times CD = \varepsilon\big(< v_i h_j >_{\text{donnée}} - < v_i h_j >_{\text{modèle}}\big), \tag{32}$$

En pratique, on approche cette fonction par une méthode appelée échantillonnage de Gibbs

(*Gibbs sampling*, voir Figure 26) et l'on utilise l'espérance du produit des unités cachées et visibles relatives aux données et celles relatives au modèle.

où ε désigne le taux d'apprentissage.

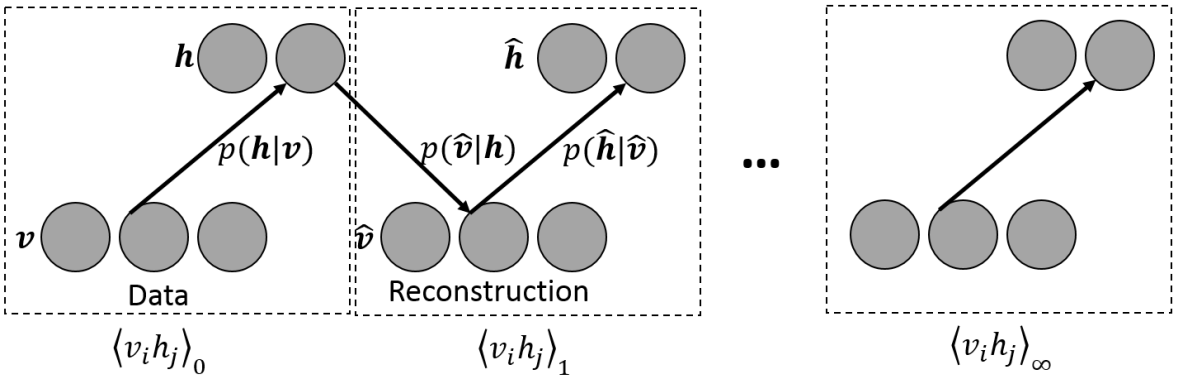


Figure 26 – Processus d'apprentissage d'une RBM. Il est possible de calculer une approximation de la divergence contrastive à partir des deux premières itérations. Le vecteur $\widehat{\boldsymbol{v}}$ désigne l'estimation du vecteur $\boldsymbol{v}$, le vecteur $\widehat{\boldsymbol{h}}$ désigne l'estimation du vecteur $\boldsymbol{h}$.

La méthode d'échantillonnage de Gibbs consiste en une succession d'étapes [80] :

- Initialiser les unités visibles à partir d'un vecteur d'apprentissage, initialiser les poids aléatoirement
- Mettre à jour les unités cachées connaissant les unités visibles en utilisant l'équation (26)
- Mettre à jour les unités visibles connaissant les unités cachées en utilisant l'expression (27)
- Mettre à jour en parallèle les unités cachées connaissant les unités visibles en utilisant à nouveau l'expression (26)
- Mettre à jour les poids d'après l'équation (31)

L'approximation pourtant très brutale de n'effectuer qu'une seule fois (on arrête le calcul à $\langle v_i h_j \rangle_1$) les étapes décrites ci-dessus permet d'approximer rapidement le calcul de la divergence contrastive [81] pour l'entrainement de RBM. La construction d'architectures profondes peut se faire par empilement de RBM.

1.4.4 Empilement de RBM

Une stratégie d'apprentissage pour un réseau de neurones profond (*Deep network*) consiste à empiler des RBM appris couche par couche, en partant des entrées puis en utilisant la sortie

de la couche i comme entrée pour la couche $i + 1$. Ainsi, une fois qu'un RBM est entraîné, un autre RBM peut être empilé à la suite du premier RBM afin de créer un modèle multicouche.

Les autoencodeurs, décrits dans [82], sont des structures composées de deux parties : un encodeur et un décodeur. Les autoencodeurs profonds (ou *Deep Auto-Encoders*, DAE), décrits par [79] et [83], sont des autoencodeurs construits avec des architectures profondes. Le nombre de neurones dans la dernière couche du décodeur est égal à la dimension de l'entrée du réseau. Le but d'un autoencodeur est de trouver une représentation codée d'une entrée pouvant être décodée avec précision. Un tel réseau est entraîné de sorte à trouver une représentation des données d'entrée et apprendre le lien entre une entrée et sa représentation cachée. Notons $\boldsymbol{x}$ l'entrée d'un autoencodeur, h_i la $i^{\text{ème}}$ couche cachée et $f_{\varphi i}$ la fonction d'encodage de la couche i pour un ensemble de paramètres φ donné. Notons également $\hat{\boldsymbol{x}}$ la reconstruction de l'entrée $\boldsymbol{x}$ par le décodeur et $f'_{\varphi i}$ la fonction de décodage liée à la couche i. Un exemple d'autoencodeur est montré Figure 27.

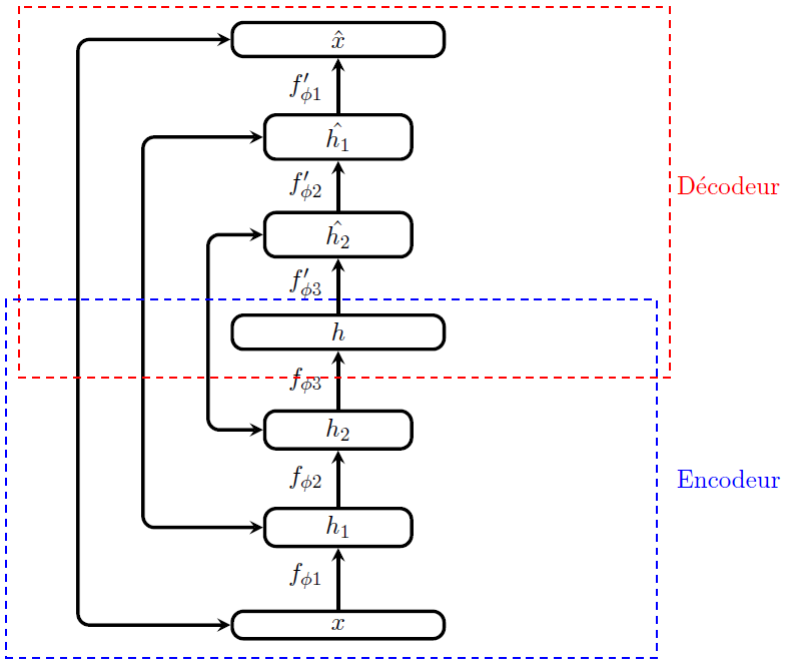


Figure 27 – Un exemple d'autoencodeur. Sa sortie $\hat{\boldsymbol{x}}$ est la reconstruction de l'entrée $\boldsymbol{x}$ à partir de la représentation cachée $\boldsymbol{h}$.

L'utilisation d'une telle méthode nécessite l'acquisition de nombreuses données.

1.5 Enregistrement de données articulatoires

1.5.1 Les méthodes d'enregistrement de données articulatoires

Notre objectif est de piloter un modèle du conduit vocal, qui pourrait être utilisé à des fins pédagogiques : par exemple pour apprendre à articuler et prononcer correctement des phonèmes qui n'existent pas dans notre langue maternelle, ou bien pour apprendre des techniques de chant. Pour ce faire, nous avons besoin de quantifier et mesurer le comportement des articulateurs. Ceci nous permettra d'étudier le rapport entre les dynamiques anatomiques du conduit vocal et ses productions sonores. En particulier, nous nous intéressons aux mouvements de la langue.

Le conduit vocal peut être étudié grâce à de nombreux capteurs. Certains d'entre eux offrent la possibilité de visualiser les mouvements des articulateurs au cours de la phonation, en voix parlée ou en voix chantée. Chaque méthode possède des avantages et des inconvénients qui la rendent plus adaptée à une tâche ou à une autre [84]. Les caractéristiques de différents capteurs permettant de visualiser le conduit vocal sont rassemblées dans le Tableau 2.

Tableau 2 : Méthodes d'analyse du conduit vocal (tableau de synthèse, d'après [84]).

Instrument	Avantage(s)	Inconvénient(s)
Endoscopie rigide	Images de bonne qualité	Très invasif, nécessite en général une anesthésie
Fibroscopie	Visualisation directe du larynx	Invasif, peut nécessiter une anesthésie
Caméra externe	Non-invasif	Pas de vue interne du conduit vocal
IRM	Vue très détaillée du conduit vocal	Fréquence d'imagerie trop faible, nécessite l'immobilité, très coûteux, potentiellement dangereux.
Radiographie X	Images des os	Potentiellement dangereux, nécessite l'immobilité, très coûteux
Articulographie électromagnétique (EMA)	Résolution spatiale < mm, mouvement des articulateurs en fonction du temps (lèvres, langue, mâchoire, velum)	Invasif, difficile à calibrer
Nasographie/transillumination	Informations à propos de la nasalité	Invasif
Électropalatographie	Position de la langue par rapport à celle du palais	Position de la langue incomplète, inconfort
Ultrason (échographie)	Mouvements de la langue en temps réel	Possible inconfort, pas de référence de la position de la langue
Electroglottographie	Temps réel, non-invasif	Peu d'informations disponibles

Certaines de ces méthodes, comme l'endoscopie rigide, la fibroscopie et la nasographie sont très invasives et ne peuvent pas être utilisés pendant la pratique du chant. D'autres, comme les rayons X ou l'IRM, nécessitent une immobilité et sont potentiellement dangereux sur le long

terme, ce qui signifie que ces techniques ne peuvent être utilisées que comme référence en position de repos et l'exposition d'un sujet doit être limitée. L'échographie semble donc un bon compromis pour la visualisation du conduit vocal en temps réel [85].

1.5.2 L'échographie

L'échographie est une technique d'imagerie qui utilise des ondes acoustiques de très hautes fréquences (ondes ultrasonores). Une onde ultrasonore est par définition une onde dont la fréquence est supérieure à la limite maximale des fréquences audibles pour l'oreille humaine, qui est de 20 kHz. Dans le domaine de l'imagerie médicale, les fréquences sont de l'ordre du Méga Hertz. Les échographes que nous utilisons produisent des ondes de fréquences comprises entre 4 et 8 MHz. Parce que ces ondes se réfléchissent sur la surface des objets qu'elles rencontrent, elles peuvent être utilisées pour visualiser divers organes. La possibilité d'obtenir des images en temps réel ainsi que le caractère non-invasif de cet instrument de mesure en fait l'outil privilégié depuis le début des années 80 pour observer les mouvements de la langue. Les images ultrasonores peuvent fournir une coupe sagittale de la surface de la langue (voir Figure 28), sur lesquelles le contour supérieur de la langue est très apparent. Il s'agit donc d'un outil bien adapté pour notre objectif, puisqu'il nous permettrait de modéliser les mouvements du contour de la langue sous forme de coordonnées de points. Comme nous l'avons vu en section 1.1, la position de la langue est corrélée au deuxième formant, et par conséquent est un indicateur fiable de la production des voyelles orales.

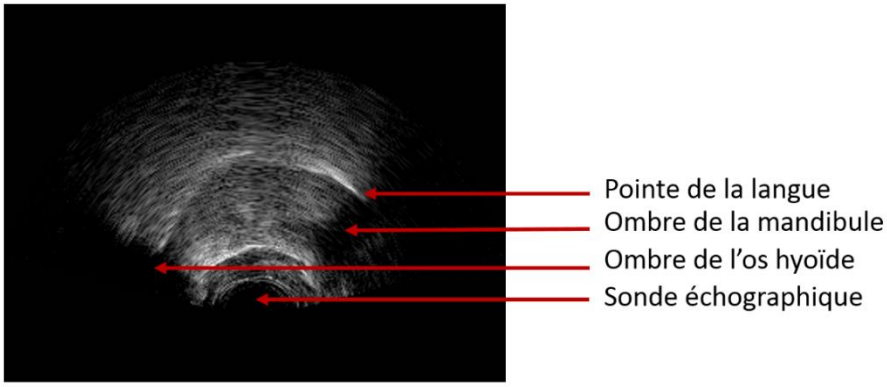


Figure 28 - Un exemple d'image ultrasonore de la langue. L'utilisation d'une sonde échographique placée sous le menton permet d'obtenir une coupe sagittale de la langue.

1.5.2.1 Principes physiques de l'imagerie ultrasonore

Dans la technique de l'échographie (détaillée dans [84], qui a inspiré cette description) des ondes ultrasonores sont émises dans la direction du milieu que l'on souhaite imager. Ces ondes sont des ondes dites de pression et leur propagation est à l'origine d'un phénomène de compression puis de dilatation du milieu traversé, phénomène qui se propage dans le milieu. Chaque milieu est caractérisé par une vitesse c de propagation de l'onde, qui dépend de sa densité et de son élasticité. La vitesse de propagation d'une onde ultrasonore est de 1480 m/s dans l'eau, contre 1540 m/s dans les tissus mous. En revanche, dans les tissus osseux, la vitesse de propagation d'une onde ultrasonore atteint 3000 m/s. Dans l'air, cette vitesse est de 340 m/s. On définit l'impédance acoustique Z d'un milieu comme le produit entre la densité ρ du milieu et la vitesse de propagation c d'une onde dans ce milieu.

$$Z = \rho\, c, \tag{33}$$

Deux phénomènes se produisent à l'interface de deux milieux : réflexion et réfraction (voir Figure 29). Dans le cas d'une réflexion, le faisceau est réfléchi d'un angle identique à l'angle d'incidence. Pour une réfraction, le faisceau incident est dévié d'un angle dont la valeur dépend du rapport entre les vitesses de propagation de l'onde dans les milieux traversés. Ainsi, à l'interface entre deux milieux d'impédance acoustique Z_1 et Z_2, si l'on note $r = \frac{Z_1}{Z_2}$, les rapports entre l'intensité incidente I_0, l'intensité réfléchie I_r et l'intensité transmise I_t sont donnés ci-dessous :

$$\frac{I_r}{I_0} = \left(\frac{r-1}{r+1}\right)^2, \tag{34}$$

et

$$\frac{I_t}{I_0} = \frac{4\,r}{(1+r)^2}. \tag{35}$$

Ainsi, plus le rapport r entre les impédances des milieux est élevé, plus la réflexion de l'onde est importante.

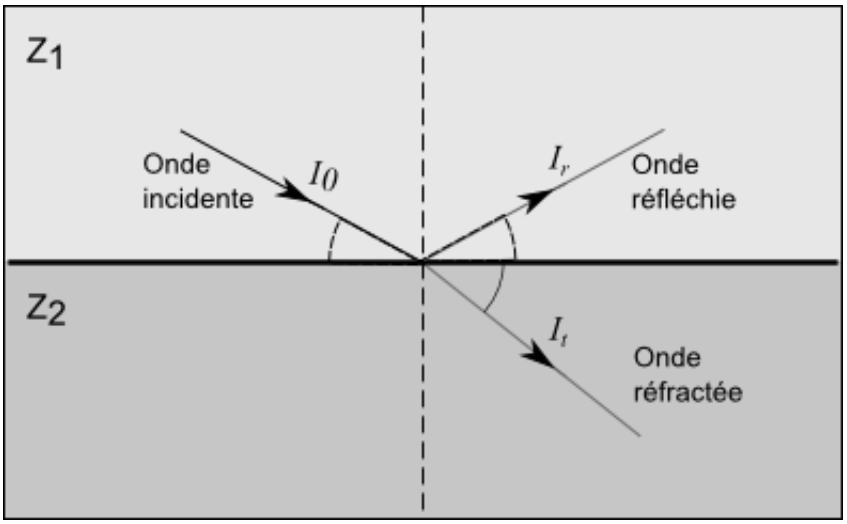


Figure 29 - Illustration des principes de réflexion et de réfraction de l'onde ultrasonore. A l'interface entre les deux milieux, une onde incidente d'intensité I_0 est réfléchie en une onde I_r et réfractée en une onde I_t, d'après [84].

Aux phénomènes de réflexions et de réfractions viennent s'ajouter les phénomènes de diffusion. Une onde de fréquence f, se propageant dans un milieu à une vitesse c, possède une longueur d'onde λ définie comme $\lambda = c/f$. Si la longueur d'onde est très petite devant la dimension du milieu, comme c'est le cas dans les tissus humains, les phénomènes de réflexion et réfraction prédominent. En revanche, si le milieu traversé est très hétérogène, comme par exemple un tissu spongieux, il possède localement des impédances acoustiques différentes. Les milieux traversés sont de grandeur proche de la longueur d'onde et il y a alors un phénomène de diffusion. Il s'agit d'une réémission isotrope d'une partie de l'onde incidente. En outre, le phénomène d'absorption a pour conséquence l'atténuation de l'énergie transportée par l'onde dont l'intensité décroît exponentiellement avec la profondeur de pénétration dans les tissus. La profondeur d'exploration au-delà de laquelle l'onde est trop atténuée est inversement proportionnelle à la fréquence de l'onde.

1.5.2.2 Fonctionnement d'un transducteur ultrasonore

Un transducteur ultrasonore, comme présenté dans [84], est un dispositif permettant la conversion d'un signal électrique en une onde ultrasonore et réciproquement, en utilisant l'effet piézoélectrique. Les matériaux piézoélectriques ont en effet la propriété de se polariser électriquement sous l'effet d'une contrainte mécanique et de se déformer lorsqu'un champ électrique leur est appliqué. Si un matériau piézoélectrique est soumis à un champ électrique alternatif, il subit une alternance périodique de compression et dilatations, ce qui produit une onde de pression. Un transducteur échographique utilise l'effet piézoélectrique à la fois en émission et en réception : des signaux sinusoïdaux modulés par des impulsions électriques de

commande sont transformés en onde ultrasonore et les échos ultrasonores (issus des réflexions) sont convertis en courants électriques. L'onde ultrasonore est modulée par des impulsions brèves, c'est ce que l'on nomme une émission pulsée. Ainsi, l'onde ultrasonore n'est pas émise en continu, il y a un temps d'attente entre chaque émission. L'onde émise par le transducteur est transmise dans les différents milieux étudiés et se propage dans les tissus. Lors du passage entre deux milieux d'impédances acoustiques différentes, les phénomènes de réflexions et de diffusions sont à l'origine d'échos qui se propagent en direction du transducteur. Durant le temps d'attente, le transducteur est en mode récepteur et peut donc convertir ces échos en signal électrique. La distance d_{cible} entre l'émetteur et l'interface d'où provient l'écho est déduite du temps de vol t_{vol}, durée qui sépare l'émission de l'onde de la réception de l'écho.

$$d_{cible} = c\ \frac{t_{vol}}{2} \tag{36}$$

La vitesse de propagation de l'onde dans les tissus mous est de 1540 m/s. Comme une nouvelle émission ne peut avoir lieu tant que les échos n'ont pas été détectés, la durée entre deux émissions est fonction de la profondeur d'exploration. Ainsi, il y a un choix à faire entre une fréquence d'émission élevée et une grande profondeur d'exploration. Un transducteur échographique possède une centaine d'éléments piézoélectriques disposés de façon linéaire ou bien convexe. La sonde que nous avons utilisée est une sonde microconvexe pourvue de 128 éléments piézoélectriques. Dans le mode d'affichage de l'échographe le plus couramment utilisé, le temps de vol de l'écho et la position de l'élément piézoélectrique sur le transducteur permet de déterminer la position d'un point dans l'image. Des niveaux de gris permettent de représenter l'amplitude du signal électrique fourni par l'élément piézoélectrique. Pour un système échographique, il y a deux types de résolution spatiale : la résolution axiale et la résolution latérale. La résolution axiale concerne la résolution dans l'axe du faisceau ultrasonore, tandis que la résolution latérale est la résolution dans un plan perpendiculaire au faisceau. La résolution temporelle est la fréquence de répétition des images et dépend de la profondeur d'exploration maximale souhaitée.

1.5.3 L'électroglottographie

L'électroglottographie permet de mesurer un corrélat du signal de source du conduit vocal, ce qui est très utile dans notre cas, car il n'existe pas de mesure directe de l'activité de la source glottique. L'utilisation d'un électroglottographe (EGG) nous permet ainsi d'estimer les

paramètres de source indépendamment du filtrage opéré par le conduit vocal. L'électroglottographie repose sur la mesure de la différence de potentiel électrique entre deux électrodes placées au niveau du cou d'un sujet [17]. La mesure de cette tension permet, pour un courant constant, d'avoir accès à l'impédance électrique du cou. Cette impédance est fonction de l'ouverture glottique : elle augmente lorsque l'air peut circuler au niveau de la glotte car l'air est moins bon conducteur que les tissus humains. En effet, une ouverture glottique, étant par définition une diminution du contact entre les plis vocaux, sera caractérisée d'un point de vue électrique par une augmentation de l'impédance et donc de la tension entre les deux électrodes. A l'inverse, une fermeture glottique résulte d'un contact plus grand entre les plis vocaux et se traduit donc par une diminution de la tension entre les deux électrodes. Un électroglottographe est constitué d'un générateur qui fournit un courant alternatif dont la fréquence est de l'ordre du MHz, de deux électrodes et d'un circuit de démodulation de fréquence. L'ensemble est complété par un filtre passe-haut de fréquence de coupure comprise entre 5 et 40 Hz, qui permet d'éliminer les basses fréquences parasites. En effet, les mouvements du sujet, les contractions des muscles dans la zone du cou, ou le débit sanguin dans les artères et les veines ajoutent des artefacts basse fréquence sur le signal et n'indiquent en rien l'activité glottique. L'impédance mesurée varie à la fréquence de vibration des cordes vocales. La tension recueillie subit une modulation à cette fréquence. On retrouve dans la période de ce signal la période fondamentale du son émis. Le signal EGG est par ailleurs très riche en harmoniques et de l'énergie est visible dans ce spectre jusqu'à environ 20 kHz. Le signal EGG donne une information sur le contact entre les plis vocaux.

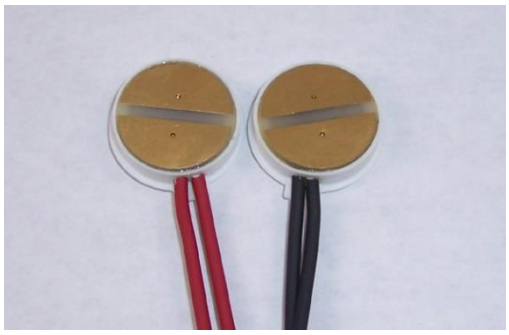


Figure 30 - Les deux électrodes d'un électroglottographe. Ces électrodes sont maintenues en position sur le cou du locuteur par un collier élastique.

1.5.4 Choix du matériel

Nous souhaitons visualiser de façon non-invasive les mouvements des articulateurs ainsi que de façon plus générale les gestes vocaux du chanteur pendant la pratique du chant. Diverses

études ont montré que la combinaison de plusieurs capteurs permet d'acquérir des données articulatoires (voir [86], [87] et [88]) et même de permettre le développement d'interfaces de parole silencieuses (voir [89], [90], [91]). Un microphone permet d'enregistrer le son produit par le chanteur. Afin d'étudier les mouvements de la langue, nous avons choisi l'imagerie échographique. Les mouvements des lèvres peuvent être enregistrés à l'aide d'une caméra. Un électroglottographe nous donne accès à des informations sur la source glottique. Nous avons choisi de rajouter deux capteurs, un accéléromètre positionné au niveau du nez afin de mesurer la nasalité du son et une ceinture de respiration placée au niveau du torse.

Afin de compenser les mouvements des chanteurs lors de leurs performances, une partie des capteurs ont été fixés sur un casque [92] (voir Figure 31), tandis que les autres sont directement placés en contact avec une partie du corps du chanteur. La sonde échographique, la caméra ainsi que le microphone sont fixés sur le casque. L'accéléromètre est placé directement sur le nez du chanteur, un « collier » permet de maintenir les électrodes de l'électroglottographe en place et une ceinture de respiration permet de mesurer l'amplitude des mouvements de respiration au niveau de la poitrine [93].

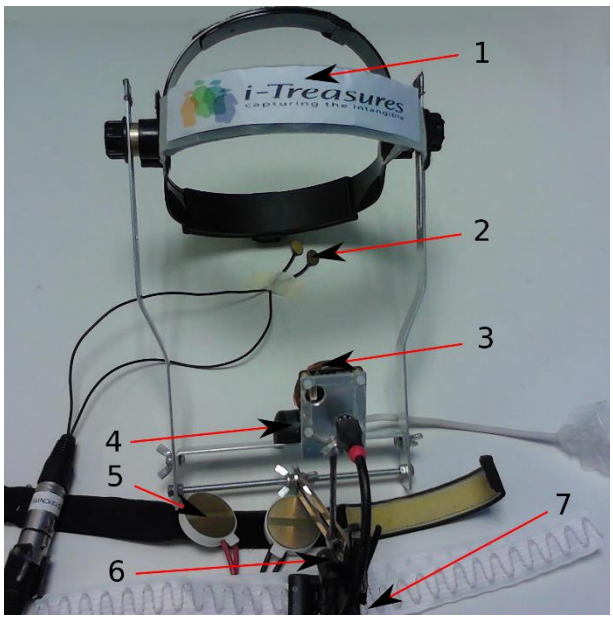


Figure 31 - Le casque d'acquisition des données du conduit vocal. 1. Casque permettant de fixer les capteurs. 2. Capteur piézoélectrique, placé sur le nez du sujet. 3. Caméra. 4. Sonde ultrasonore. 5. Electroglottographe 6. Microphone 7. Ceinture de respiration.

1.5.4.1 Cas de l'échographie

Pour l'imagerie de la langue, le choix d'une sonde convexe dont la fréquence d'émission est comprise entre 4 et 8 MHz est approprié. Afin de suivre de manière précise l'activité de la

langue, nous avons choisi un système d'acquisition à 60 Hz. Afin d'obtenir une coupe sagittale de la langue, la sonde doit être placée sous le menton et rester en contact avec la mâchoire tout au long de la production de son. L'utilisation de gel permet de limiter la présence d'air entre la mâchoire et la sonde. La difficulté est de maintenir de façon constante le contact entre la mâchoire et la sonde. Plusieurs approches sont possibles. L'approche choisie dans le système HATS [94] consiste à maintenir fermement la tête et la sonde dans une position donnée. L'ajout d'un petit coussinet rempli de gel sur la sonde permet de ne pas trop perturber la production et autorise un léger mouvement de la mâchoire. Le coussinet conserve le contact acoustique et se déforme en fonction des mouvements de la mâchoire inférieure. Cependant, ce genre de système est assez contraignant pour le locuteur. Il est également possible de n'imposer des contraintes que sur une partie du système, en ne fixant que la sonde à un support et en laissant au locuteur la possibilité de bouger sa tête. La mâchoire inférieure étant contrainte par la sonde, c'est la partie supérieure du crâne qui peut basculer vers l'arrière. On peut aussi maintenir immobile la tête du sujet (appui du front et du menton) mais laisser la sonde libre, en la tenant à la main par exemple. La sonde suit ainsi les mouvements de la mâchoire inférieure. D'autres types de systèmes ont été développés, en utilisant un casque sur lequel est fixée la sonde afin de maintenir le contact entre la sonde et la mâchoire du locuteur. Dans ces systèmes, la sonde et la tête peuvent se déplacer l'une par rapport à l'autre et il peut alors s'avérer intéresser de compenser l'un ou l'autre des mouvements. A l'inverse, le casque développé au laboratoire possède une plateforme de support ajustable pour le capteur ultrasonore, afin de le maintenir en contact avec le menton. La sonde que nous avons utilisée est une sonde microconvexe pourvue de 128 éléments piézoélectriques, conçue et réalisée afin d'en diminuer la taille et le poids. La taille ainsi que la fréquence de ces éléments piézoélectriques constituent des caractéristiques qui influent largement sur les propriétés de la sonde. Le champ d'émission de la sonde est de 140° permettant une bonne visualisation du mouvement de la langue. L'échographe que nous utilisons est le Terason T3000, un système léger et portable qui permet néanmoins l'enregistrement d'images de bonne qualité via la connexion Firewire d'un ordinateur.

1.5.4.2 Les autres capteurs

Le casque permet l'acquisition simultanée des données sur le conduit vocal, grâce à une sonde ultrasonore, mais aussi une caméra pour visualiser les mouvements des lèvres et un microphone pour enregistrer le signal acoustique. Afin de s'affranchir du problème des

variations d'éclairages, qui pourrait impacter l'efficacité des algorithmes de traitement des images, la caméra est équipée d'un filtre permettant de filtrer la lumière visible ainsi que d'un anneau de LEDs infrarouges. Un micro-cravate de la marque Audio-Technica Pro 70 est également fixé sur le casque afin d'enregistrer le signal acoustique. Par ailleurs, nous avons également choisi d'inclure trois autres capteurs qui ne sont pas fixés sur le casque. Ces autres capteurs sont en effet directement maintenus sur le corps du chanteur. Il y a un accéléromètre positionné sur le nez du chanteur, permettant de mesurer son activité nasale au cours de la phonation. Cet accéléromètre enregistre les vibrations au niveau du nez, d'où peuvent être extraits des marqueurs de nasalité. Un électroglottographe (EGG) (de la marque Glottal Enterprises Inc., modèle EG2-PCX2) est attaché au cou du chanteur. Ce capteur permet d'enregistrer un signal temporel dont les variations permettent de déterminer de façon fiable les ouvertures et fermetures glottiques. Un autre capteur, une ceinture de respiration, positionnée au niveau de la poitrine du chanteur, permet de mesurer la fréquence et l'amplitude des respirations du chanteur.

1.5.5 Informations contenues dans les données

Nous souhaitons obtenir des images sagittales de la langue. Pour cela, la barrette de transducteurs piézoélectriques de la sonde doit être placée dans le sens de la longueur de la langue. Cette coupe permet de visualiser l'interface entre la partie supérieure de la langue et l'air. A gauche et à droite de l'image se trouvent des ombres acoustiques, qui sont dues à la présence d'os, à savoir l'os hyoïde à gauche et l'os de la mâchoire à droite. En effet, les tissus osseux réfléchissent presque entièrement le faisceau ultrasonore et l'onde n'est pas transmise. Il arrive parfois que l'ombre acoustique de la mâchoire masque une partie de la langue. Ainsi, il n'est pas possible d'avoir une information fiable sur la position de la pointe de la langue. De même, la position du palais n'est pas directement visible, elle peut seulement être déduite du contact avec la langue. Un mouvement de déglutition permet de repérer la position du palais.

Nous souhaitons obtenir des images permettant de détecter le degré d'ouverture des lèvres. Les variations de niveau de gris entre les lèvres et l'intérieur de la bouche ou les dents doivent être suffisantes pour permettre de segmenter l'image de façon automatique, indépendamment de l'éclairage. L'association du filtre permettant de filtrer la lumière visible et de l'anneau de LED infrarouges permet d'obtenir des images de luminosité constante. Nous obtenons donc

des images en noir et blanc, dont l'intensité lumineuse est stable. Sur ces images, l'information essentielle est la forme des lèvres.

Le microphone doit permettre l'acquisition d'un signal acoustique malgré les mouvements et les gestes des chanteurs. Le système de fixation du microphone sur le casque permet d'enregistrer des signaux acoustiques pour lesquels la distance lèvres-microphone est constante. Ces signaux doivent être synchronisés avec les autres signaux enregistrés à l'aide de la carte son [92].

Le signal électroglottographique nous intéresse pour extraire des informations sur ses portions pseudopériodiques, en particulier pour les sons voisés. Sur ce signal, il est possible de détecter la période fondamentale ainsi que les instants de fermeture glottique et les paramètres de qualité vocale qui en découlent. Cependant, les conditions d'enregistrements n'excluent pas la présence d'artefacts dans le signal électroglottographique, principalement dus à des mouvements des muscles du cou.

Le signal issu de l'accéléromètre contient les informations de nasalité du chant. En effet, ce signal est quasi-nul pour des sons oraux et possède une amplitude significative dans le cas de sons nasaux.

Le signal de respiration est un signal qui a la particularité d'avoir une fréquence bien plus faible que les autres signaux enregistrés. Alors que les fréquences du signal acoustique, du signal électroglottographique et du signal issu de l'accéléromètre sont de l'ordre de la centaine de Hz, le signal de respiration est de l'ordre du Hz. La carte audio n'étant pas prévue pour enregistrer des fréquences aussi basse (elle inclut un filtre passe-bas à 1 Hz), nous avons dû faire une modulation d'amplitude afin d'enregistrer des informations de respiration.

1.5.6 Acquisition de données

1.5.6.1 Logiciels

Le système d'acquisition doit être capable d'enregistrer de façon synchrone l'ensemble des données et en particulier les images échographiques et la vidéo à 60 Hz. Une plateforme d'acquisition permettant à la fois l'acquisition et la visualisation des données en temps réel

des données a été développée en utilisant le logiciel RTMaps, de la société Intempora Inc. Les données peuvent être enregistrées localement ou bien transmises sur un réseau. La taille des images ultrasonores est de 320x240 pixels et les images des lèvres sont de taille 640x480 pixels. Une carte son USB (AudioBox44VSL) à quatre entrées permet l'acquisition synchrone de l'EGG, du microphone, de l'accéléromètre et de la ceinture de respiration. La sortie de la carte son est interfacée avec le reste du système d'acquisition. Les quatre entrées analogiques de la carte sont échantillonnées à 44100 Hz avec un encodage sur 16 bits.

1.5.6.2 Synchronisation des capteurs

Il est extrêmement important de s'assurer que les données acquises par l'ensemble des capteurs sont synchrones. Pour ce faire, nous utilisons un test de synchronie avec un événement que l'ensemble des capteurs enregistre [92]. Une seringue emplie de gel utilisé pour les échographies est coincée sur un poids à ressort. Lorsque le ressort se détend, la seringue éjecte une goutte de gel qui atteint la partie supérieure de la sonde ultrasonore et court-circuite les deux électrodes de l'EGG. La vibration induite par le corps de la seringue est enregistrée par l'accéléromètre, tandis que le microphone enregistre le bruit du poids frappant la seringue. Enfin, la caméra est placée de façon à filmer la goutte de gel une fois éjecté. Nous déterminons où le gel est détecté pour chaque capteur et calculons ainsi l'éventuel délai d'acquisition.

1.5.6.3 Construction des bases de données

Le système d'enregistrement a été testé dans le cadre du *Human Beat Box*, du *Cantu in Paghjella* et de la musique byzantine. A chaque fois, une séance d'enregistrement a permis à la fois de valider l'utilisation du casque en fonction du style de chant et de définir un protocole d'acquisition de données. Au cours d'une session d'enregistrement, il est important de vérifier la cohérence des signaux observés ainsi que leur synchronie, à l'aide du protocole décrit section 1.5.6.2. La base de musique byzantine collectée comprend des voyelles isolées en voix parlée, en voix chantée et des extraits de chants byzantins dans plusieurs styles. La première base de chants corses comprend des voyelles isolées parlées et chantées, puis des associations CV (consonnes-voyelles) faisant partie des consonnes et des voyelles les plus représentées dans la langue corse et enfin trois chants corses. Le cas du HBB s'est avéré particulièrement difficile à enregistrer, notamment à cause des larges mouvements de la mâchoire du chanteur qui empêchaient la sonde de rester stable au cours de l'acquisition.

Ces bases de données nous permettent de construire des modèles articulatoires des techniques de production vocale. L'importance des mouvements de la langue pour la production des sons nous a conduits à considérer en priorité cette modalité.

2 Extraction du contour de langue à partir d'images échographiques

2.1 Introduction

Nous souhaitons utiliser des technologies de l'information et de la communication afin de préserver les techniques de chants rares qui font partie du patrimoine immatériel. Parmi les informations que nous pourrions fournir comme soutien pédagogique pour l'apprentissage de ces techniques, des informations sur les mouvements de la langue pourraient permettre d'améliorer la technique d'articulation d'un chanteur. Puisque les images échographiques sont peu lisibles par une personne non entraînée à la lecture de telles images, il nous a semblé utile d'extraire le contour supérieur de la langue afin de proposer des informations plus directes sur les mouvements de la langue impliqués dans les techniques de chant.

Les images de la langue obtenues par imagerie ultrasonore permettent d'obtenir de façon non invasive une coupe sagittale de la surface de la langue, sur lesquelles le contour supérieur de la langue est apparent. Il est courant de considérer que la position de ce contour supérieur de la langue peut être repérée par les pixels les plus bas de cette zone où les pixels sont très proches du blanc. Cependant, la présence de bruit multiplicatif de type *speckle* (chatoiement) rend la tâche d'extraction de contour délicate. Dans cette section, nous cherchons à extraire le contour de la langue à partir des images échographiques sous forme de coordonnées de points appartenant au contour. En effet, déterminer le contour de la langue manuellement est particulièrement long et incompatible avec l'automatisation du traitement des données. La détermination automatique des contours de la langue sur des images échographiques est une tâche complexe qui nécessite une grande robustesse aux changements de positions de la sonde échographique. Dans de nombreuses situations, une intervention humaine se révèle nécessaire afin de corriger les écarts d'étiquetage. Le contour de la langue peut être directement extrait trame par trame sur chaque image échographique, ou bien suivi d'image en image sur une séquence. Certaines de ces méthodes utilisent des connaissances a priori sur la forme du contour ou la physique des mouvements de la langue [95], [96] et [97], par exemple en imposant un lissage spatial sur le contour ou bien en interdisant des modifications trop abruptes entre deux trames consécutives. Différentes méthodes d'extraction du contour de langue à partir des images échographiques ainsi que leurs intérêts et leurs limites sont discutés section 2.2.

2.2 Méthodes d'extraction du contour de langue à partir d'images échographiques

2.2.1 Méthodes d'extraction du contour

2.2.1.1 Méthodes simples de traitement d'image pour l'extraction du contour

Les méthodes simples de détection du contour de langue sur des images échographiques utilisent généralement les différences d'intensité entre les pixels appartenant au contour de la langue et les pixels en dehors du contour. Ces méthodes utilisent généralement des paramètres comme une région d'intérêt et un seuil de détection qui dépendent du locuteur, de la position de la sonde, voire de la session d'enregistrement. Ces méthodes sont généralement assez simples à mettre en œuvre et peu gourmandes en temps de calcul mais les paramètres de ces algorithmes doivent être ajustés en fonction des images. Un prétraitement des images permet de réduire les images à la zone où se trouve la langue. Un seuillage permet de conserver les pixels d'intensité la plus importante. Ensuite, pour chaque image, un contour est détecté en parcourant l'image colonne par colonne. Pour chaque colonne, plusieurs pixels candidats sont sélectionnés et c'est soit le voisinage, soit le passé qui permet de sélectionner le meilleur de ces candidats. En combinant les informations extraites à partir du contour détecté sur une image et celui prédit par les contours des images précédentes, l'estimation de la position du contour de la langue devient assez précise. Un post-traitement consiste à lisser la courbe ainsi obtenue. Cependant, l'utilisation d'interpolation de type polynomiale ou de splines cubiques peut introduire des effets de bord, en particulier au niveau de la pointe de la langue. Il y a donc un risque que les contours obtenus perdent en réalisme et donc ne permettent plus l'identification du mouvement sous-jacent de la langue. En outre, le paramétrage de la région d'intérêt ainsi que du seuil de binarisation sont à ajuster en fonction des images. L'utilisation de ce type d'algorithmes peut devenir délicate pour des images échographiques très bruitées, pour lesquelles certains points du contour ne seront pas détectés ou certains points en dehors du contour seront détectés comme faisant partie du contour. Cette méthode requiert donc souvent des ajustements et des corrections manuels.

2.2.1.2 Une méthode d'apprentissage statistique pour l'extraction du contour

Pour extraire les coordonnées de points appartenant au contour de la langue à partir d'images échographiques, une méthode décrite dans [98] propose d'utiliser des réseaux profonds (*Deep Learning*). Dans cette méthode, l'architecture utilisée est celle de réseaux de neurones de types autoencodeurs, c'est-à-dire des réseaux entraînés pour trouver une représentation des entrées de sorte que l'entrée puisse être reproduite à partir de cette représentation. L'auteur propose de construire une architecture capable d'apprendre la relation entre une image échographique et une image de contour associée et ainsi de pouvoir construire une image représentant le contour de la langue sur une image échographique. Cette méthode possède l'avantage d'être purement automatique et permet de déterminer le contour de la langue sans connaissances a priori sur le phonème prononcé ou sur le locuteur. C'est une technique qui n'utilise pas non plus les contours précédents pour calculer le contour courant, qui peut donc être calculé sur des images isolées et pas seulement des séquences. Une fois que le réseau de *Deep Learning* est entraîné, déterminer le contour d'une image est simple et rapide. La qualité des contours extraits par cette méthode est équivalente à celle obtenue par étiquetage manuel. Cependant, cette méthode présente le défaut de nécessiter l'étiquetage manuel complet de la base d'apprentissage. On estime à plus de 50 heures le temps nécessaire à un expert pour étiqueter une base d'apprentissage de 10 000 exemples. En parallèle avec les méthodes d'extraction du contour de la langue, des méthodes de suivi de contour sur une séquence d'images, s'appuyant sur l'argument que le contour varie peu entre deux images consécutives, ont été développés.

2.2.2 Méthodes de suivi de contour

2.2.2.1 Méthodes de contours actifs

Un modèle de contour actif est une structure formée de points mobiles répartis sur une courbe en deux dimensions, placée sur une zone d'intérêt d'une image. La courbe se déplace (comme un serpent, d'où la dénomination « snake ») en suivant la forme des objets de l'image. La méthode « Snake » permet de déterminer le contour d'un objet dans une image par minimisation d'une fonction de coût. Cette méthode a été fréquemment employée afin d'extraire le contour de la langue dans des images échographiques, comme présenté dans [99] et [100]. La première étape consiste à concevoir une fonction dont les minima locaux englobent l'ensemble des contours recherchés. Les contours sont recherchés par la

minimisation de cette fonction à partir d'une courbe initiale. La fonction se déplace ainsi comme un serpent (d'où l'appellation « Snake ») vers le contour le plus proche sous l'influence d'un champ de forces créé par le gradient. Cependant, il n'est possible de détecter le contour en utilisant la méthode Snake que si le contour est suffisamment apparent, sans quoi le Snake peut être attiré par un gradient élevé qui n'a aucun rapport avec le contour. L'idée est alors d'utiliser des méthodes fondées sur le mouvement, en supposant que l'image précédente permet d'obtenir des informations sur l'image suivante. Cette méthode est très efficace si le contour est bien visible, mais si le contour disparaît sur certaines images, le contour obtenu peut être erroné et nécessite une intervention humaine afin de réinitialiser le suivi et éviter de propager ces erreurs dans la suite de la séquence. Certaines améliorations ont été proposées, comme le Snake contraint, notamment en utilisant un prétraitement des images afin d'améliorer la qualité du suivi du contour en augmentant la visibilité du contour dans les images d'entrée [101]. Le contour est initialisé pour chaque trame en utilisant les informations du flux optique entre deux trames consécutives et deux capteurs électromagnétiques collés sur la langue. Cette méthode est plus efficace qu'un Snake classique mais nécessite néanmoins elle aussi des ajustements manuels.

2.2.2.2 Modèles d'apparence active

Les méthodes utilisant des modèles d'apparence active [102], proposent d'utiliser des informations a priori sur la forme de la langue. Ces modèles incorporent des informations sur la géométrie de la langue et permettent d'estimer de façon robuste et précise le contour de la langue sur une séquence d'images, même en cas de mauvaise visibilité du contour sur les images. Il s'agit généralement de connaissances a priori sur les variations de forme et de textures. Dans [103], il s'agit d'extraire le contour sur la partie visible du contour sur les images et d'extrapoler le contour sur les parties non visibles. Le modèle d'intensité des images autour de la zone du premier contour, appelé modèle de texture, est entraîné sur des trames étiquetées manuellement. Un modèle bayésien permet d'estimer les paramètres du modèle pour chaque trame. Cependant, bien que permettant un suivi robuste du contour, ces techniques nécessitent des connaissances a priori sur les variations de forme et de texture de la langue et donc l'utilisation d'autres techniques d'observations de la langue (radiographie, avec les risques et les inconvénients que cela comporte).

2.2.2.3 Suivi robuste du contour

Un algorithme robuste de suivi de contour sur des séquences d'images échographiques est présenté dans [104]. Cet algorithme a la particularité de proposer une meilleure robustesse du suivi du contour de la langue sur de longues durées. Le bruit sur les images échographiques peut causer des discontinuités dans le contour visible à l'échographie. Afin de gérer cette difficulté, cette méthode utilise des contours actifs avec une contrainte de similarité. Afin de compenser l'accumulation d'erreurs de suivi inhérentes au fait de suivre le contour sur une séquence d'image et non de le déterminer trame par trame sans utilisation d'a priori, cette méthode propose l'utilisation de réinitialisation automatique du contour fondée sur un index de similarité. Cette méthode donne de bons résultats sur des séquences d'images ultrasonores, même pour des durées de quelques minutes. Cette réinitialisation automatique remplace les réinitialisations manuelles nécessaires dans la plupart des autres méthodes de suivi de contour [105]. Les résultats démontrent que cette méthode permet d'améliorer la robustesse des contours actifs en cas de segments manquants et permet d'automatiser la réinitialisation du contour afin d'éviter la propagation des erreurs de suivi du contour.

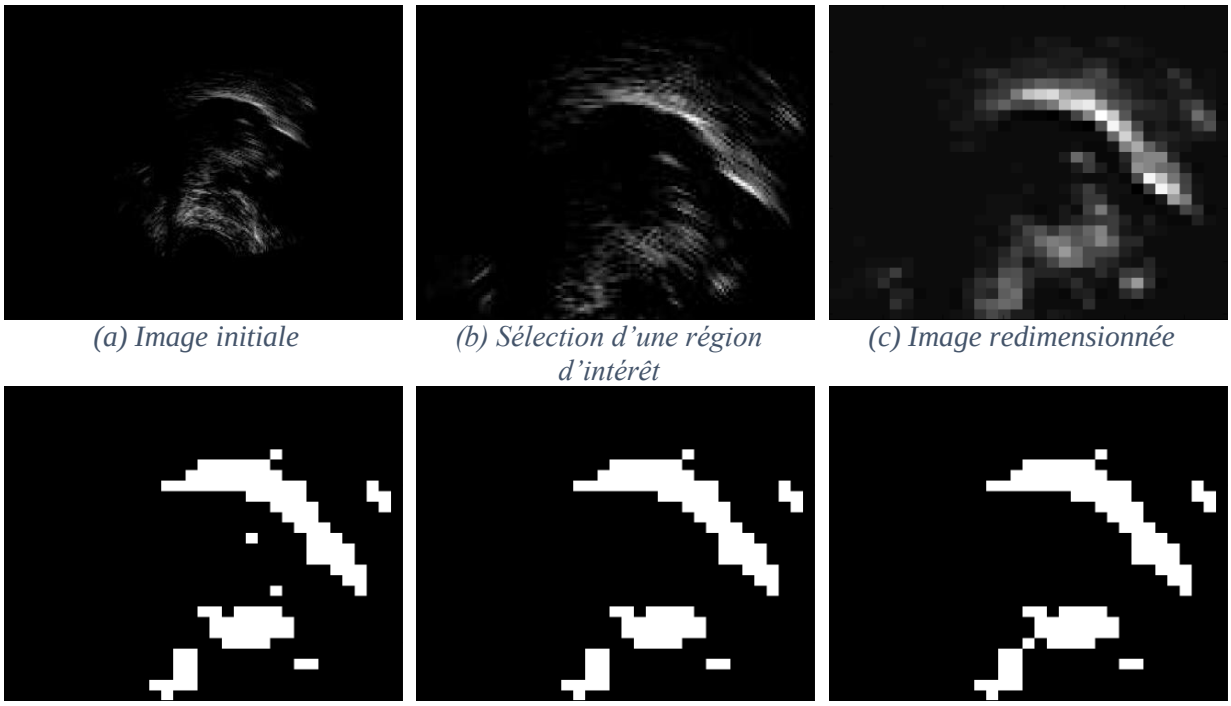
Nous avons choisi d'utiliser une méthode permettant d'extraire le contour des images plutôt que de le suivre sur une séquence. Nous souhaitons en effet éviter le risque de dégradation de la qualité des contours au cours du temps des méthodes présentées en section 2.2.2.1 et 2.2.2.2 et s'affranchir du calcul des CW-SSIM qui rallonge le temps de détermination du contour. Afin d'assurer une plus grande robustesse des performances et de tirer profit de la quantité de données dont nous disposons, nous avons choisi un modèle de réseaux profonds. Pour cela, quelques prétraitements sont nécessaires.

2.3 Prétraitement des images échographiques

2.3.1 Traitement des images échographiques

La méthode de *Deep Learning* utilise des réseaux multicouches avec plusieurs couches cachées, l'intérêt d'utiliser plusieurs couches cachées étant d'obtenir des réseaux efficaces grâce à la combinaison des différents niveaux de non-linéarités. Un réseau de neurones profond possède une topologie beaucoup plus complexe qu'un perceptron multicouche et chaque couche possède ses propres hyper-paramètres. De plus, la taille des images ultrasonores que nous voulons utiliser comme entrées du réseau est relativement importante (240 x 320 pixels). Le nombre de neurones dans les couches cachées étant généralement du

même ordre de grandeur que la dimension des entrées, il est nécessaire de réduire la taille des entrées. En outre, nous n'utilisons dans notre algorithme que des images binaires. En effet, les machines de Boltzmann ont initialement été conçues pour des problèmes binaires et, bien qu'il existe des modèles de DBN ayant pour entrées des images avec des valeurs continues entre 0 et 1, il nous a semblé légitime d'utiliser des images binarisées pour détecter des contours (donc des seuils d'intensité) sur nos images. Cependant, la réduction de dimensions engendre une perte de qualité des images, dont nous essayons de limiter l'impact (voir Figure 32). La figure ci-dessous décrit les différentes opérations que nous effectuons de façon automatique sur les images avant de les utiliser comme entrées du réseau.



(a) Image initiale *(b) Sélection d'une région d'intérêt* *(c) Image redimensionnée*

(d) Image binarisée *(e) Suppression des points isolés* *(f) Connexion entre pixels voisins*

Figure 32 – Prétraitements effectués sur les images échographiques afin de réduire la taille des entrées. Pour une image initiale (a) de taille 240x320 pixels, nous sélectionnons une région d'intérêt de 100x170 pixels comme montré en (b). Ensuite, l'image est redimensionnée en une image de 30x33 pixels. L'image est ensuite binarisée comme montré en (d). Ensuite, les points isolés, considérés comme du bruit, sont supprimés comme montré sur la figure (e). Finalement, afin d'éviter les sauts dans l'image à cause de la binarisation, les pixels voisins sont reconnectés entre eux comme montré en (f). Ces images sont ensuite représentées comme des vecteurs ligne.

2.3.2 Utilisation d'un contour initial pour l'apprentissage

L'application de notre méthode nécessite deux étapes [106] :

- Une étape d'apprentissage au cours de laquelle le réseau est entraîné pour reproduire exactement ce qu'il reçoit en entrée. Cette entrée est constituée d'une image ultrasonore traitée comme montré sur la Figure 32 et d'une image binaire représentant seulement le contour de la langue.

- Une seconde étape au cours de laquelle le réseau d'apprendre à reconstruire le contour d'une image à partir de l'image elle-même et non plus la concaténation des vecteurs représentant l'image d'entrée et l'image de son contour.

Si nous utilisons ce même réseau entraîné à la fois sur des images de contour et des images échographiques, il n'est pas évident que le réseau soit capable de produire une image de contour si on ne lui en fournit pas en entrée. La méthode décrite par [98] propose d'estimer les contours à partir de l'image échographique seule, en s'appuyant sur le fait que la représentation apprise par le réseau entraîné sur les deux types d'images contient la relation entre ces deux types de données. Ainsi le décodeur est capable, à partir de cette représentation cachée, de reconstruire à la fois l'image échographique et l'image de contour. L'hypothèse est donc que si l'on parvient à construire un encodeur capable de créer un codage caché similaire à celui fourni par l'encodeur précédent mais à partir des images échographiques seules, alors le décodeur sera capable de le décoder et reconstruire les deux types d'entrée. Cet encodeur est obtenu de façon « translatée » par rapport à l'encodeur d'origine : le premier RBM est remplacé par un tRBM (*translational RBM*). Ainsi, si le réseau a correctement appris durant la phase d'apprentissage, il devra être capable d'attribuer à chaque image ultrasonore réduite et binarisée un contour qui correspond à la forme de la langue.

2.3.3 Outil d'extraction du contour initial

Pour notre base d'apprentissage, nous utilisons un contour initial extrait de façon automatique à l'aide d'un algorithme de traitement d'images qui localise et prédit la surface de la langue sur chaque image ultrasonore. Cet algorithme comprend à la fois une détection et une prédiction de la position de la langue. Chaque image échographique est prétraitée afin de centrer la détection du contour sur la partie de l'image qui contient l'information pertinente. Il

est important de noter que d'un locuteur à l'autre, voire d'une session d'enregistrement à une autre, les amplitudes de mouvement ainsi que les régions d'intérêt peuvent être différentes. Ces différences justifient la nécessité d'ajuster les seuils de prétraitement de l'outil d'extraction du contour initial en fonction des données. Sur les images prétraitées (voir Figure 33), la détection du contour de chaque image est faite colonne par colonne, de gauche à droite. Pour chaque colonne, de haut en bas, chaque pixel blanc suivi d'un pixel noir est considéré comme candidat à l'appartenance au contour. Un exemple de situation dans laquelle plusieurs pixels sont candidats est donné Figure 34. Un seul point par colonne est sélectionné, donc une décision est prise pour savoir quel point candidat appartient au contour. Le long d'une colonne donnée, si plusieurs pixels blancs sont suivis d'un pixel noir, la sélection du meilleur candidat est faite en comparant l'image courante à l'image précédente (voir Figure 35). Ceci suppose que si un point se trouvait dans le contour précédent, il est conservé. Si aucun point du contour précédent ne correspond à un des pixels candidats, le meilleur candidat est déduit des candidats des colonnes précédentes (voir Figure 36). Nous récupérons grâce à cet algorithme un ensemble de coordonnées $(\boldsymbol{x}, \boldsymbol{y})$ pour chaque image qui correspondent à cette surface.

Figure 33 – Exemple d'image utilisée en entrée de l'algorithme automatique de détection de contours dont la sortie est utilisée comme contour initial. Cette image est obtenue après découpage de l'image, seuillage et filtrage. La sélection des régions recherchées pour l'extraction du contour utilise des connaissances a priori sur la physiologie de la langue.

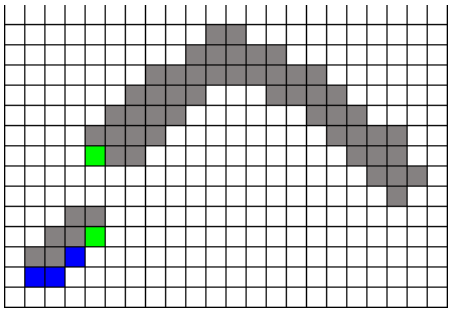


Figure 34 – Un exemple de cas où plusieurs pixels (en vert) sont candidats à l'appartenance au contour (pixels bleus). Par la suite, l'image courante sera nommée i.

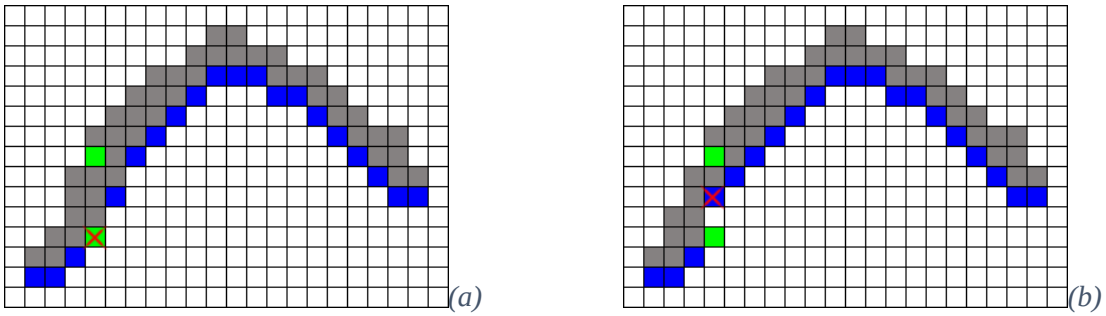


Figure 35 – Deux exemples d'images précédentes possibles pour l'image i montrée Figure 34. Sur la colonne qui nous intéresse, les pixels candidats pour l'image i sont affichés en vert sur les images i – 1. Le pixel marqué d'une croix rouge est le pixel appartenant au contour de l'image i – 1 pour la colonne qui nous intéresse. Dans le cas présenté en (a), une décision peut être prise grâce au contour de l'image i – 1 car un des pixels candidats pour le contour de l'image i appartenait au contour de l'image i – 1. En revanche, dans le cas proposé en (b), aucun des pixels candidats en i ne faisait partie du contour de l'image i – 1. D'autres critères sont alors pris en compte pour la décision.

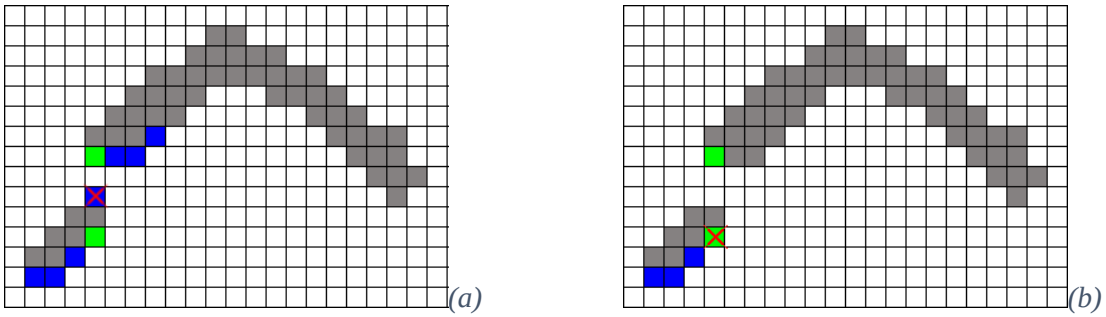
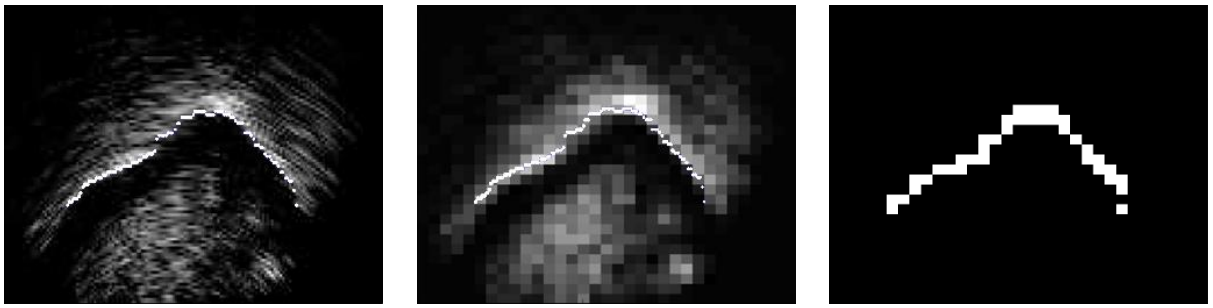


Figure 36 – Sélection du pixel appartenant au contour de l'image i si plusieurs pixels sont candidats mais qu'aucun d'entre eux n'appartient au contour de l'image précédente i – 1. Dans le cas (a), le pixel choisi comme appartenant au contour de l'image i, marqué d'une croix rouge, est prédit par rapport à la position des pixels précédents du contour de l'image i (en bleu) par régression linéaire. Il n'appartient pas à la sélection des pixels candidats, marqués en vert. Dans le cas (b), le pixel choisi comme appartenant au contour de l'image i, marqué d'une croix rouge, est choisi comme étant le pixel candidat (en vert) le plus proche du pixel sélectionné pour la colonne précédente. En pratique, la décision est faite comme montré en (a) sauf si une régression linéaire n'est pas possible. Dans ce cas, la décision est prise comme en (b).

Les coordonnées des contours utilisés comme contour initial sont ensuite converties en images binaires comme décrit à la Figure 37 pour pouvoir être utilisées par le réseau de neurones. Nous utilisons en entrée de notre autoencodeur des coordonnées de contours extraits de façon automatique. Ces contours sont ensuite convertis en images, comme décrit à la Figure 37. Pour chaque exemple, chaque pixel de l'image échographique ainsi que chaque pixel de l'image de contours est représenté par une entrée du réseau.



(a) Contour initial *(b) Contour redimensionné* *(c) Image de contour binarisée.*

Figure 37 – Conversion des coordonnées des contours en images binaires. La première image (a) montre les coordonnées des contours utilisés comme étiquetage pour la base d'apprentissage. Ces contours correspondent à une région d'intérêt de 100x170 pixels obtenus à partir de l'algorithme automatique. Ensuite, les coordonnées des contours sont sous-échantillonnés pour correspondre au changement d'échelle (30x33 pixels) et affichés figure (b). Enfin, l'image (c) est une image de taille 30x33 pixels où la valeur 1 a été affectée aux pixels appartenant au contour défini par les coordonnées de la figure (b).

2.4 Utilisation d'un autoencodeur profond pour l'extraction automatique du contour de la langue

2.4.1 Description de la phase d'apprentissage

Lors de la phase d'apprentissage, nous utilisons un autoencodeur constitué d'un encodeur à 3 couches cachées et d'un décodeur symétrique. Chaque exemple de la base d'apprentissage est représenté en entrée du réseau par un vecteur contenant les intensités des deux images binarisées et réduites. Un schéma de la phase d'apprentissage montrant le type d'entrées et de sorties est donné Figure 38.

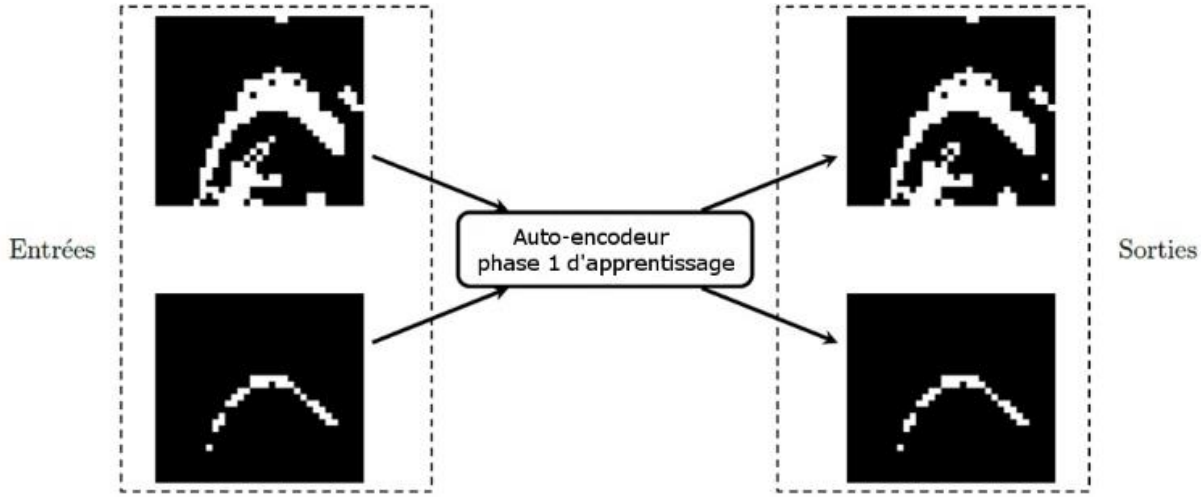


Figure 38 – Exemple d'entrées et de sorties de l'autoencodeur au cours de la première phase d'apprentissage.

Avant l'apprentissage du réseau par empilement de RBM successives, les poids du réseau

sont initialisés aléatoirement. L'encodeur est entraîné de couche en couche de manière gloutonne (*greedy*) par empilement de RBM. Les entrées jouent tout d'abord le rôle d'unités visibles et la première couche correspond aux unités cachées d'une première RBM, puis une seconde machine de Boltzmann est entraînée en prenant la première couche comme couche visible et la deuxième comme couche cachée et ainsi de suite. Chaque RBM est ainsi entraînée à partir des représentations cachées de la RBM précédente. La partie décodeur, quant à elle, est obtenue en propageant de manière symétrique les poids de l'encodeur.

2.4.2 Reconstruction du contour à partir de l'image ultrasonore seule

Une fois que le réseau est entraîné et que sa structure est validée, on peut l'appliquer à des données provenant d'une base de test indépendante. Cependant, nous souhaitons obtenir le contour de la langue d'une image échographique sans donner d'autre entrée au réseau que cette image. Une illustration du réseau est donnée Figure 39 :

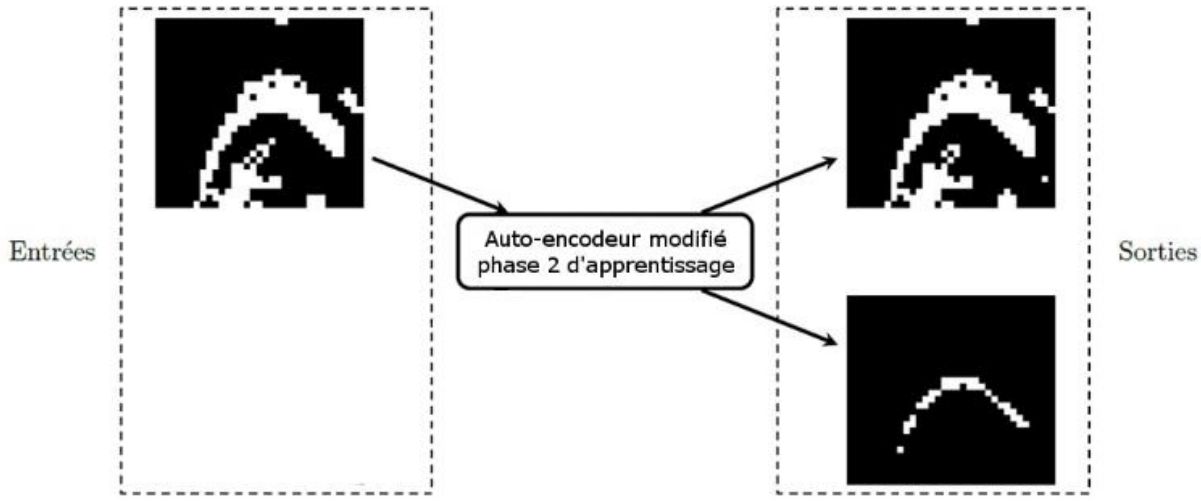


Figure 39 – Exemple d'entrées et de sorties de l'autoencodeur modifié utilisé pour la deuxième phase de l'apprentissage.

Un encodeur entraîné sur la concaténation des données échographiques et le contour apprend la relation qui existe entre les données de sorte que le décodeur est capable de reconstruire l'entrée à partir de la représentation cachée extraite des données. Si l'on construit un nouvel encodeur entraîné seulement sur les images échographiques et capable de produire la même représentation cachée que l'encodeur initial, alors il est possible d'utiliser le décodeur du réseau initial pour reconstruire les contours de la langue à partir de la représentation cachée apprise par le nouvel encodeur. Cet encodeur modifié, développé par [98], porte le nom de

translational DBN. Tout d'abord, un premier autoencodeur est entraîné à reconstruire image et contour à partir de l'image et du contour en entrée. Ensuite, on créée une RBM qui prend en entrée l'image seule et qui a le même nombre d'unités cachées que la première couche du réseau précédent. L'apprentissage est réalisé à l'aide de la divergence contrastive, mais le calcul des probabilités $p(h_i|\boldsymbol{v})$ est construit avec les paramètres du premier réseau tandis que les $p(v_i|\boldsymbol{h})$, $p(h_i|\widehat{\boldsymbol{v}})$et $p\left(\hat{v}_i\middle|\widehat{\boldsymbol{h}}\right)$ sont calculés à partir du second, d'où l'idée de translation. Une fois ce réseau entraîné, on peut l'utiliser sur des données qui n'ont pas été apprises. Notons $\boldsymbol{x}$ l'entrée d'un autoencodeur, h_i la $i^{\text{ème}}$ couche cachée et $f_{\varphi i}$ la fonction d'encodage de la couche i pour un ensemble de paramètres φ donné. Les grandeurs surmontées d'un $\widehat{\ }$ désignent les grandeurs relatives au décodeur. Soit θ un autre ensemble de paramètres, et soit $g_{\theta i}$ la fonction d'encodage associée. Un schéma du réseau initial et de celui modifié (translaté) sont donnés Figure 40.

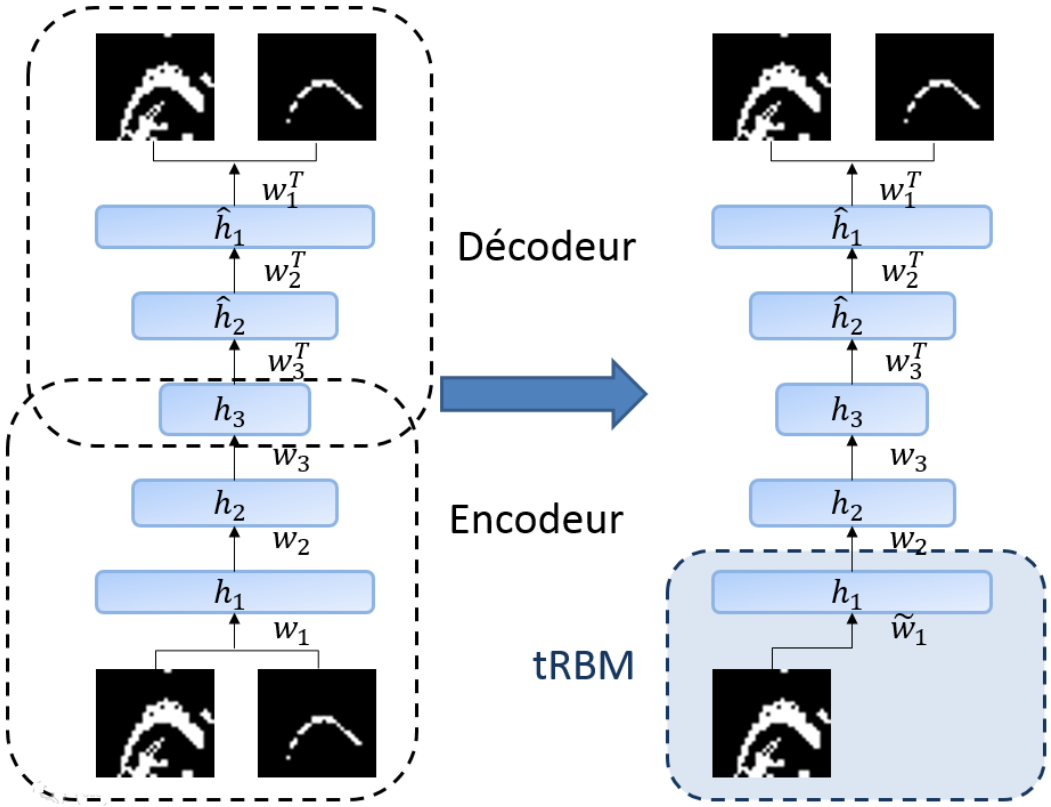


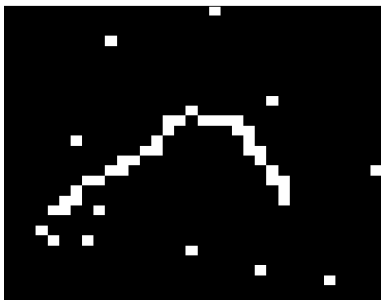
Figure 40 – Autoencodeur d'origine (à gauche) et autoencodeur modifié (à droite) pour extraire automatiquement les contours de langue à partir d'une image échographique prétraitée. La partie inférieure désigne l'encodeur, alors que la partie supérieure désigne le décodeur.

L'utilisation d'un RBM « translationnels » et de RBM empilés en autoencodeurs a un lien avec les autoencodeurs débruitants décrits au chapitre 3. En effet, il s'agit d'imposer une contrainte au réseau sur la représentation à extraire des données en corrompant (ici, en supprimant) une partie de l'information. Les résultats d'extraction du contour sont détaillés

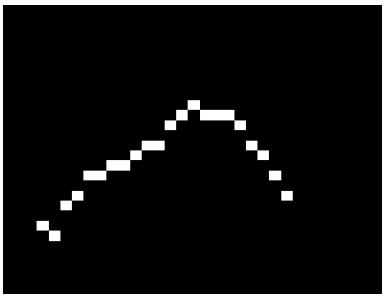
section 2.5. Les images de contour reconstruites doivent ensuite être converties en coordonnées de points appartenant au contour de la langue.

2.4.3 Conversion des images de sortie en contours

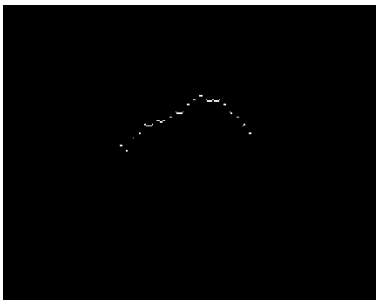
La sortie de notre réseau correspond pour chaque exemple à deux images : la reconstruction de l'image échographique réduite ainsi que la reconstruction d'une image de contour. Nous souhaitons convertir ces images de contours en une liste de coordonnées $(\boldsymbol{x}, \boldsymbol{y})$ pour chaque image. Il est possible de convertir les images de contours binaires de taille 30x33 pixels en une série de coordonnées correspondant aux images échographiques initiales, de taille 240x320 (voir Figure 41). En pratique, nous observons que les images de sortie représentant les contours (en bas à droite sur la Figure 39) sont assez fidèles en termes de forme mais quelques pixels sont erronés. Cela se traduit soit par des pixels isolés supplémentaires (situés hors de la région du contour) ou à l'inverse des discontinuités dans le contour. Pour améliorer la qualité de ces images, nous effectuons des traitements simples, identiques à ceux effectués en entrée après la binarisation, à savoir la suppression des pixels isolés et la reconnexion entre deux pixels adjacents. Une fois ces traitement effectués, nous nous assurons qu'il y ait au plus un pixel par colonne. Si ce n'est pas le cas, nous conservons celui qui a le plus de voisins appartenant au contour. Ensuite, nous redimensionnons l'image de contours de 30x33 pixels afin qu'elle atteigne la taille d'origine de 240x320 pixels par interpolation bicubique. Chaque pixel blanc de cette image reconstituée définit un couple de coordonnées $(\boldsymbol{x}, \boldsymbol{y})$. Afin que les contours obtenus soient plus réalistes, un dernier traitement est effectué : un lissage de la courbe de coordonnées. Il s'agit d'une régression locale utilisant un modèle polynomial du second degré. Par ailleurs, dans cette régression, un poids plus faible est affecté aux points aberrants (trop éloignés du reste de l'ensemble des points). Cela permet d'obtenir un contour continu sans rencontrer les effets de bord liés à une
interpolation polynomiale classique.



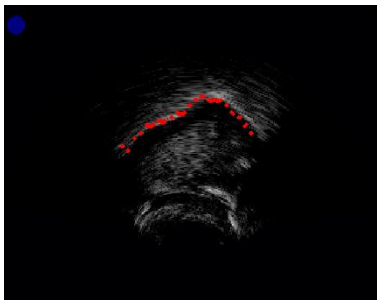
(a) Sortie de l'autoencodeur



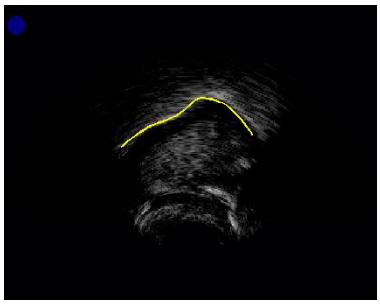
(b) Figure nettoyée



(c) Redimensionnement



(d) Conversion en coordonnées



(e) Lissage.

Figure 41 – Les différentes étapes du post-traitement effectué pour la conversion des images de contours en coordonnées. L'image (a) montre une figure obtenue en sortie de l'autoencodeur, de 30x33 pixels. La figure (b) montre cette sortie nettoyée après différents traitements. Après redimensionnement et mise à l'échelle, nous obtenons la figure (c) de 240x320 pixels. Les pixels sont ensuite convertis en coordonnées de points par rapport à l'image échographique d'origine (d). Enfin, la figure (e) montre le contour obtenu après lissage.

Cette méthode nous permet d'extraire le contour de la langue à partir d'images échographiques provenant d'une base de test indépendante. Nous évaluons ensuite la qualité de ces contours reconstruits en les comparants à des contours extraits manuellement ainsi qu'aux résultats fournis par d'autres auteurs.

2.5 Méthodes pour l'évaluation des résultats de reconstruction du contour

2.5.1 Critères d'évaluation

L'évaluation de la qualité de la reconstruction du contour de la langue nécessite de trouver des critères d'évaluation. Il est utile de comparer un contour à un contour de référence. De façon générale, un contour de langue extrait de façon satisfaisante est une ligne qui suit de façon physiologiquement réaliste le bord inférieur [85] de la courbe brillante représentant le contour de la langue sur l'image échographique. Il est important d'extraire toute la surface de la langue visible sur l'image échographique sans ajouter d'artefacts [85]. Afin d'évaluer la qualité des contours obtenus par la méthode de *Deep Learning*, nous avons entraîné le réseau

sur une base de 17000 images puis sélectionné de façon aléatoire 150 autres images échographiques issues de la même session d'enregistrement pour tester l'extraction du contour. Nous avons dessiné manuellement le contour sur ces 150 images et avons comparé les coordonnées de ces contours obtenus manuellement avec ceux obtenus par la méthode de *Deep Learning*. Puis nous avons comparé, sur ces mêmes 150 images, les coordonnées obtenues par la méthode de *Deep Learning* à ceux que nous avons utilisés comme contour initial pour l'apprentissage, afin d'obtenir une évaluation de la fidélité de la reconstruction à l'entrée (voir exemple Figure 42).

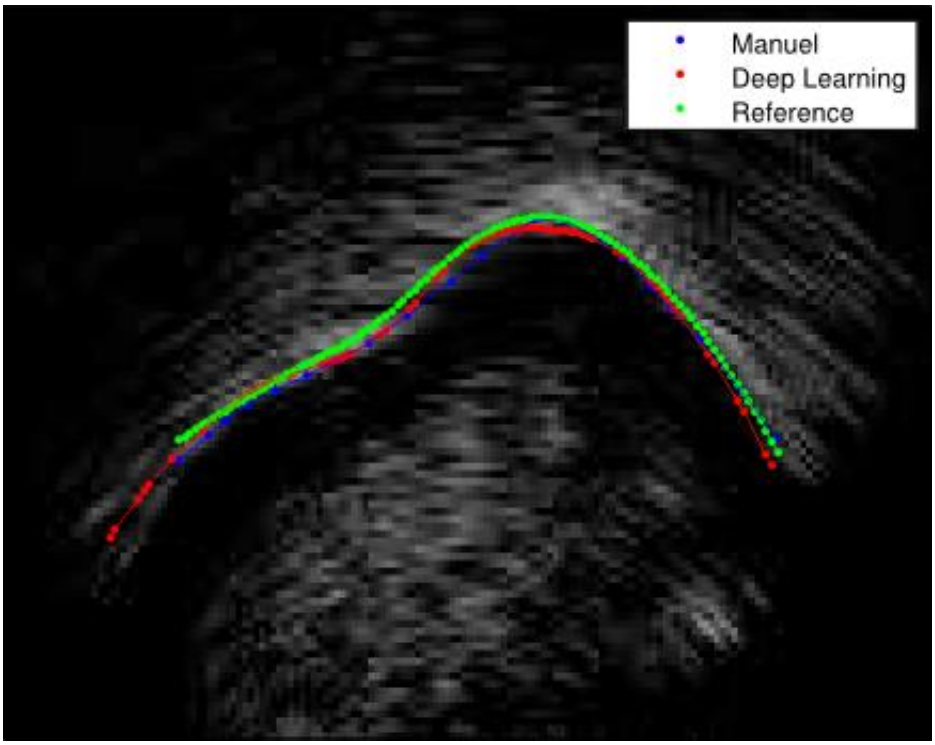


Figure 42 – Comparaison entre une courbe de contour extrait manuellement (en bleu), l'algorithme de Deep Learning *(en rouge) et le contour initial (référence) pour l'apprentissage.*

Cependant, les coordonnées des contours ainsi obtenus ne possédant pas systématiquement le même nombre de points et les points qui définissent un des contours n'ayant pas les mêmes abscisses que les points décrivant l'autre contour, la comparaison ne peut pas être faite directement. Dans [100], une métrique est proposée afin de comparer chaque pixel d'une courbe donnée au pixel le plus proche (en termes de distance L_1) sur la courbe avec laquelle est faite la comparaison. Cette métrique, nommée *Mean Sum of Distances* (MSD), permet d'évaluer en pixels la distance moyenne d'une courbe à une autre, même si les points qui décrivent l'une des deux courbes n'ont pas les mêmes abscisses que les points qui décrivent l'autre courbe. Soit un contour $\boldsymbol{U}$ composé d'un ensemble de points définis par leurs

coordonnées 2D $(u_1, \dots, u_n)$ et un contour $\boldsymbol{V}$ composé d'un ensemble de points définis par $(v_1, \dots, v_m)$. Le MSD est défini comme suit :

$$MSD(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{\mathrm{n} + \mathrm{m}} \left(\sum_{i=1}^{m} \min_j |v_i - u_j| + \sum_{i=1}^{n} \min_j |u_i - v_j| \right) \qquad (37)$$

Comme le montre la Figure 43, les abscisses des points de ces courbes ne sont pas les mêmes. Ces comparaisons permettent d'évaluer la qualité du contour reconstruit en fonction de l'architecture choisie, selon la base de données.

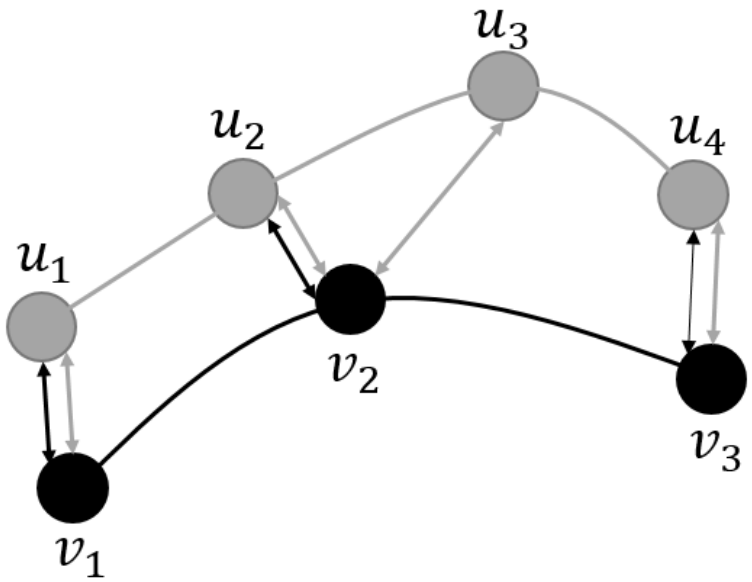


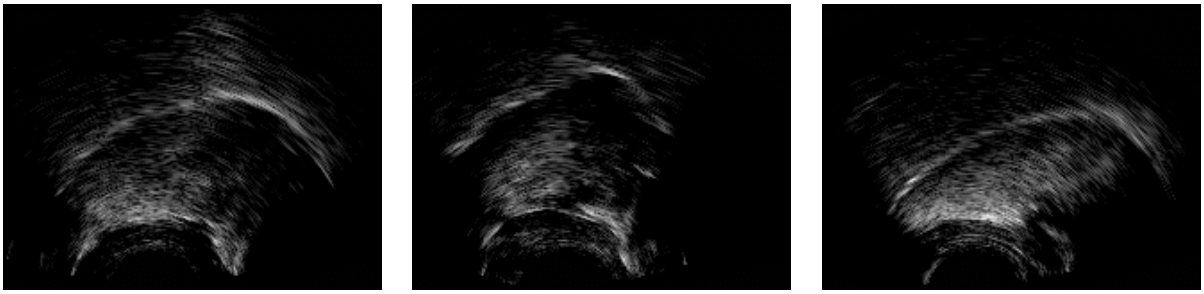
Figure 43 – Représentation simplifiée de deux sous-parties de deux contours. Les quatre grandeurs $\boldsymbol{u_1}$, $\boldsymbol{u_2}$, $\boldsymbol{u_3}$ et $\boldsymbol{u_4}$ représentent les coordonnées (x, y) de trois points adjacents du contour gris. De même, les trois grandeurs $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ et $\boldsymbol{v_3}$ représentent les coordonnées (x, y) de trois points adjacents du contour noir. La comparaison de deux courbes de contour en utilisant le MSD permet de comparer ces contours même s'ils n'ont pas le même nombre de points.

2.5.2 Base de données et applications

Nous disposons de données comprenant des enregistrements audio, vidéo et échographiques de trois locuteurs, deux hommes et une femme, en langue française (voir Figure 44). Chaque locuteur a enregistré entre 40 et 60 listes, comprenant chacune 50 phrases [107]. Une phrase compte environ entre 300 et 500 images échographiques. À la différence de [98], nous n'utilisons pas des phonèmes isolés mais des phrases complètes, ce qui augmente la diversité des formes de langue proposées.

Les premiers tests que nous avons effectués [108] ne concernaient qu'une locutrice, pour laquelle nous avons choisi un ensemble de 50 phrases, desquelles sont issues les 17 000 images échographiques de la base d'apprentissage et validation et les 150 images de test présentées section 2.5.1.

Nous avons ensuite cherché à évaluer les performances de notre algorithme d'extraction de contours à base de *Deep Learning* en testant ses capacités à extraire des informations à partir de données échographiques de plusieurs locuteurs différents. Pour cela, nous avons sélectionné aléatoirement 16 000 images provenant d'une base constituée de 50 phrases (c'est-à-dire entre 15 000 et 20 000 images par locuteur) prononcées par trois locuteurs. Les performances de notre méthode dépendent largement du choix des hyperparamètres de l'architecture. Cependant, la qualité des images, largement dépendante du locuteur, et en particulier la présence ou non d'ombres dans le contour, s'est avérée particulièrement critique.



(a) *(b)* *(c)*

Figure 44 – Exemples d'images échographiques provenant de trois locuteurs différents. La figure (a) ainsi que la figure (c) correspondent à des locuteurs, tandis que la figure (b) correspond à une locutrice. Sur ces images, nous pouvons voir que chaque locuteur a une forme de langue différente. De plus, d'un locuteur à l'autre, les amplitudes de mouvement ainsi que les régions d'intérêt sont différentes. Ces différences rendent impossible l'utilisation directe de notre outil d'extraction du contour initial pour l'apprentissage, qui nécessite une calibration pour chaque locuteur.

2.6 Choix de l'architecture de l'autoencodeur

Chaque exemple de la base d'apprentissage est présenté au réseau comme un vecteur contenant les intensités normalisées des deux images binarisées, soit 1980 unités pour les 1980 pixels plus une pour introduire un biais. Plusieurs jeux d'hyperparamètres ont été testés, à savoir le nombre de couches cachées, le nombre d'unités par couche, le nombre de passes d'apprentissage (le nombre de fois où les données de la base d'apprentissage sont présentées au réseau) et la taille des mini-batches, qui sont des sous-ensembles des données d'apprentissage, regroupant généralement entre 10 et 100 exemples chacun. Nous avons fondé notre choix sur l'erreur de reconstruction en base de validation, définie comme la moyenne

quadratique de la différence entre les composantes x_i des N vecteurs d'entrée de dimensions $1 \times I$ et les composantes $\widehat{x_i}$ de leurs N vecteurs reconstruits de mêmes dimensions.

$$E_{reco} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_i^n - \widehat{x_i}^n)^2} \quad (38)$$

Ces tests ont été effectués sur une base de données de 17 000 exemples, dont nous avons utilisé 15 000 images en apprentissage et 2000 en validation.

2.6.1 Profondeur du réseau

L'empilement de RBMs permet d'augmenter le niveau d'abstraction du modèle. Cependant, il est important de trouver la profondeur la plus pertinente pour extraire des descripteurs utiles. Pour cela, nous avons testé plusieurs architectures avec des profondeurs différentes. Dans nos tests, nous avons considéré une architecture avec 1000 unités par couche, 50 itérations du procédé d'apprentissage et des mini-batches de taille 1000 et testé les performances de structures avec respectivement 2, 3 et 4 couches cachées. L'erreur de validation la plus basse correspond à une structure à 3 couches cachées, comme présenté Tableau 3.

Tableau 3 - Influence du nombre de couches cachées sur l'erreur de validation.

Nombre de couches cachées	Erreur de validation
2	0.39
3	0.38
4	0.44

2.6.2 Complexité du réseau

Dans les modèles d'apprentissage statistique classiques, il est important d'avoir plus d'exemples d'apprentissage que de paramètres du modèle afin d'éviter le surajustement. Cependant, dans les architectures profondes, il est courant d'utiliser de nombreuses unités en couche cachée [72]. Nous avons donc testé un nombre variable d'unités par couche pour un modèle à trois couches cachées, comme montré Tableau 4. Dans le cadre de nos travaux, nous avons choisi le nombre d'unités caches en cherchant à minimiser l'erreur de validation tout en limitant le temps de calcul.

Tableau 4 - Influence du nombre d'unités par couche sur l'erreur de validation.

Nombre d'unités caches par couche	Erreur de validation
500	0.41
1000	0.38
2000	0.37

2.6.3 Taille des mini-batches

L'utilisation de mini-batches, qui sont des regroupements de la base d'apprentissage en petits sous-ensembles, permet de diminuer les durées d'apprentissage car la mise à jour des poids n'intervient qu'une fois par mini-batch et non pour chaque exemple. Cependant, trouver une taille optimale de mini-batches n'est pas immédiat. D'après [79], il convient de découper la base d'apprentissage en sous-ensembles de 10 à 100 exemples. L'idée défendue dans [79] est que la taille des mini-batches dépend du nombre de classes présentes dans la base d'apprentissage. Dans notre cas, puisqu'il ne s'agit pas d'un problème de classification, il est impossible d'estimer le nombre de « classes » différentes de la base d'apprentissage. Nous avons donc comparé l'erreur de validation pour différentes tailles de mini-batches, allant de 10 à 200 exemples.

Tableau 5 - Influence de la taille des mini-batches sur l'erreur de validation.

Taille des mini-batches	Erreur de validation
10	0.65
50	0.53
100	0.38
200	0.40

Les résultats montrent que pour un réseau à 3 couches cachées avec 1000 unités par couches, 50 itérations et des mini-batches de taille 10, l'erreur atteint 0.65, puis elle diminue à 0,38 pour des mini-batches de taille 100 et augmente au-delà.

2.6.4 Nombre d'itérations

Nous avons utilisé une procédure semblable afin de tester le nombre d'itérations de la boucle d'apprentissage (mise à jour des poids) nécessaires afin d'obtenir une erreur de validation la

plus basse possible. Il est important de conserver un nombre d'itérations suffisamment bas afin de garder un temps de calcul raisonnable mais tout en atteignant des performances satisfaisantes. Nous avons utilisé un réseau à 3 couches cachées avec 1000 unités par couche et des mini-batches de taille 100. Nous avons testé un apprentissage avec 5, 50 et 250 itérations. On observe que le fait d'utiliser un nombre trop élevé d'itérations dégrade les performances en base de validation (phénomène de surapprentissage) et augmente considérablement le temps de calcul. Le compromis temps de calcul-performances semble bien respecté en limitant le nombre d'itérations à 50.

Tableau 6 - Influence du nombre d'itérations sur l'erreur de validation.

Nombre d'itérations	Erreur de validation
5	0.41
50	0.38
250	0.40

2.7 Qualité du contour reconstruit

Nous avons finalement choisi une structure avec un encodeur à trois couches cachées suivi d'un décodeur symétrique avec 2000 unités cachées par couche, des mini-batches de 100 exemples et 50 itérations. En plus de la comparaison entre les coordonnées des contours issus du *Deep Learning* et d'une part les coordonnées des contours utilisés comme entrée, d'autre part les coordonnées des contours extraits manuellement, nous avons voulu comparer les coordonnées des contours utilisés comme entrée et les coordonnées des contours extraits manuellement, afin de mieux interpréter les résultats (voir Tableau 7). Comme dans la section précédente, nos résultats portent sur un seul locuteur. Quelques exemples de contours extraits par la méthode de *Deep Learning* sont donnés Figure 45.

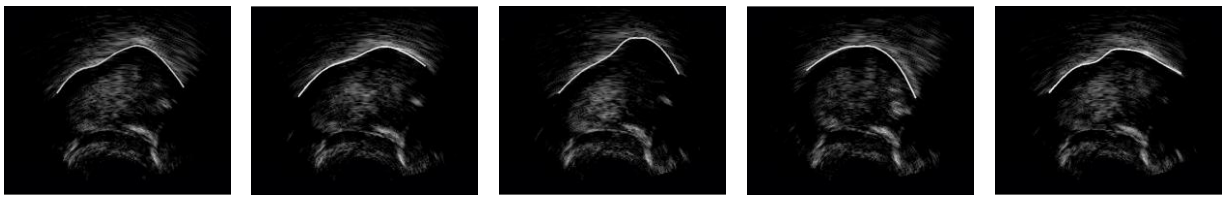


Figure 45 - Quelques exemples de contours extraits en utilisant notre autoencodeur profond.

Les résultats de comparaisons entre les coordonnées des contours extraits en utilisant le *Deep Learning* (DL), les contours initiaux utilisés comme référence (Ref) pour l'apprentissage,

détaillé section 2.5 et l'étiquetage manuel (man). Les résultats sont donnés dans le Tableau 7. Ils montrent que les contours obtenus avec les méthodes DL, Ref et Man sont comparables. L'autoencodeur, bien qu'il ne traite qu'une image à la fois, est capable d'atteindre des résultats comparables à ceux de l'algorithme utilisé comme référence, qui utilise les informations temporelles pour extraire le contour de langue sur une séquence. Ce constat suggère que l'architecture de *Deep Learning* a intégré des informations provenant des contraintes de continuité imposées dans l'algorithme de référence.

Nous avons souhaité comparer les résultats obtenus par notre méthode à ceux de la littérature. Dans [100], le contour fourni par EdgeTrak, qui utilise la méthode Snake, est comparé au contour fourni par deux experts différents. Cependant, afin de comparer les valeurs de MSD, exprimées en pixels, indépendamment des résolutions des images, nous avons converti ces valeurs en millimètres. Dans [100], les 67 images comparées (issues d'une séquence d'enregistrement sur un locuteur) étaient de taille 112,9 x 89,67 mm. La comparaison entre un expert 1 et un expert 2 donne un MSD de 0.85 mm (2.9 pixels avec l'équivalence 1px = 0.295 mm), la comparaison entre l'expert 1 et EdgeTrak donne un MSD de 0.67 mm, tandis que la comparaison entre l'expert 2 et EdgeTrak donne un MSD de 0,86 mm. Dans [98], qui utilise un autoencodeur avec étiquetage manuel de la base d'apprentissage, après 5 validations croisées, le MSD moyen calculé sur 8640 images est de 0.73 mm. Les valeurs de MSD obtenues dans nos conditions expérimentales, calculées avec l'équivalence 1 px = 0,35 mm, données Tableau 7, sont assez proches de ces valeurs. Ceci nous permet de conclure que les résultats obtenus par notre méthode d'autoencodeur profond entraîné sont de qualité voisine aux valeurs rapportées par d'autres auteurs, tout en permettant une plus grande automatisation du processus d'extraction de contour. Comme dans [98], il est nécessaire de prétraiter les images pour en réduire la dimension, il est également nécessaire de les post-traiter afin de convertir les images en listes de coordonnées de contours. Cependant la méthodologie que nous avons employée diffère de [98] dans la mesure où nous utilisons une extraction de contour semi-automatique pour la base d'apprentissage et non une extraction manuelle. Par ailleurs, nous travaillons avec des données différentes ce qui rend la comparaison entre les différentes méthodes difficile.

Tableau 7 – Valeurs moyennes du MSD comparant les contours provenant de l'étiquetage manuel (Manuel) à ceux utilisés comme référence (Ref), les contours provenant de l'étiquetage manuel à ceux issus du Deep Learning *(DL) et les contours utilisés comme référence à ceux issus du* Deep Learning.

	MSD Moyen (mm)
Ref vs. Manuel	0,9
Ref. vs. DL	0,8
Manuel vs. DL	1,0

Les valeurs les plus faibles de MSD concernent la comparaison entre les courbes issues du *Deep Learning* et celles utilisées comme référence. Par ailleurs, les valeurs obtenues en comparant les contours obtenus manuellement aux deux autres contours sont du même ordre de grandeur. Les performances des deux méthodes semblent donc similaires. Nos valeurs d'erreur semblent correspondre aux différences que l'on peut trouver entre deux étiquetages manuels. Par ailleurs, il est à noter que les valeurs les plus faibles sont obtenues pour la comparaison entre les coordonnées issues du *Deep Learning* et celles utilisées comme référence pour l'apprentissage, ce qui témoigne d'un apprentissage performant.

Nous avons ensuite voulu appliquer la même méthode d'extraction de contours à des données de trois locuteurs différents. Pour cela, nous avons réalisé un apprentissage sur 15 000 images tirées aléatoirement parmi les 50 phrases sélectionnées par locuteur. Si les données de chaque locuteur sont prises séparément (apprentissage et test sur un seul locuteur), le MSD moyen (calculé pour 5 phrases, soit 5500 images de test) entre les coordonnées utilisées comme contour intial pour l'apprentissage et celles issues du *Deep Learning* est proche d'1 mm, quel que soit le locuteur. Néanmoins, lorsque l'on sélectionne le même nombre d'images mais en mélangeant aléatoirement les données des trois locuteurs, le score de MSD moyen passe à 1,9 mm. Cette dégradation de la qualité des contours reconstruits peut être due à une augmentation de la complexité de la tâche demandée, puisque les formes de contour de langue sont beaucoup plus variées et que certaines images sont assez pauvres (présence d'ombres sur le contour). Une augmentation à la fois de la taille de la base d'apprentissage et de la qualité des images enregistrées permettrait d'augmenter les performances en multi-locuteur. Il paraît également cohérent d'augmenter la taille des mini-batches afin de tenir compte de la plus grande variabilité des images.

2.8 Discussion

L'utilisation d'un autoencodeur profond afin d'extraire automatiquement le contour de la langue d'une image échographique semble donner des résultats prometteurs. En effet, lors des tests sur un locuteur, les erreurs de reconstruction sont proches des différences entre deux contours extraits manuellement. Elles sont également comparables aux erreurs de reconstruction rapportées par certains auteurs. Ces résultats justifient l'utilisation de contours extraits de façon automatique comme contour intial au lieu d'extraire manuellement le contour de la langue dans de larges bases de données, ce qui est très fastidieux. De plus, si notre algorithme obtient des performances similaires à celles obtenues à l'aide d'un outil utilisant des informations temporelles, nous pouvons considérer que notre réseau est capable d'apprendre de nouvelles contraintes extraites à partir des entrées, sans utiliser directement d'informations temporelles.

Au cours de notre travail, nous avons choisi, ajusté et validé la structure sur une base de validation. Pour un travail futur, fournir à l'algorithme des bases d'apprentissage très variées, composées de phrases, mots ou phonèmes prononcés par plusieurs locuteurs et dans des modalités (parole ou chant) différentes permettrait de tester la robustesse de l'algorithme aux changements de conditions expérimentales. Si ces résultats s'avèrent concluants, il est envisageable de généraliser cette méthode à l'étude des mouvements des lèvres grâce aux captures vidéo réalisées pendant les interprétations des chanteurs. En effet, le type d'algorithmes présenté ici pourrait être utilisé pour extraire le contour des lèvres à partir de quelques points saillants sur les images. En outre, cette méthode d'extraction de caractéristiques pourrait aussi servir pour des tâches de classification afin de reconnaître les phonèmes prononcés. En effet, l'utilisation des seuls descripteurs issus des mouvements de la langue ne suffirait a priori pas à distinguer des phonèmes entre eux et les informations issus des lèvres pourraient apporter un complément d'information.

Nous avons donc réussi à obtenir une extraction de contour de performance équivalente à l'état de l'art, qui s'affranchit en même temps de la nécessité d'étiqueter des points manuellement [108]. Toutefois la complexité des calculs nécessaire pour mettre en œuvre la méthode, ainsi qu'un manque de stabilité quant aux contours extraits, selon le locuteur et les conditions expérimentales, ne nous ont pas permis d'arriver à un outil bien adapté à une

interface temps réel dans un contexte d'apprentissage de l’articulation. Plutôt de poursuivre dans cette approche, nous avons opté pour une méthode permettant de fournir à l'apprenti un outil qui accède directement aux paramètres acoustique du chant, sans passer par l'étape d'extraction du contour, et qui incorpore également les lèvres du chanteur dans son analyse. Après avoir considéré les séquences d’images échographiques seules, nous avons souhaité extraire des informations en utilisant la combinaison des images de la langue et des lèvres, afin d’obtenir un modèle articulatoire.

3 Synthèse vocale à partir des mouvements des articulateurs

3.1 Introduction

Les images des mouvements de la langue et des lèvres nous permettent d'accéder à des informations articulatoires relatives à la réalisation du geste vocal. Dans ce chapitre, nous cherchons à déterminer comment les mouvements des articulateurs influencent la modification du son émis par la source glottique en agissant sur le filtre du conduit vocal. Nous proposons donc d'utiliser les informations articulatoires dont nous disposons afin de synthétiser des extraits de voix chantée. Cette synthèse nous permettrait de vérifier la corrélation entre le geste articulatoire et la production acoustique,en proposant un cadre général pour l'étude des relations acoustico-articulatoires en voix chantée. .

Notre objectif est d'étudier l'influence des articulateurs sur la production du son. La tâche consiste donc à synthétiser de la voix chantée à partir d'images montrant les mouvements de la langue et des lèvres. L'imagerie des gestes articulatoires, auxquels nous avons accès grâce aux images ultrasonores et vidéos traduit en images les différentes contraintes acoustiques appliquées au flux d'air après passage par les plis vocaux [109]. Le signal acoustique peut donc être vu comme la représentation en termes de mise en vibration de l'air de ces mouvements des articulateurs. Nous cherchons à piloter un synthétiseur vocal (modélisé comme l'association d'une source et d'un filtre), à partir de descripteurs des mouvements de la langue et des lèvres (captés par un échographe et une caméra). Une approche directe [110] consiste à chercher à associer l'espace visuel à l'espace acoustique en cherchant des correspondances. Ces correspondances peuvent être établies par des fonctions de transformation permettant de passer d'un espace à un autre. Ce type de conversion n'utilise pas des descripteurs issus d'a priori linguistiques, acoustiques ou phonétiques. Le signal acoustique est synthétisé en utilisant un modèle source-filtre (voir section 1.3.1.3). Ce modèle implique de connaître à la fois les coefficients du filtre du conduit vocal permettant la synthèse mais aussi le signal d'excitation, que nous déduisons du signal électroglottographique.

Cependant la conversion des mouvements articulatoires en signal acoustique peut aussi être moins directe [110]. La parole peut être décrite à différents niveaux, qu'ils soient acoustiques,

phonétiques et phonologiques, lexicaux, syntaxiques et sémantiques. Ces niveaux permettent de découper le signal en unités élémentaires qui ont un sens. Ainsi, chaque image représente une unité élémentaire visuelle, tandis que du point de vue acoustique, une unité élémentaire sera une trame. Le domaine de la phonétique travaille au niveau du phonème. Synthétiser de la voix parlée ou chantée à partir des seules images des mouvements de la langue et des lèvres est une tâche difficile, puisqu'il manque de l'information, notamment en ce qui concerne l'activité laryngée, les mouvements du voile du palais et la nasalité. Il est donc intéressant d'intégrer des informations supplémentaires dans le modèle [111]. On peut par exemple faire correspondre les informations visuelles à des informations phonétiques au lieu de les faire correspondre directement à des informations acoustiques [112]. Il y a donc une étape de décodage visuo-phonétique [113], qui consiste à identifier dans une séquence d'images la séquence de phonèmes la plus probable (voir [114] et [115]). Cette approche permet d'introduire des connaissances linguistiques a priori sur la séquence observée. De plus, des connaissances linguistiques et syntaxiques permettent l'ajout de contraintes sur le vocabulaire autorisé. Cette approche peut être réalisée à l'aide de modèles de Markov cachés. Ensuite, la synthèse peut être effectuée par concaténation de signaux élémentaires pré-enregistrés ou à l'aide de chaînes de Markov cachées. Nous choisirons une approche directe pour la robustesse de la synthèse proposée aux articulations imprécises. En effet, le type de son étudié étant le chant et en particulier le chant corse, il y a des hypo-articulations (production de parole avec un effort moindre, rendant imprécis les gestes articulatoires) qui rendent difficile la reconnaissance de phonèmes. Dans [20], l'auteur relève en effet dans les voyelles chantées du *Cantu in Paghjella* une confusion entre les voyelles /i/ et /e/ d'une part, et /u/ et /o/ d'autre part.

Dans notre approche, nous testons d'abord comment la position de la langue ou celle des lèvres modifie le son produit. Pour cela, nous utilisons uniquement les informations visuelles et cherchons à les convertir directement en signal acoustique en fixant au préalable la durée ainsi que la hauteur du son, ainsi que des paramètres de qualité vocale permettant d'influer sur le timbre de la voix produite. Dans cette approche, les informations visuelles ne suffisent pas à déterminer si le son est voisé ou non. Par défaut, nous considérons que l'ensemble de ces sons sont voisés. Cette conversion nous permet d'écouter et mettre en évidence l'influence d'une variation de l'un ou l'autre des articulateurs sur le son produit.

Une fois cette phase de conversion mise en place, nous pouvons effectuer une synthèse vocale plus complète sur des chants entiers. Pour cela, nous avons besoin d'informations sur l'activité glottique et en particulier la hauteur du son, les instants de silence et de voisement, ainsi que la qualité vocale du chanteur.

Dans cette partie, nous utilisons les coefficients LSF (*Line Spectral Frequencies*), dérivés des coefficients LPC (voir section 1.3.1.4), pour modéliser le conduit vocal. Afin de chercher la correspondance entre l'espace des articulations acoustiques et l'espace des coefficients du filtre du conduit vocal, nous utilisons un réseau de neurones de type perceptron multicouche (voir [68]). La fonction d'excitation dépend de la présence ou non de voisement de la trame à synthétiser. Les paramètres de la fonction d'excitation sont extraits directement du signal EGG.

Nous décrirons dans les sections 3.3.1 et 3.3.2 les deux méthodes d'extraction de descripteurs que nous avons utilisées : une méthode fondée sur l'analyse en composantes principales permettant de décrire l'ensemble des images dans l'espace des *EigenLips* et *EigenTongues* et une méthode utilisant un autoencodeur profond. Pour cela, il est d'abord nécessaire d'aligner temporellement les images et les coefficients LSF à prédire.

3.2 Calcul des variables à prédire : prétraitements du signal acoustique

3.2.1 Ordre de prédiction LPC

Nous souhaitons prédire une valeur de LSF (voir section 1.3.1.4) pour chaque image échographique et optique. Pour cela, nous calculons les valeurs des LSF à prédire durant la phase d'apprentissage supervisé. Ces LSF de référence sont calculés à l'aide du signal acoustique. Ce dernier est échantillonné à 44,100 kHz. Cette fréquence d'échantillonnage est adaptée à la richesse spectrale de la voix chantée et au contenu musical (par opposition aux enregistrements de parole pure). D'après [116], l'ordre de prédiction LPC doit être proportionnel au taux d'échantillonnage. Dans [89], le taux d'échantillonnage des enregistrements de voix parlée est de 11,025 kHz et l'ordre de prédiction LPC est de 12. Pour une fréquence d'échantillonnage de 44,100 kHz, il semble adapté d'utiliser un ordre de prédiction LPC de 48 pour une application en voix chantée. Afin d'illustrer la nécessité

d’adapter l’ordre de prédiction LPC, nous avons comparé les spectres LPC et FFT d’une fenêtre de signal pour différents ordres de prédiction LPC : 12 et 48 (Figure 46 et Figure 47). L’ordre de prédiction LPC de 12 semble insuffisant pour décrire le contenu fréquentiel de nos enregistrements de voix chantée.

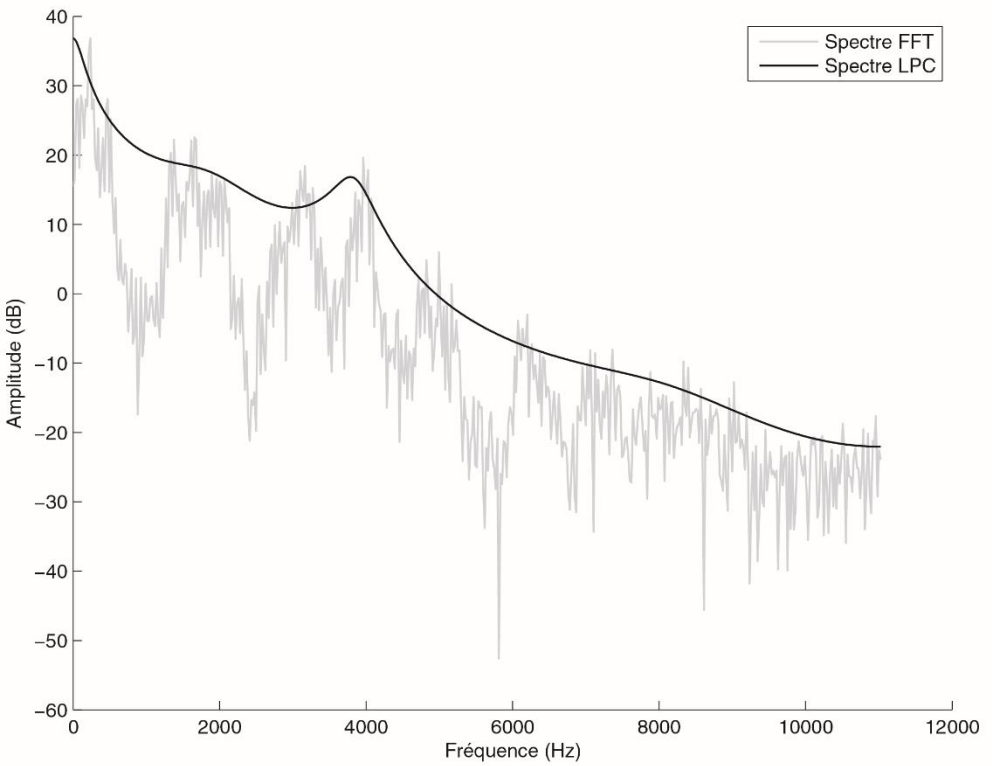


Figure 46 - Spectre FFT (en gris) et enveloppe LPC (en noir) calculée sur une trame de signal en utilisant un ordre de prédiction LPC 12.

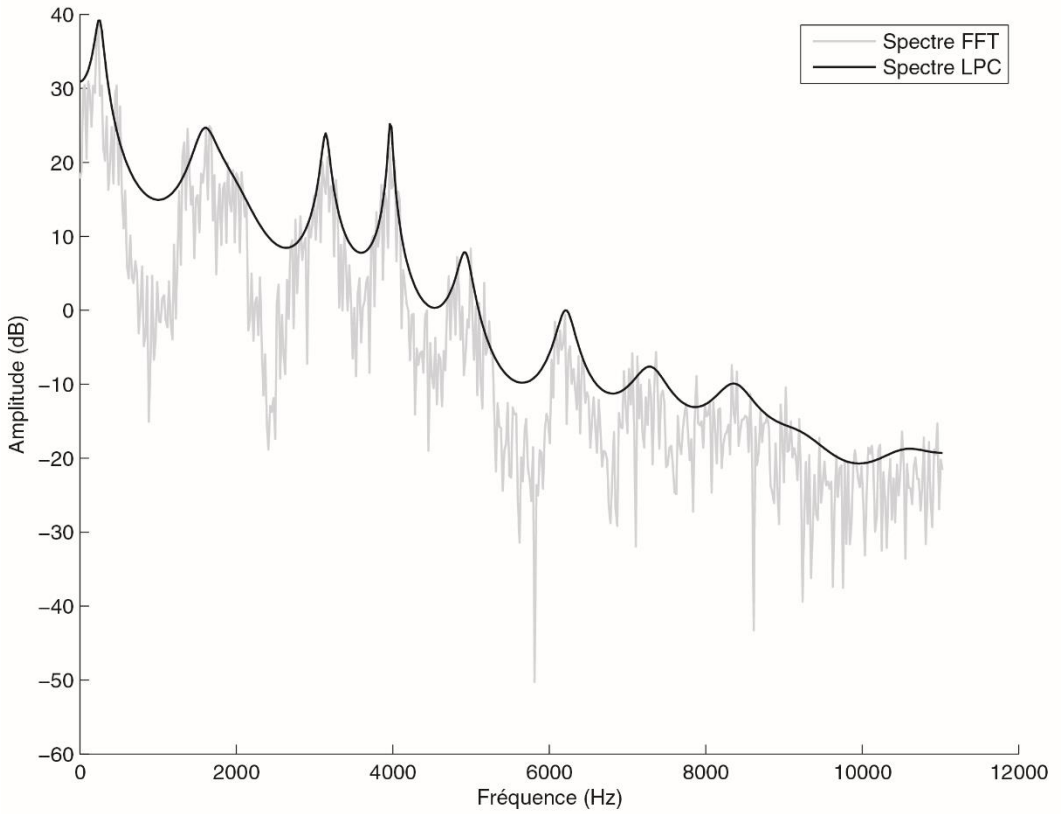


Figure 47 - Spectre FFT (en gris) et enveloppe LPC (en noir) calculée sur une trame de signal en utilisant un ordre de prédiction LPC 48.

Utiliser un ordre de prédiction aussi élevé implique une augmentation importante du temps de calcul, chaque LSF étant prédit séparément. Nous avons donc choisi, pour cette étude, de sous-échantillonner le signal acoustique d'un facteur 4, après utilisation d'un filtre anti-repliement (filtre passe-bas de type Tchebychev d'ordre 8).

3.2.2 Calcul des LSF de référence à partir du signal acoustique

afin d'accentuer sélectivement les hautes fréquences du signal acoustique et détecter les formants à hautes fréquences [116], nous avons dans un premier temps utilisé un filtre de préaccentuation de coefficient $\alpha = 0{,}95$ sur le signal acoustique complet avant d'effectuer le codage LPC. Il s'agit d'un filtre passe-haut du premier ordre à réponse impulsionnelle finie, défini par la relation :

$$y[n] = x[n] - \alpha \cdot x[n-1] \tag{39}$$

Afin d'éviter les distorsions, il convient d'inverser ce filtrage avant d'effectuer la synthèse. Après préaccentuation du signal acoustique, nous avons appliqué un fenêtrage du signal en utilisant des fenêtres de Hamming de largeur 33,3 ms avec un recouvrement de 50 %. Ainsi,

nous avons découpé le signal acoustique et récupéré une trame toutes les 16,7 ms, de façon à disposer d'un ensemble de 12 LSF, d'une image de langue et d'une image de lèvres toutes les 16,7 ms. Nous pouvons noter que la durée des fenêtres permet a priori d'utiliser l'approximation de stationnarité du signal acoustique [117].

3.2.3 Détection du voisement

Le modèle de source étant dépendant du voisement, nous utilisons un modèle de source pour les sons voisés et un modèle différent pour les sons non voisés [118]. Pour cela, il est nécessaire de déterminer pour chaque trame si elle est voisée ou non. Nous avons utilisé une méthode simple reposant sur l'autocorrélation de chaque trame [119]. Un signal voisé étant pseudo-périodique, l'autocorrélation d'une trame voisée est autocorrélation nettement supérieure à celle d'une trame non voisée ou d'un silence. Un seuillage très bas sur la valeur de cette autocorrélation permet de discriminer les trames qui ne sont pas voisées, en acceptant le risque de détecter comme voisée une trame non voisée. L'utilisation de l'autocorrélation permet également de détecter la fréquence fondamentale d'un son, par évaluation de la pseudo-périodicité. Cependant, la périodicité du signal acoustique n'est pas complète et l'identification de voisement peut être biaisée. Par ailleurs, il est difficile de déterminer un seuil de détection à cause des variations des valeurs de fréquence fondamentale et de formants.

3.2.4 Filtrage des LSF

Les LSF calculés à une cadence de 16,7 ms ont la particularité de présenter des discontinuités et des variations rapides qui n'ont pas de rapport avec la réalité acoustique de l'articulation. Dans [120], il est suggéré d'utiliser des fenêtres plus larges, ou bien de lisser les variations de chaque LSF au cours du temps à l'aide d'un filtre passe-bas. Nous ne souhaitons pas augmenter la largeur des fenêtres afin de conserver la synchronisation entre les LSF et la vidéo, donc nous avons choisi de filtrer les signaux de LSF par un filtre de Butterworth du premier ordre et de fréquence de coupure $F_s/4$, où F_s désigne la fréquence d'échantillonnage. La détermination des LSF à partir du signal acoustique nous a permis de déterminer les sorties attendues du modèle pour les images échographiques et optiques.

3.3 Construction de modèles multimodaux de l'articulation

Dans cette partie, nous souhaitons combiner les informations provenant de deux articulateurs que sont la langue et les lèvres. L'objectif est d'extraire des descripteurs à l'aide des observations des mouvements de ces deux types d'articulateurs afin de prédire la valeur des coefficients du filtre du conduit vocal. Nous utilisons dans un premier temps une méthode linéaire fondée sur l'analyse en composantes principales. Dans un second temps, nous cherchons à développer une méthode non linéaire et pour laquelle nous n'imposons pas le type de descripteurs extraits, en utilisant un autoencodeur profond. Pour ce faire, il est possible d'extraire séparément des descripteurs à partir des deux types de modalités, à savoir les images de langue et les images de lèvres. Cependant, comme les phénomènes d'articulations sont assez complexes il peut être préférable d'extraire une description en combinant des informations optiques et échographiques. Ces types d'images étant très différents en termes de répartition de l'intensité des pixels, le fait de leur appliquer les mêmes transformations ne garantit pas un équilibre entre les deux types d'informations.

3.3.1 Une approche linéaire : projection dans l'espace des *EigenLips* et *EigenTongues*

En s'appuyant sur les travaux de [121] décrivant l'utilisation *d'EigenFaces* pour la reconnaissance de visages, les auteurs de [122] ont proposé un modèle utilisant *EigenLips* et *EigenTongues* afin de convertir les informations articulatoires issues de la langue et des lèvres en informations acoustiques. Cette décomposition en *EigenLips* et *EigenTongues* utilise l'analyse en composantes principales (PCA). Soit E un ensemble d'apprentissage comprenant M images de taille $N \times N$. Ces images peuvent être considérées comme des vecteurs ligne de dimension N^2. L'ensemble des images d'apprentissage peuvent être représentées par une matrice A de taille $N^2 \times M$. Soit C la matrice de covariance de A.

$$C = \frac{1}{M}(AA^T) \tag{40}$$

La décomposition en valeurs propres de la matrice de covariance de A s'écrit :

$$R^T C\, R = \Lambda, \tag{41}$$

Où R représente la matrice des vecteurs propres et Λ la matrice des valeurs propres. Les vecteurs propres sont également appelés « composantes principales ». L'espace des vecteurs propres est ordonné selon la direction de variance observée dans E décroissante. La dimension des vecteurs propres est N^2. Il est donc possible de représenter ces composantes principales sous forme d'images, à savoir les EigenTongues et EigenLips. Nous pouvons ensuite projeter une nouvelle image I de taille N^2 dans la base des vecteurs propres et extraire n descripteurs visuels α_k, pour k compris entre 1 et n.

$$\alpha_k = \sum_{i=1}^{N^2} I_i R_{ik} \tag{42}$$

L'hypothèse que les axes de plus grande variance, sur lesquels les données sont plus dispersées, représentent le signal utile implique que l'utilisation des premières composantes principales suffit au codage d'une image.

Une image peut donc être représentée en termes de projection dans l'espace des *EigenTongues*, dont quelques exemples sont présentés Figure 48. Cette méthode permet d'extraire une représentation qui permet de coder (voir Figure 49) certaines informations contenues dans les images comme par exemple la position de la langue, de l'os hyoïde ou de certains muscles.

Figure 48 - Représentations des espaces des EigenLips *et des* EigenTongues. *Sur la ligne du haut, de gauche à droite, les quatre premiers* EigenLips. *Sur la ligne du bas, de gauche à droite, les quatre premiers* EigenTongues.

Figure 49 - Exemples d'images et de leur reconstruction en utilisant les 100 premiers descripteurs. Sur la ligne du haut, une image de lèvres issue de la base de validation (à gauche) et sa reconstruction utilisant les 100 premiers EigenLips *(à droite). Sur la ligne du bas, une image ultrasonore issue de la base de validation (à gauche) et sa reconstruction utilisant les 100 premiers* EigenTongues *(à droite).*

L'analyse en composantes principales permet d'obtenir une compression particulièrement efficace, en déterminant un ensemble de vecteurs propres estimés à partir des exemples de la base d'apprentissage. Les performances de cette représentation dépendent donc de la constitution et de l'équilibre des bases de données.

3.3.2 Une approche non linéaire : Autoencodeurs profonds

L'objectif de notre autoencodeur est d'extraire des descripteurs pertinents pour la prédiction des LSF. Dans une telle architecture, les hyperparamètres de l'autoencodeur sont ajustés de façon à minimiser l'erreur de reconstruction entre la sortie et l'entrée. Cependant, sans imposer de contraintes au réseau, celui-ci ne va pas forcément extraire des descripteurs pertinents pour notre tâche, De plus, les images ultrasonores sont bien plus bruitées que les images des lèvres. En effet, elles sont caractérisées par un chatoiement (*speckle noise*). Une amélioration possible consiste à utiliser un autoencodeur débruitant (*denoising autoencoder*) pour augmenter la robustesse de la sélection de descripteurs. Le principe est d'ajouter artificiellement du bruit sur les images d'entrée et d'apprendre au réseau à reconstruire l'image débruitée. Le réseau apprend ainsi à traiter des images très bruitées. Ce principe est décrit dans [123] et [124]. L'autoencodeur débruitant peut également être utilisé pour imposer des contraintes lors de l'apprentissage pour extraire une nouvelle représentation des données [125]. En effet, au lieu de contraindre la représentation des données par une contrainte de parcimonie par exemple, l'autoencodeur débruitant permet de modifier le critère de reconstruction ; le but d'un autoencodeur débruitant est de nettoyer des données partiellement corrompues. Pour un tel autoencodeur, extraire une bonne représentation des données revient à extraire une représentation que l'on peut obtenir de façon robuste à partir des données

corrompues et qui est utile pour reconstruire les données non corrompues correspondantes. Tout d'abord, les entrées $\boldsymbol{x}$ sont corrompues en $\tilde{\boldsymbol{x}}$. Comme avec un autoencodeur classique, $\tilde{\boldsymbol{x}}$ est associé à la sortie $\boldsymbol{y} = f_\theta(\tilde{\boldsymbol{x}}) = s(W\tilde{\boldsymbol{x}} + b)$. A partir de cette sortie, nous reconstruisons une estimation de l'entrée $\boldsymbol{z} = g_\theta{}'(\boldsymbol{y})$. La Figure 50 illustre le processus. L'erreur de reconstruction est calculée par l'écart entre la sortie $\boldsymbol{z}$ et l'entrée non corrompue $\boldsymbol{x}$ dont $\boldsymbol{z}$ est censée être la plus proche possible.

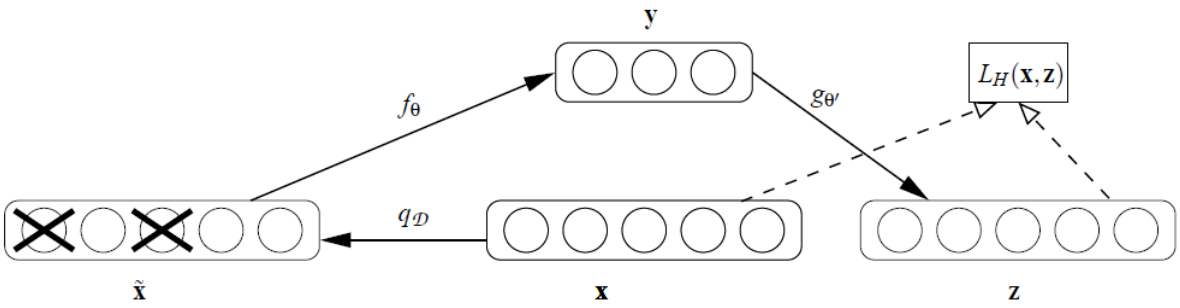


Figure 50 - L'architecture débruitante, d'après [123]. Un exemple $\boldsymbol{x}$ est corrompu en $\tilde{\boldsymbol{x}}$. L'autoencodeur associe $\tilde{\boldsymbol{x}}$ à $\boldsymbol{y}$ via la fonction d'encodage f_θ et vise à reconstruire $\boldsymbol{x}$ via la fonction de décodage g'_θ. La reconstruction $\boldsymbol{z}$ est censée être la plus proche possible de l'entrée non corrompue $\boldsymbol{x}$. Les unités barrées dans $\tilde{\boldsymbol{x}}$ représentent la corruption des données (dans cet exemple par suppression de certaines unités).

Cette approche nous amène à poser deux hypothèses de travail. Premièrement, un plus haut niveau de représentation devrait être plus stable et robuste à la corruption des données. D'autre part, on s'attend à ce que la tâche de débruitage nécessite l'extraction de descripteurs qui capturent une structure utile dans la distribution des données d'entrée. Dans cette approche, ce n'est pas le débruitage en lui-même qui est recherché, c'est l'extraction de descripteurs plus robustes et permettant un meilleur niveau de représentation.

Il est possible d'empiler des étages débruitants de la même façon que l'on empile des RBM pour initialiser un réseau. Il est à noter que la corruption des entrées est utilisée uniquement pour l'apprentissage initial de chaque couche afin d'extraire des descripteurs utiles. Une fois que la fonction f_θ est apprise, elle est appliquée sur des données non corrompues. A fortiori aucune donnée corrompue n'est utilisée pour produire une représentation qui sera utilisée comme donnée d'entrée pour la couche suivante. Une fois qu'un empilement de tels autoencodeurs a été construit, l'étage d'abstraction la plus élevée peut être utilisé comme entrée d'un algorithme supervisé comme les SVM pour une tâche de classification ou les MLP pour une tâche de régression. Dans la suite du manuscrit, nous utilisons des autoencodeurs avec l'étape de débruitage.

3.3.3 Gestion de la multimodalité

Les images de la langue et celles des lèvres sont de natures très différentes. Par conséquent, il paraît assez délicat de simplement les concaténer en entrée d'un modèle d'apprentissage statistique. Un prétraitement séparé sur chacune des modalités semble pertinent. Pour cela, nous utilisons la méthode décrite dans [125] et [126]. Dans ces articles, une méthode de *Deep Learning* est employée afin d'extraire des descripteurs à partir de la combinaison de deux modalités différentes que sont des vidéos des lèvres d'une part et des spectrogrammes audio d'autre part. Cette méthode implique de corréler des informations provenant de multiples sources. Ainsi, nous cherchons à extraire des informations sur le conduit vocal à partir des données articulatoires provenant à la fois de la langue et des lèvres, en utilisant une représentation partagée entre les deux modalités. On pourrait imaginer une structure, présentée Figure 51 qui permettrait d'extraire des descripteurs provenant de chaque modalité prise séparément.

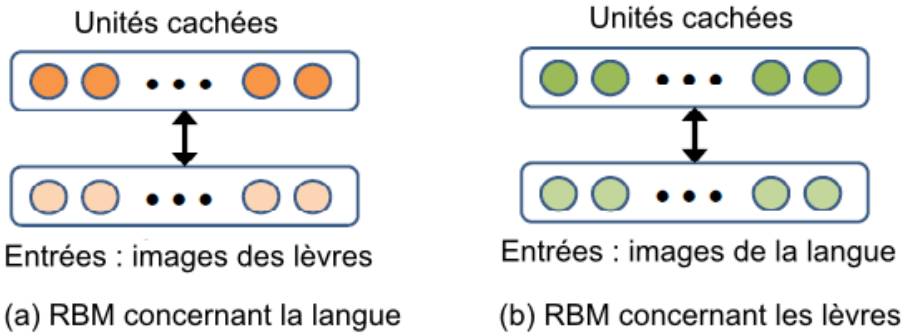


Figure 51 - Les deux RBM permettant d'extraire des descripteurs de la langue et des lèvres utilisés séparément, d'après [125].

Cependant, cette méthode ne permet pas d'extraire des relations entre les deux articulateurs. Une autre méthode possible, illustrée Figure 52, est d'extraire des descripteurs à partir des deux modalités concaténées.

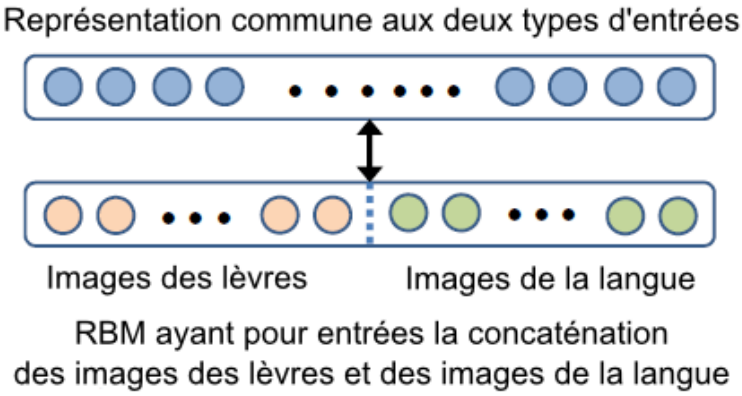


Figure 52 - RBM permettant d'extraire des descripteurs issus de la langue et des lèvres par concaténation des entrées de chaque modalité, d'après [125].

Néanmoins, puisque les relations entre les mouvements de la langue et ceux des lèvres ne sont pas linéaires et que les deux types de modalités sont très différents, il est assez difficile pour un RBM d'extraire des représentations conjointes. Dans [125], cette méthode de simple concaténation des entrées (voir Figure 52) conduit à des unités cachées qui ne semblent pas contenir d'informations communes aux deux modalités. Aussi, une autre méthode d'apprentissage est présentée. Premièrement, chacune des modalités est tout d'abord traitée séparément en étant utilisée comme entrée d'un RBM. Ce RBM permet d'extraire un modèle de chacune des modalités. La représentation des données d'entrée obtenue après ce premier étage facilite l'extraction de variables multimodales. Le premier étage a pour but d'extraire des descripteurs spécifiques à chaque type de données tandis que le second vise à construire un codage issu des deux types d'information. Le schéma d'une telle structure est donné Figure 53.

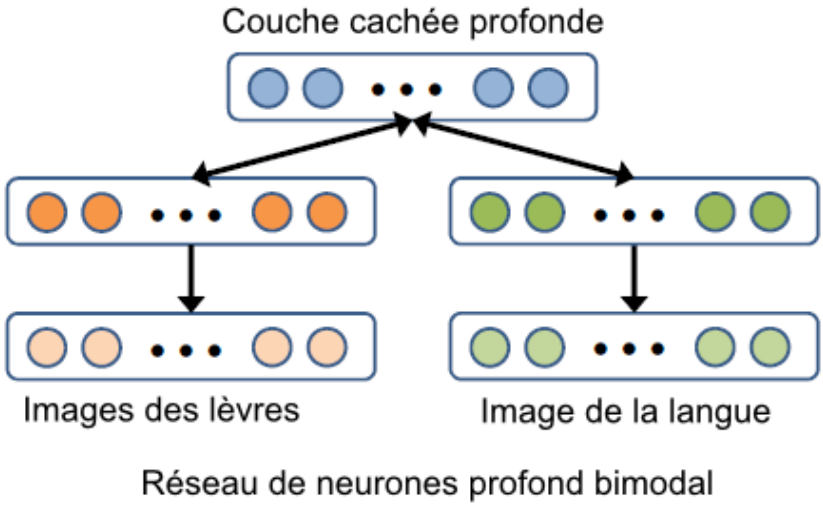


Figure 53 - Exemple de réseau de neurones profond bimodal. Chaque entrée est d'abord traitée séparément à l'aide de RBM séparés puis les couches cachées ainsi extraites servent d'entrée à un RBM dont le but est d'extraire une représentation commune des données, d'après [125].

Cependant, rien ne garantit que le modèle permette de capturer une représentation multimodale. Ainsi, certains descripteurs pourraient être utiles pour modéliser les lèvres seulement et d'autres pour modéliser la langue seulement. Ainsi, l'autoencodeur présenté Figure 54 est entraîné à reconstruire les deux modalités avec seulement l'un ou l'autre.

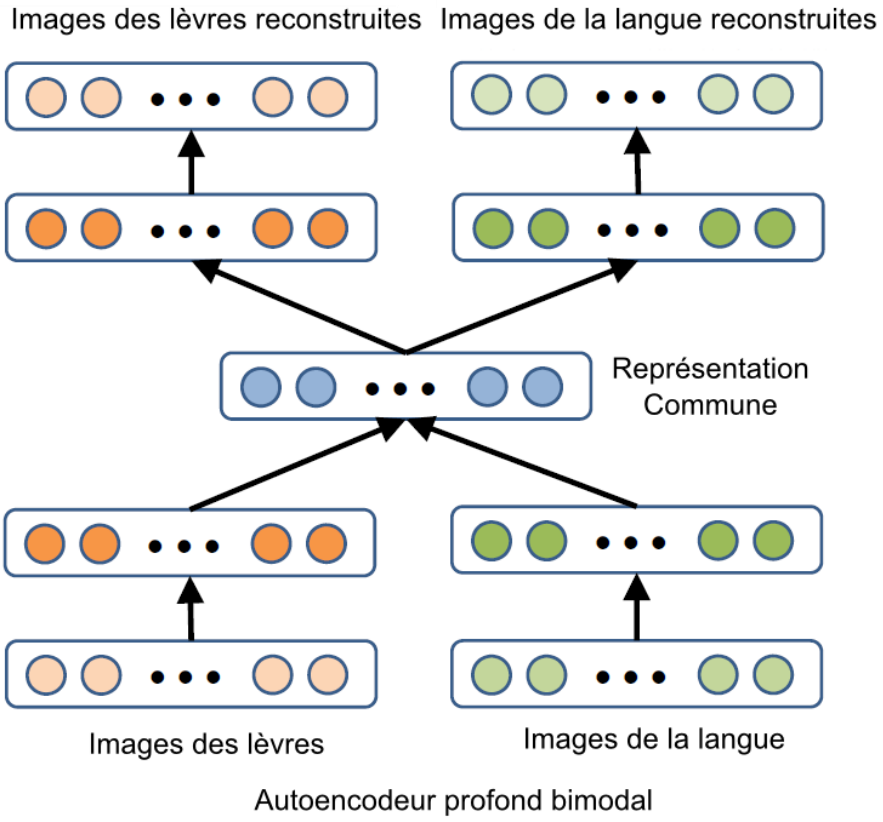


Figure 54 - Un exemple d'autoencodeur profond multimodal permettant d'extraire une représentation conjointe à partir des deux types d'entrées différentes à l'aide d'un premier étage de RBM séparés, d'après [125].

Dans notre architecture, l'autoencodeur est entraîné à l'aide d'une base de données bruitée dans laquelle un tiers des images de langue sont corrompues par du bruit de chatoiement et deux tiers des données d'entrée ne sont pas bruitées. Les images des lèvres, quant à elles, ne sont pas corrompues. Cette architecture a pour but de forcer le réseau à trouver des liens entre les images des lèvres et les images de la langue malgré le bruit.

3.3.4 Sélection de descripteurs

La méthode d'autoencodeur multimodal peut permettre d'extraire de nombreux descripteurs. Afin de sélectionner les plus pertinents pour représenter nos données, nous avons choisi

d'utiliser un algorithme de sélection de variable. Nous utilisons pour cela un algorithme de classement des variables appelé *Orthogonal Forward Regression* (OFR) (voir [127]).

Dans l'espace vectoriel dit des observations, chaque entrée est représentée par un vecteur dont les composantes sont les N observations de cette entrée et chaque sortie est représentée par un vecteur dont les composantes sont les N mesures de cette grandeur. Si l'on utilise un modèle linéaire, la contribution de la $i^{\text{ème}}$ entrée est d'autant plus corrélée à la sortie que l'angle entre le vecteur représentant l'entrée i et le vecteur représentant la sortie est petit. En effet, si cet angle est nul, la sortie est colinéaire à l'entrée et celle-ci explique entièrement la sortie par proportionnalité. À l'inverse, si cet angle est de $\pi/2$, l'entrée et la sortie sont complètement décorrélées pour un modèle linéaire. Les entrées sont ainsi classées par ordre de pertinence en calculant le carré du cosinus de l'angle entre le vecteur de l'entrée i et la sortie, cette valeur devant être la plus proche possible de 1. La procédure d'orthogonalisation suivante permet de classer les entrées par ordre de pertinence décroissante.

Dans un premier temps, on sélectionne la variable la plus corrélée avec la sortie. Ensuite, on projette le vecteur de sortie ainsi que les vecteurs de toutes les autres entrées sur le sous-espace orthogonal à l'entrée sélectionnée. Cette procédure est itérée dans tout le sous-espace, jusqu'à ce que tous les descripteurs soient classés ou en utilisant le critère d'arrêt de la variable sonde.

La seconde étape de la sélection de variables consiste à sélectionner des entrées selon le classement des variables obtenu. Il est important de conserver les variables pertinentes, qui contiennent l'information, et d'éliminer les variables non pertinentes, car celles-ci peuvent dégrader les performances du modèle. L'ensemble des variables candidates est complété par une variable aléatoire, dite variable sonde, par conséquent décorrélée de la sortie. Pour chaque nouveau descripteur candidat, on peut estimer la probabilité de faux positif par la méthode de la variable sonde, et arrêter le classement lorsque cette probabilité devient supérieure au seuil fixé. Ce seuil traduit le risque de conserver des variables qui ne sont pas pertinentes, ou bien d'éliminer des variables en réalité pertinentes. En pratique, nous avons ainsi une méthode permettant de classer les descripteurs selon leur capacité de prédiction de la sortie désirée, à savoir les LSF.

Pour chacun des $k = 12$ coefficients, nous appliquons tout d'abord la procédure d'OFR aux $N_k = 100$ descripteurs déterminés en utilisant l'autoencodeur multimodal ou bien la projection dans l'espace des *EigenLips* et des *EigenTongues*. Nous obtenons ainsi $\widehat{N_k} \leq N_k$ descripteurs pour un risque que nous fixons à 15%. Ensuite, nous conservons le premier tiers de ces descripteurs. Le nombre, $\frac{\widehat{N_k}}{3}$, de descripteurs sélectionnés dépend donc de la qualité de la prédiction des descripteurs en fonction de la tâche.

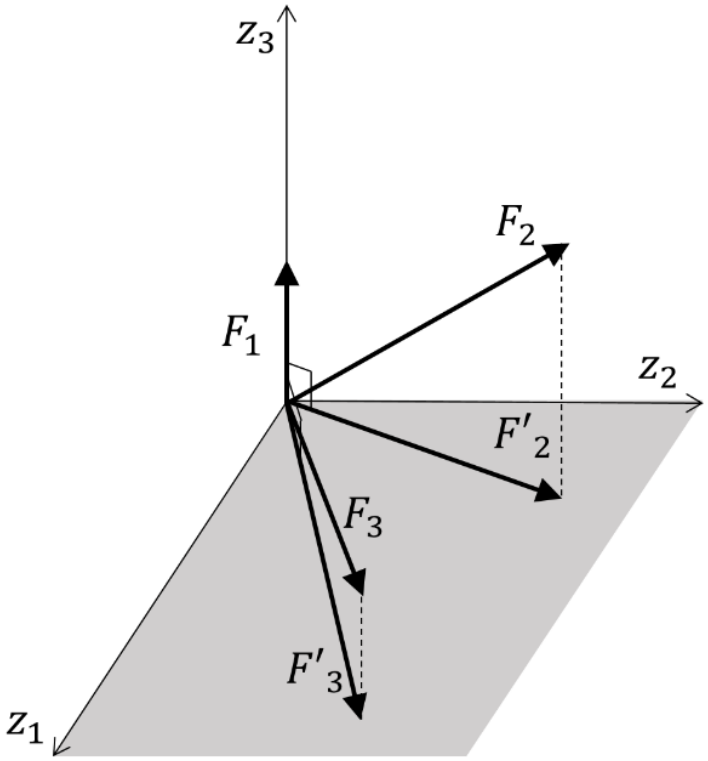


Figure 55 - Projection des descripteurs F_i orthogonalement au descripteur le mieux classé.

3.3.5 Prédiction des valeurs des LSF

Ces descripteurs nous permettent de prédire la valeur des coefficients du filtre du conduit vocal. Pour cela, nous utilisons un réseau de neurones de type perceptron multicouche (MLP) de sorte que chacun des 12 coefficients LSF par trame est prédit par un MLP. Il s'agit de réseaux de neurones non bouclés dont les neurones cachés ont une fonction d'activation sigmoïde. Pour chaque LSF à prédire, nous cherchons à minimiser l'erreur entre la sortie du réseau et la sortie désirée, qui est la valeur de LSF calculée depuis le signal acoustique. Nous cherchons le modèle le plus adapté à nos données en faisant varier le nombre de couches cachées et en faisant plusieurs initialisations des poids du réseau.

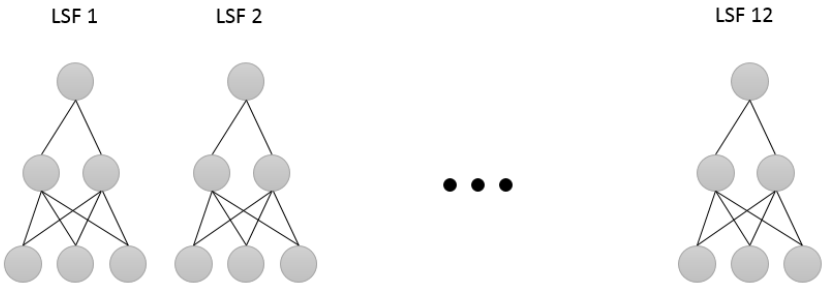


Figure 56 - Illustration des 12 réseaux de neurones de type perceptrons multicouches dont la fonction est de prédire la valeur des LSF à partir des descripteurs sélectionnés par OFR. Chaque perceptron possède une couche cachée avec une fonction d'activation sigmoïde puis une sortie linéaire.

Pour chaque modalité, les données ont été divisées en une base d'apprentissage et de validation de 35000 images, présentées de manière aléatoire, et une base de test de 5000 images consécutives. Les 35000 images de la base d'apprentissage et de validation ont été utilisées pour entraîner l'autoencodeur ainsi que pour trouver les *EigenTongues* et *EigenLips*. Ces modèles ont ensuite été utilisés sur la base de test sans réapprentissage. Les descripteurs ainsi extraits ont été aléatoirement répartis et 60 % d'entre eux ont été utilisés pour entraîner les perceptrons à l'aide de l'algorithme de Levenberg-Marquardt et les 40 % restants ont été utilisés pour la validation. Les perceptrons multicouches ont ensuite été utilisés sur la base de test indépendante, sans réapprentissage. Nous avons optimisé le nombre d'unités cachées de chaque MLP de façon à minimiser l'erreur de validation. Les 12 MLP utilisés pour la prédiction des LSF ont chacun une trentaine d'entrées, entre 1 et 7 unités cachées et une sortie, chacune correspondant à l'un des 12 LSF. Nous avons testé 50 initialisations des paramètres du modèle. Les valeurs des LSF n'étant pas comprises entre 0 et 1, les neurones de la couche cachée ont une fonction d'activation sigmoïde et le neurone de sortie a une fonction d'activation linéaire.

3.3.6 Comparaison entre les méthodes

Afin de comparer l'approche utilisant les *EigenLips* et *EigenTongues* et celle utilisant l'autoencodeur multimodal, nous avons choisi d'utiliser des structures comparables. Ainsi, nous travaillons dans chacun des cas sur les mêmes bases d'apprentissage et de validation, avec des images identiquement redimensionnées. Chacune des méthodes nous permet d'extraire 100 descripteurs, pour lesquels nous appliquons la méthode de sélection de variable supervisée rapportée ci-dessus (voir section 3.3.4). Ces descripteurs sont ensuite utilisés dans un cas comme dans l'autre comme entrée d'un perceptron multicouche par LSF, l'architecture

de ces perceptrons étant déterminée de façon à minimiser l'erreur de validation pour chaque méthode. Nous cherchons à comparer les descripteurs extraits par la méthode d'autoencodeur multimodal, méthode non linéaire d'extraction de descripteurs, à ceux extraits par la méthode des *EigenLips* et *EigenTongues*. Nous détaillerons dans la section3.3.6 comment nous avons effectué cette comparaison, en utilisant les informations acoustiques obtenues à partir des informations articulatoires afin de synthétiser des signaux de voix chantée.

3.4 Méthodes de synthèse vocale

Une fois les LSF prédits, nous pouvons modéliser la façon dont la configuration du conduit vocal influence le flux d'air. Ces informations, combinées à un modèle de source, permettent de synthétiser de la voix. Pour synthétiser les sons voisés, nous utilisons et comparons différents types de signaux de source, à savoir les résidus de prédiction LPC, du bruit blanc, la dérivée temporelle du signal électroglottographique et un modèle d'onde de débit glottique. Pour les sons non voisés, nous utilisons simplement un bruit blanc dont la durée est raccourcie.

3.4.1 Utilisation de signaux d'excitation

Les résidus de prédiction LPC, dont un exemple est donné Figure 18, constituent le signal de source le plus simple et aussi de meilleure qualité car ils contiennent toutes les informations qui ne sont pas contenus dans les coefficients de prédiction.

Un bruit blanc est souvent utilisé afin de générer des sons non voisés. Cependant, l'utilisation d'un bruit blanc comme signal de source quel que soit le voisement du son permet d'obtenir un signal de parole proche de la voix murmurée. Ce type de signal d'excitation a l'avantage d'être très simple à mettre en œuvre et de ne pas nécessiter de connaissances a priori. Cependant, la qualité médiocre de la voix synthétique obtenue en utilisant un bruit blanc en guise de source ne permet pas son utilisation pour de la voix chantée.

L'utilisation de l'EGG comme signal de source est discutable en synthèse, car le signal EGG ne permet pas d'accéder directement la source glottique. En effet, même si l'EGG, après dérivation, permet de déterminer les instants d'ouverture et de fermeture glottique (voir Figure 57) avec davantage de précision que ne le permet le signal acoustique, c'est un signal

qui comporte beaucoup d'artefacts, en particulier liés aux mouvements de la tête et du cou.

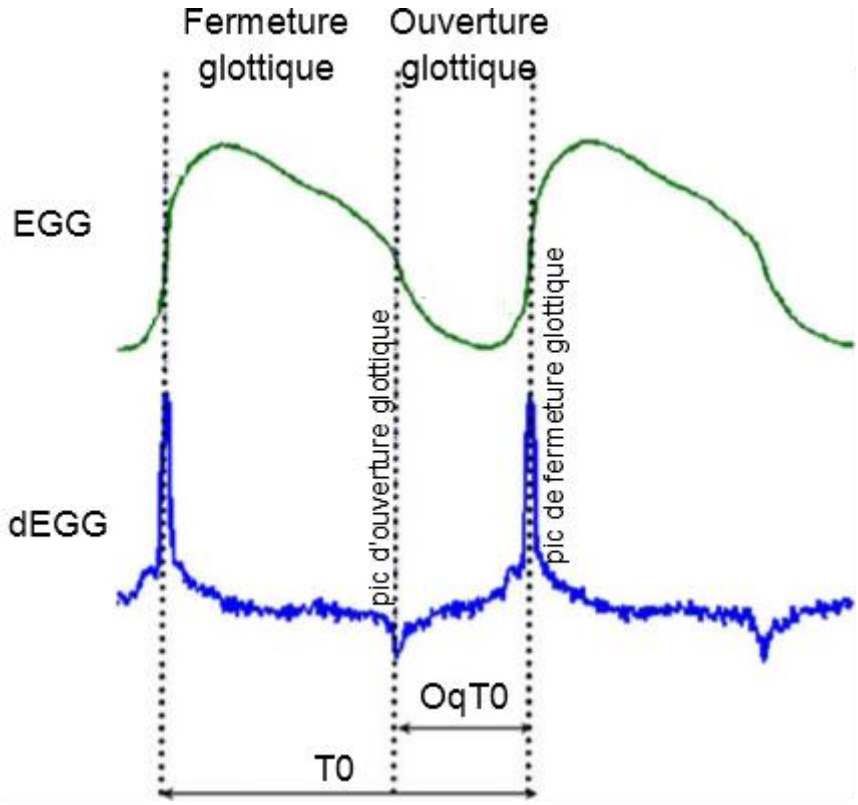


Figure 57 - En vert, un exemple de signal EGG et sa dérivée en bleu. Les pics positifs de la dérivée de l'EGG correspondent à des fermetures glottiques tandis que les pics négatifs du signal de dEGG correspondent à des ouvertures glottiques. L'identification des instants d'ouverture et de fermeture glottique permet de déterminer la période fondamentale ainsi que le quotient ouvert.

3.4.2 Construction de signaux d'excitation

L'EGG peut être utilisé pour déterminer les instants d'ouverture et de fermeture glottique. Il permet donc de calculer de manière assez précise les valeurs des paramètres de différents modèles d'onde de débit glottique comme le modèle LF ou le modèle CALM. Dans nos travaux, nous avons choisi d'utiliser l'EGG pour déterminer l'évolution au cours du temps des paramètres du modèle CALM, permettant une synthèse vocale avec une qualité vocale évoluant au cours du temps.

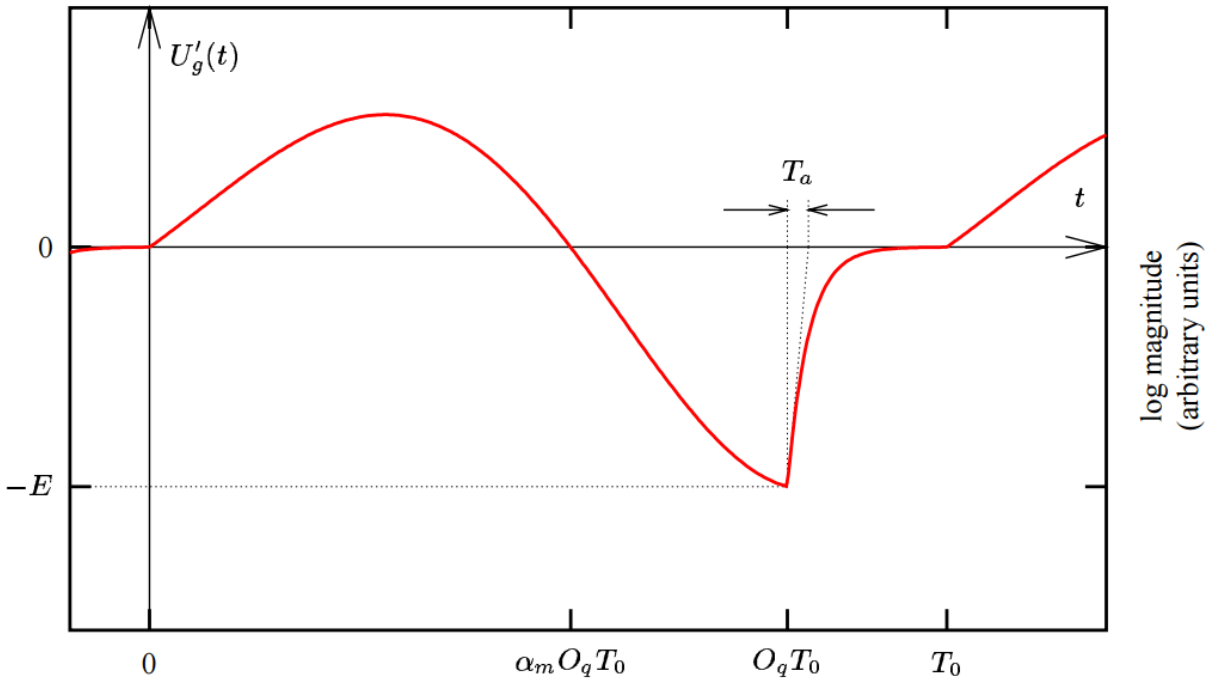


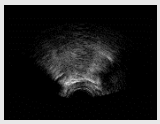
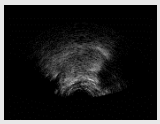
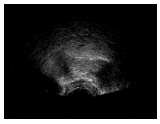
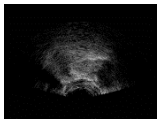
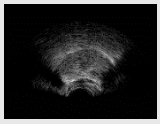
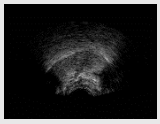
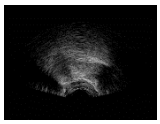
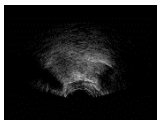
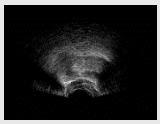
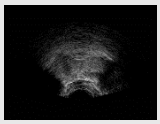
Figure 58 – Modèle d'onde de débit glottique dérivée. Sur cette figure sont représentés les paramètres du modèle CALM.

Ces méthodes de synthèse produisent des sons de qualités différentes, que nous cherchons à évaluer.

3.5 Application à une base de voyelles chantées isolées

Dans un premier temps, nous avons utilisé notre modèle de synthèse vocale sur une base de voyelles chantées isolées. Cette base est constituée par la répétition par un chanteur des voyelles (tenues sur la durée et chantées sur la même note) /i/, /o/, /e/, /O/, /a/. Dans cette base, cette succession de voyelles est répétée 4 fois. Elle comprend au total 6036 images de chaque type, soit une durée de 100,6 secondes, c'est-à-dire 1 minutes et 40 secondes de chant. Les images de cette base ont la particularité d'être très stables au cours du temps, c'est-à-dire que pour une voyelle donnée, la position prise par la langue comme celle prise par les lèvres évoluent peu. Le Tableau 8 montre des images de la langue et des lèvres pour chacune des voyelles de la base pour des trames données, lors de la première répétition des voyelles. Afin de travailler uniquement sur la voix chantée, nous avons supprimé les silences sur cette base. L'évaluation des résultats de prédiction des LSF peut être faite par observation directe des valeurs des LSF. La Figure 59 et la Figure 60 montrent la superposition des LSF théoriques avec les LSF estimés par la méthode d'autoencodeur sans ajout de bruit.

Tableau 8 - Constitution de la base de voyelles isolées. Nous pouvons remarquer qu'il y a peu de variabilité entre les images enregistrées lors de la réalisation d'une même voyelle..

Voyelle	Langue en début de voyelle	Langue en fin de voyelle	Lèvres en début de voyelle	Lèvres en fin de voyelle	N° de la trame montrée en début de voyelle	N° de la trame montrée en fin de voyelle
/i/					0078	0205
/o/					0368	0483
/e/					0735	0836
/O/					0994	1093
/a/					1275	1388

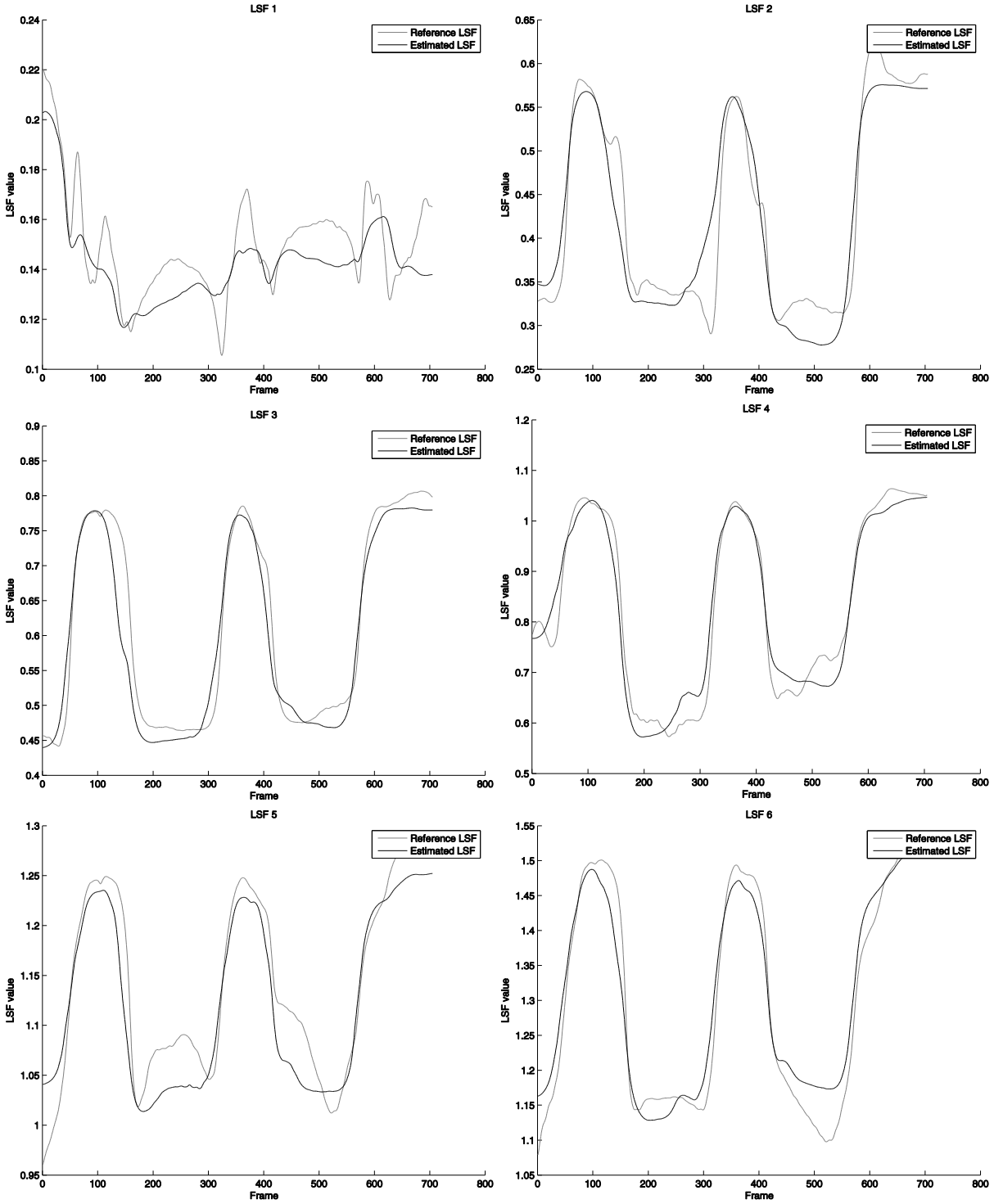


Figure 59 - Comparaison entre les valeurs de référence et les estimations des six premiers LSF en utilisant l'autoencodeur multimodal sur la base de voyelles isolées. Ces figures indiquent une bonne prédiction des LSF et donc des pertes de qualité vocaliques faibles.

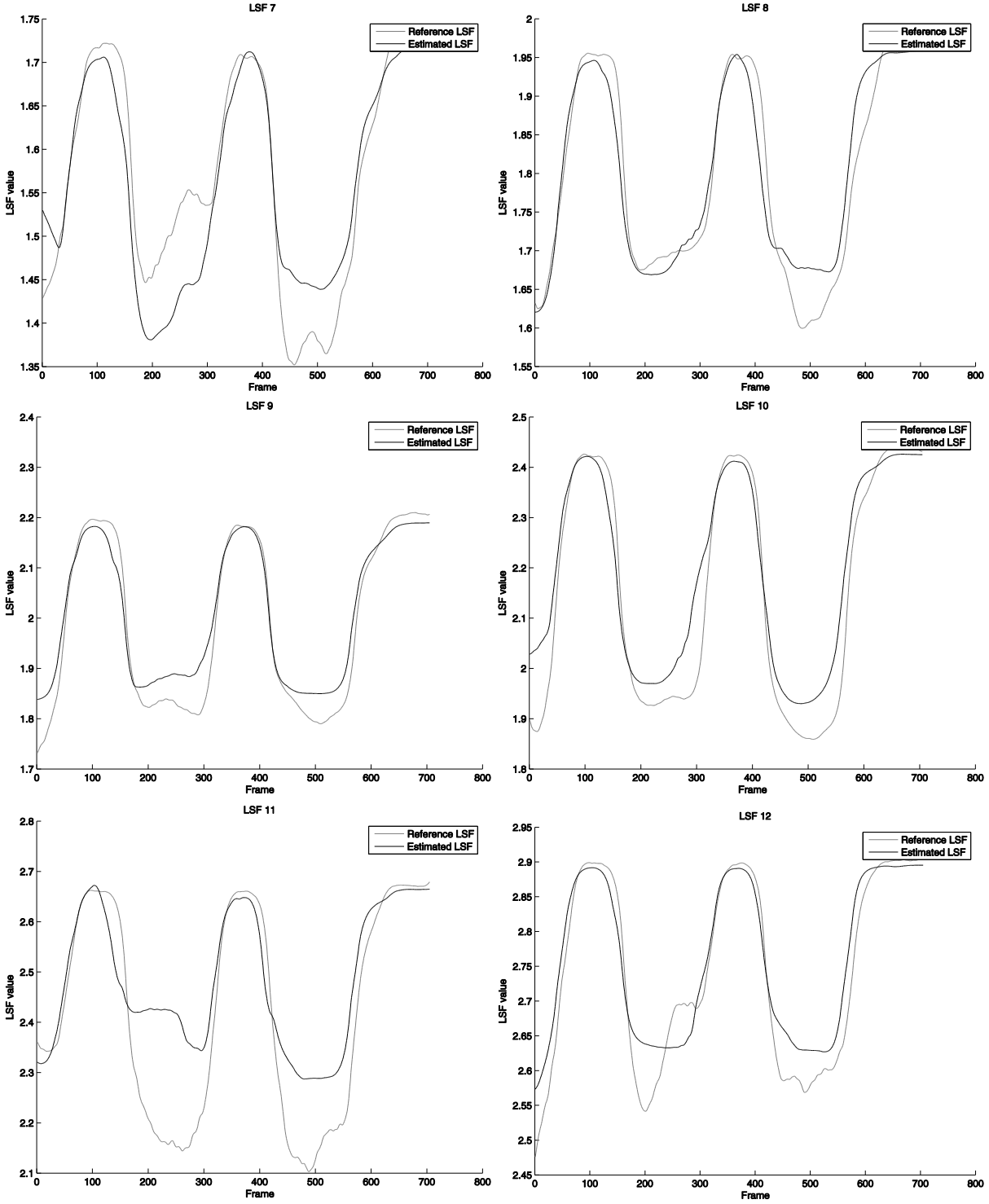


Figure 60 - Comparaison entre les six derniers LSF de référence et les LSF estimés par l'autoencodeur multimodal sur la base de voyelles isolées. La prédiction des six derniers LSF est un peu moins précise que celle des six premiers.

Il est également possible d'utiliser un score dit de distorsion spectrale moyenne, qui traduit la déformation du spectre du signal induite par les écarts de valeur des LSF par rapport aux valeurs de référence, dans des bandes de fréquences données [128]. Le calcul de la distorsion spectrale est donné par l'équation suivante :

$$SD = \left\langle \sqrt{\frac{1}{(n_1 - n_0)} \sum_{k=n_0}^{n_1-1} \left(10 \log_{10} \left(\left| \frac{A\left(e^{\frac{j2\pi k}{N}}\right)}{A'\left(e^{\frac{j2\pi k}{N}}\right)} \right|^2 \right) \right)^2} \right\rangle \quad (43)$$

Dans cette équation, A et A' désignent respectivement les polynômes LPC issus des LSF d'origine (calculés à partir du signal acoustique) et les LSF estimés. N désigne l'ordre de la FFT et vaut 512. Les valeurs de n_0 et n_1 déterminent les bandes de fréquences sur lesquelles est effectué le calcul. Dans nos calculs, $n_0 = 6$ et $n_1 = 200$, ce qui donne, pour des fréquences allant de 129 à 4307 Hz, des bandes fréquentielles de largeur 21,5 Hz. Une distorsion « transparente » pour l'oreille est une distorsion de 1 dB. Nous utilisons également une note d'opinion moyenne différentielle (ou ΔMOS, pour *Differential Mean Opinion Score*). Ce score est défini d'après une distorsion de référence de valeur 1 dB par les expressions suivantes [128] :

$$MOS = 3.56 - 0.8\,SD + 0.04\,SD^2 \quad (44)$$

$$\Delta MOS(SD) = MOS(SD) - MOS(1dB) \quad (45)$$

Ces deux métriques sont issues des technologies de la communication et l'information, et servent à évaluer les performances de codage et transmission de la parole. Par ailleurs, les tests de synthèse audio permettent de comparer un signal original à sa version synthétique. La valeur de la distorsion spectrale obtenue sur les exemples de la base de validation nous ont permis de sélectionner l'architecture la plus adaptée à nos besoins. Nous avons dans un premier temps cherché à faire varier le taux d'apprentissage utilisé lors de l'apprentissage des RBM (voir section 1.4.31.4.3). Les résultats sont donnés dans le Tableau 9 :

Tableau 9 - Evolution du score de distorsion spectrale en fonction des valeurs du taux d'apprentissage initial et du taux d'apprentissage final.

Taux d'apprentissage initial	Taux d'apprentissage final (à partir de nEpoch = 10)	Distorsion spectrale (validation)
0,001	0,005	3,5 dB
0,005	0,01	3,4 dB
0,01	0,05	3,3 dB
0,05	0,05	3,5 dB
0,01	0,01	3,2 dB

Le taux d'apprentissage ne semble pas être un paramètre décisif dans la qualité de la prédiction. Nous avons donc choisi un taux d'apprentissage fixe, égal à 0,01.

Nous avons ensuite cherché à modifier la structure de l'autoencodeur, d'abord en faisant varier la taille de la couche cachée du second RBM, le premier RBM correspondant à la mise en commun des informations correspondant à la langue et aux lèvres. Les résultats obtenus sont rassemblés dans le tableau suivant :

Tableau 10 - Evolution du score de distorsion spectrale en fonction de la taille de la couche cachée du 2ème RBM.

Taille de la couche cachée du 2ème RBM	Distorsion spectrale (validation)
1000	3,3 dB
2000	3,2 dB
3000	3,3 dB

D'après ces résultats, la taille du second RBM ne nous a pas semblé déterminante pour l'amélioration de la précision de l'estimation des valeurs des LSF. Nous avons ensuite cherché à modifier la taille du premier RBM, celui pour lequel les données de la langue et des lèvres sont utilisées séparément. Les résultats obtenus sont rassemblés dans le tableau suivant :

Tableau 11 - Evolution du score de distorsion spectrale en fonction de la taille de la couche cachée du premier RBM.

Taille de la couche cachée du 1er RBM	Distorsion spectrale (validation)
500	2,6 dB
1500	3,2 dB
2000	3,2 dB

Sur cette base de données, réduire la dimension du premier RBM nous a permis d'améliorer la qualité de la prédiction des LSF à partir des données d'entrée. Par ailleurs, compte tenu du peu de variabilité des images de cette base, qui peuvent être catégorisées en autant de classes que de voyelles, il est nécessaire d'utiliser des mini-batches de taille relativement réduite [79]. Le passage de l'utilisation de mini-batches de taille 100 à des mini-batches de taille 50 permet d'augmenter la précision de la prédiction des LSF, ce qui se traduit par une nette diminution de la distorsion spectrale. La comparaison entre la méthode d'autoencodeur multimodal et la méthode des *EigenLips* et *EigenTongues* donne des résultats satisfaisants en base de test. En effet, le score de distorsion spectrale est de 2,2 dB en base de test (proche de la distorsion transparente) en utilisant une extraction de descripteurs par autoencodeur multimodal, tandis que la prédiction à partir des descripteurs extraits par *EigenTongues* et *EigenLips* est de 3,0 dB sur la même base de test. Cette différence provient de la prédiction plus précise des valeurs de LSF par l'autoencodeur. Afin de confirmer l'efficacité de la prédiction des LSF en comparaison avec une méthode linéaire comme les *EigenLips* et *EigenTongues*, notre modèle gagnerait à être testé en conditions réelles, en incluant silences et fricatives, sur des chants entiers. Ce type de données permettrait d'éviter les problèmes de redondance des données qui ne sont pas complètement indépendantes. C'est ce que nous présentons dans la partie suivante.

3.6 Application à une base de chants traditionnels

Puisque la variabilité d'une base de données de voyelles tenues et isolées et très limitée, nous avons souhaité tester notre algorithme sur une base de chants entiers. Nous disposons d'une base enregistrée par un chanteur sur 5 chants en corse et en latin, incluant des répétitions. Cette base totalise 43 413 images, soit environ 12 minutes d'enregistrement.

Cette base comporte des sons voisés ainsi que des sons non voisés et des silences. Nous avons

dans un premier temps travaillé sur la seule partie voisée de la base, puis sur l'ensemble des sons. Sur une base aussi complexe, il est difficile d'extraire des descripteurs permettant de représenter l'ensemble des variations dans les données. C'est pourquoi le choix de l'architecture profonde, et en particulier celui des hyperparamètres de l'autoencodeur profond joue un rôle central dans l'extraction de descripteurs articulatoires.

3.6.1 Choix de l'architecture profonde

Comme dans le réseau permettant d'extraire le contour des images ultrasonores, les hyper-paramètres du réseau doivent être déterminés de façon à permettre l'extraction de descripteurs représentant au mieux les données d'entrée. Si l'apprentissage ne permet pas de décrire précisément les données, alors les descripteurs extraits ne permettront pas de prédire correctement les coefficients du filtre du conduit vocal. Outre l'erreur de reconstruction de l'autoencodeur, nous avons dans un premier temps utilisé une mesure de distance permettant de comparer les descripteurs aux valeurs de LSF désirés. Il s'agit en d'utiliser le score de la première variable par l'algorithme OFR afin d'optimiser le choix des hyper-paramètres. En effet, plus cette valeur est proche de 1, plus elle traduit une forte capacité de prédiction des LSF par les descripteurs extraits. Pour plus de simplicité, nous appellerons cette quantité « score d'OFR ». C'est un score rapide à calculer. Le Tableau 12 présente la valeur de ce score pour la prédiction de chacun des LSF en utilisant la méthode des *EigenLips* et *EigenTongues*.

Tableau 12 – Score d'OFR de chaque LSF en utilisant les descripteurs extraits par la méthode EigenLips/EigenTongues.

LSF	**Score d'OFR en utilisant la méthode *EigenLips + EigenTongues***
LSF 1	0,37
LSF 2	0,36
LSF 3	0,21
LSF 4	0,23
LSF 5	0,24
LSF 6	0,16
LSF 7	0,28
LSF 8	0,17
LSF 9	0,24
LSF 10	0,18
LSF 11	0,16
LSF 12	0,22

Le premier constat que nous pouvons faire à la lecture de ce tableau est que les deux premiers LSF ont de meilleures chances d'être bien prédits que les suivants. Ceci peut s'expliquer par le fait que les premiers LSF contiennent davantage d'informations pertinentes par rapport à l'articulation que les derniers. En effet, les LSF ont une relation avec les coefficients de prédiction LPC, qui sont eux-mêmes corrélés aux formants. Or, la valeur des 4 ou 5 premiers formants suffit à décrire la nature d'une voyelle. Dans certains cas, une classification n'utilisant que la valeur des deux premiers formants peut suffire à reconnaître des voyelles.

Tableau 13 - Score d'OFR de chaque LSF en utilisant les descripteurs extraits par la méthode d'autoencodeur multimodal (DAE) selon le nombre d'unités par couche cachée. Les valeurs indiquées correspondent au nombre d'unités par couche en commençant par la première couche jusqu'à la couche la plus profonde d'encodage.

Score d'OFR selon la méthode employée

	DAE 500-1980-200	DAE 500-2000-200	DAE 500-1000-500-200
LSF 1	0,37	0,39	0,40
LSF 2	0,42	0,42	0,42
LSF 3	0,18	0,16	0,19
LSF 4	0,33	0,32	0,36
LSF 5	0,29	0,31	0,31
LSF 6	0,19	0,20	0,15
LSF 7	0,28	0,29	0,28
LSF 8	0,21	0,24	0,22
LSF 9	0,32	0,27	0,31
LSF 10	0,20	0,22	0,24
LSF 11	0,24	0,23	0,27
LSF 12	0,27	0,28	0,33

Tableau 14 - Score d'OFR de chaque LSF en utilisant les descripteurs extraits par la méthode d'autoencodeur multimodal (DAE) selon le nombre d'unités par couche cachée. Les valeurs indiquées correspondent au nombre d'unités par couche en commençant par la première couche jusqu'à la couche la plus profonde d'encodage.

Score d'OFR selon la méthode employée

	DAE 100-400-200	DAE 250-500-250-500	DAE 1000-2000-200
LSF 1	0,39	0,42	0,41
LSF 2	0,40	0,42	0,41
LSF 3	0,17	0,20	0,08
LSF 4	0,33	0,35	0,24
LSF 5	0,33	0,29	0,27
LSF 6	0,18	0,22	0,15
LSF 7	0,30	0,27	0,26
LSF 8	0,25	0,22	0,21
LSF 9	0,32	0,30	0,29
LSF 10	0,24	0,23	0,27
LSF 11	0,24	0,24	0,21
LSF 12	0,25	0,30	0,28

Ce tableau ne nous permet pas particulièrement d'identifier de configuration plus adaptée à la prédiction des LSF que les autres. Cependant, il nous permet d'identifier les LSF les mieux prédictibles par notre méthode, à savoir les 5 premiers, à l'exclusion du 3ème. En particulier, les deux premiers LSF sont ceux qui ont le plus de chance d'être bien prédits. Ce sont également ceux qui ont le plus de sens d'un point de vue acoustique. Ainsi, une variation significative de la précision de l'estimation de la valeur de ces LSF devrait avoir une influence toute aussi significative sur la qualité du signal synthétique que l'on pourrait reconstruire en utilisant ces valeurs.

Cependant, l'algorithme d'OFR est couteux en temps de calcul. De même, l'étage de prédiction des LSF à l'aide de perceptrons multicouches avec sélection de la meilleure architecture (nombre d'unités cachées) prend plusieurs heures. Afin d'accélérer les tests de resynthèse pour le choix des hyper-paramètres de l'architecture profonde, nous avons opté, dans le cadre de la comparaison entre les différents paramètres, pour une méthodologie plus rapide. Ainsi, nous avons choisi de remplacer les perceptrons multicouches par une simple

régression. De même, nous avons remplacé l'algorithme d'OFR permettant de classer tous les descripteurs par un algorithme beaucoup plus rapide donnant accès uniquement au descripteur le mieux classé. Ainsi, pour la phase de validation de l'architecture, au lieu d'utiliser une sélection de variables supervisée puis de nombreux apprentissages de perceptrons multicouches, nous avons utilisé deux méthodes donnant une estimation du score de distorsion spectrale s'il était calculé pour une prédiction des LSF utilisant OFR et MLP : une régression simple n'utilisant que le meilleur descripteur, puis une régression simple utilisant l'ensemble des descripteurs. La comparaison des scores de distorsion spectrale obtenus par cette méthode permet de comparer les architectures entre elles, afin d'affiner plus rapidement le choix de l'architecture profonde. Cette méthode rapide nous a permis de constater qu'il était possible de réduire le nombre de sorties de l'autoencodeur multimodal à 100 sans impacter significativement la distorsion spectrale du signal reconstruit.

3.6.2 Construction du signal d'onde de débit glottique

La construction du signal de source utilisant un modèle d'onde de débit glottique a été effectuée en deux étapes : premièrement, le modèle a été utilisé avec des paramètres dont les valeurs étaient fixes au cours du temps. Les valeurs sont données dans le tableau suivant.

Tableau 15 - Les paramètres de source utilisés dans notre modèle avec leurs valeurs typiques et les valeurs fixées pour une synthèse vocale générique (non spécifique à un locuteur).

Paramètre	**Valeurs typiques**	**Valeurs fixées**
Coefficient d'asymétrie	0,66-0,8	0,8
Quotient ouvert	0,35-1	0,35
Fréquence fondamentale	Quelques centaines de Hz	100 Hz
Energie	-	10^{-2}

Ces valeurs ainsi fixées ne donnaient pas une synthèse très réaliste, en particulier en ce qui concerne la fréquence fondamentale. C'est pourquoi nous avons mis en place une mise à jour de ces paramètres pour chaque fenêtre de temps entourant un instant de fermeture glottique en utilisant le signal électroglottographique. Les étapes de la synthèse vocale sont précisées sur la Figure 61.

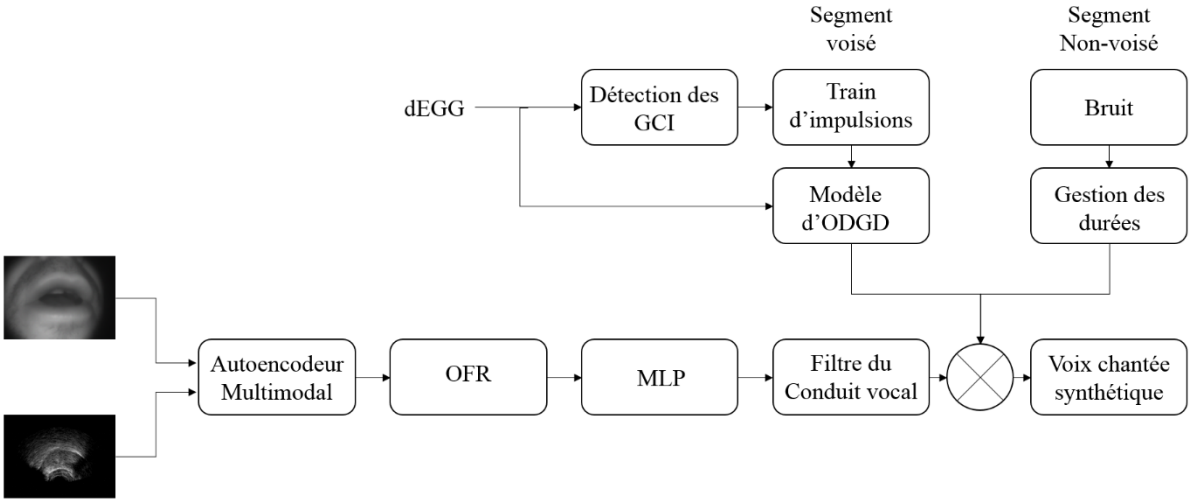


Figure 61 - Illustration schématique de la méthode de synthèse vocale à partir des données articulatoires et glottiques.

Nous utilisons dans un premier temps l'étiquetage de voisement détaillé dans la section 3.2.3. Pour les segments détectés comme étant voisés, le signal électroglottographique nous permet de déterminer les instants d'ouverture et de fermeture glottique, ce qui nous donne accès à la fréquence fondamentale du signal de voix chantée. Par ailleurs, le calcul du coefficient d'asymétrie ou du quotient ouvert, qui sont des caractéristiques du timbre de la voix d'un chanteur, découlent également de la détermination de ces paramètres.

3.6.3 Choix des descripteurs

Nous avons utilisé un autoencodeur permettant une compression des données d'entrée, en proposant une représentation sur 100 descripteurs des informations contenues sur les des deux types d'images (langue et lèvres), soit 1980 unités. Nous avons par conséquent construit notre architecture utilisant les *EigenLips* et *EigenTongues* sur le même modèle, de façon à extraire également 100 descripteurs. Sur les 100 descripteurs ainsi extraits sur la base d'apprentissage, nous avons utilisé la méthode d'OFR afin de classer les descripteurs selon leur capacité de prédiction. Pour chaque LSF, environ 80 descripteurs en moyenne étaient mieux classés que 90% des réalisation de la variable sonde, formant l'ensemble des descripteurs efficaces. Afin d'alléger l'architecture des modèles et de nous prémunir des risques de sur-ajustement, nous n'avons conservé qu'un sous-ensemble de ces descripteurs efficaces. Nous avons considéré le nombre de descripteurs efficaces conservés (par ordre de classement de l'algorithme OFR) comme un hyperparamètre que nous avons optimisé à partir du score de validation.

Cette optimisation nous a amené à conserver un tiers des descripteurs classés par l'algorithme d'OFR, soit une trentaine de descripteur par LSF. L'utilisation de méthode combinée avec une prédiction par MLP donne des résultats proches des résultats obtenus lorsque l'on utilise simplement les 100 descripteurs avec une prédiction par une régression linéaire. Ce constat valide l'intérêt d'utiliser la méthode simplifiée (présentée en section 3.6.1) pour l'optimisation des hyper paramètres de l'autoencodeur multimodal, bien que l'amélioration de la prédiction des LSF en utilisant OFR et MLP motive l'utilisation de cette méthode plus complexe.

3.6.4 Résultats de prédiction des LSF

Sur la base de données de 43 413 images, 35 000 images sont utilisées en apprentissage et 5000 en validation. Dans [89], il est rapporté que les silences et fricatives ont un effet qui tend à moyenner la valeur des LSF prédits. Afin d'éviter que la prédominance des silences et fricatives ne vienne moyenner les valeurs des LSF prédites, nous choisissons de modifier les proportions de trames voisées et non voisées de la base d'apprentissage. Ainsi, sur les 35 000 exemples de la base d'apprentissage, 30 000 correspondent à des trames voisées et 5 000 à des trames non voisées. Sur cette base de validation incluant silences et fricatives, la méthode utilisant les *EigenTongues* et *EigenLips* a obtenu un score de distorsion spectrale de 5,2 dB sur la base de test. Une illustration de la reconstruction des LSF par la méthode des *EigenLips* et *EigenTongues* est donnée Figure 62 et Figure 63.

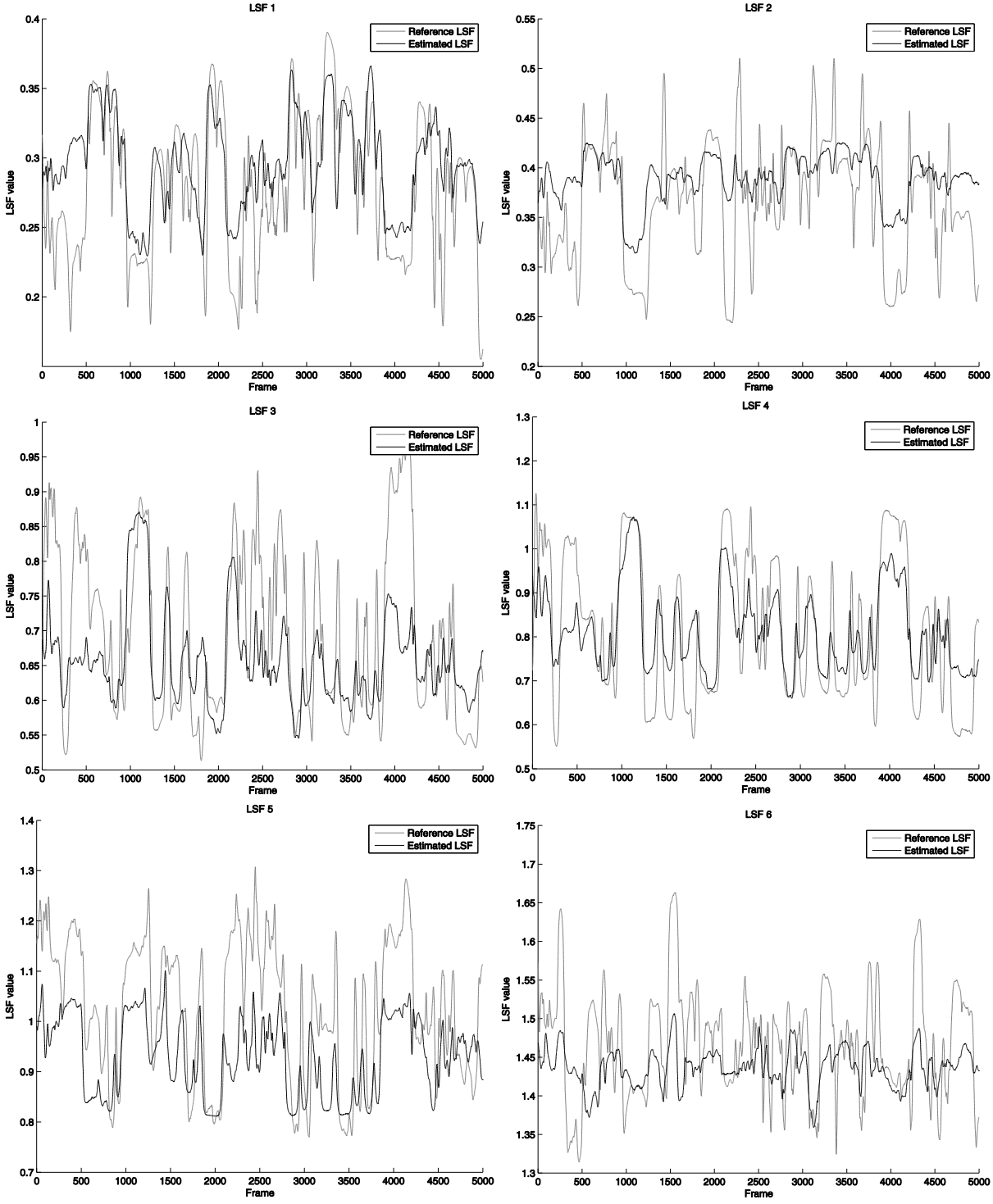


Figure 62 - Comparaison entre les six premiers LSF de référence et les LSF estimés par le modèle EigenLips/EigenTongues *sur la base de chants traditionnels.*

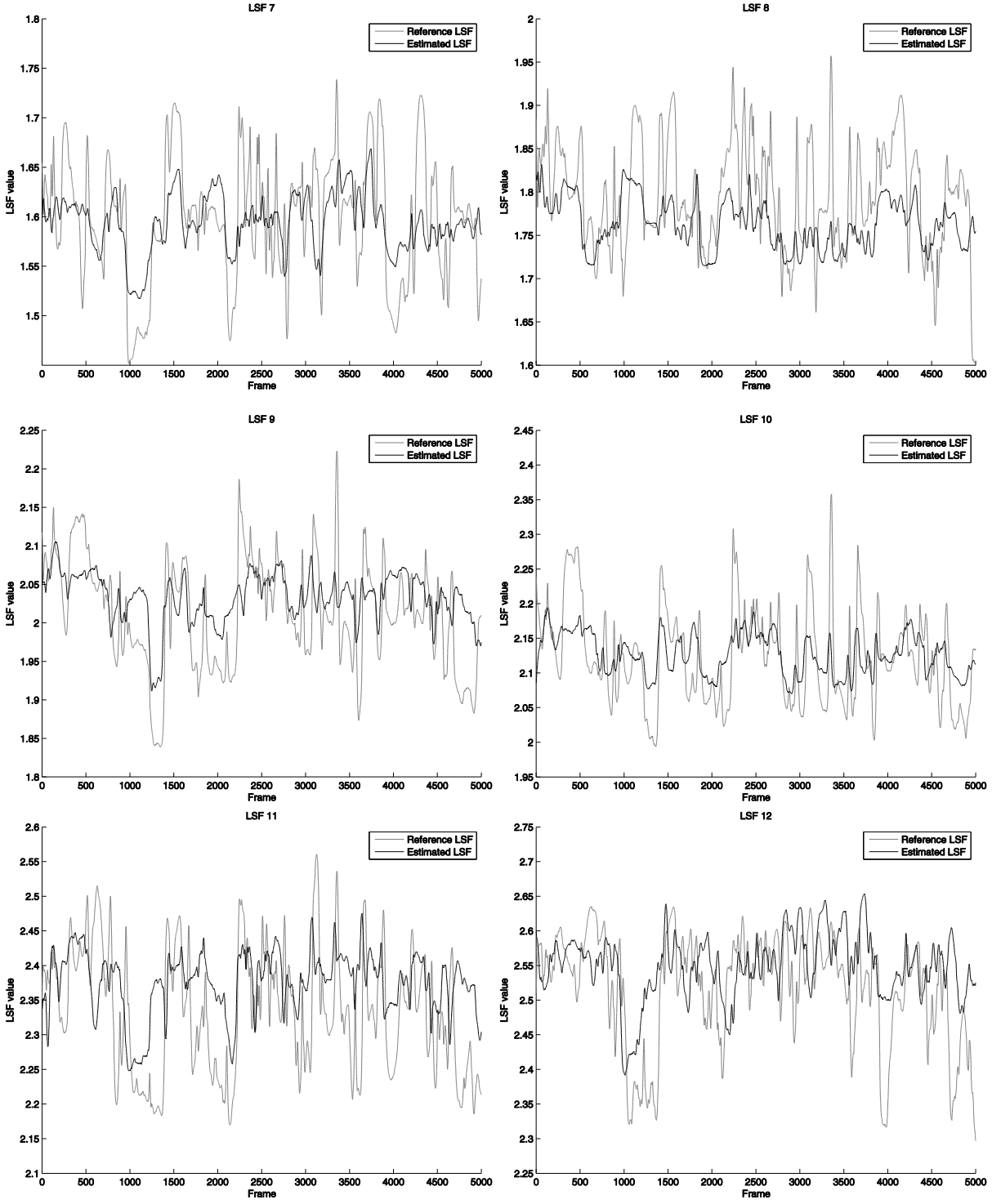


Figure 63 - Comparaison entre les six derniers LSF de référence et les LSF estimés par le modèle EigenLips/EigenTongues *sur la base de chants traditionnels.*

Il apparaît très clairement que les premiers LSF sont mieux prédits que les derniers, ce qui est en accord avec la discussion détaillée au paragraphe 3.6.1. La méthode de *Deep Learning* quant à elle obtient des résultats de prédiction plus précis, en particulier lors des transitions

abruptes, avec un score de distorsion spectrale de 4,3 dB sur la base de test, en incluant les silences et les fricatives.

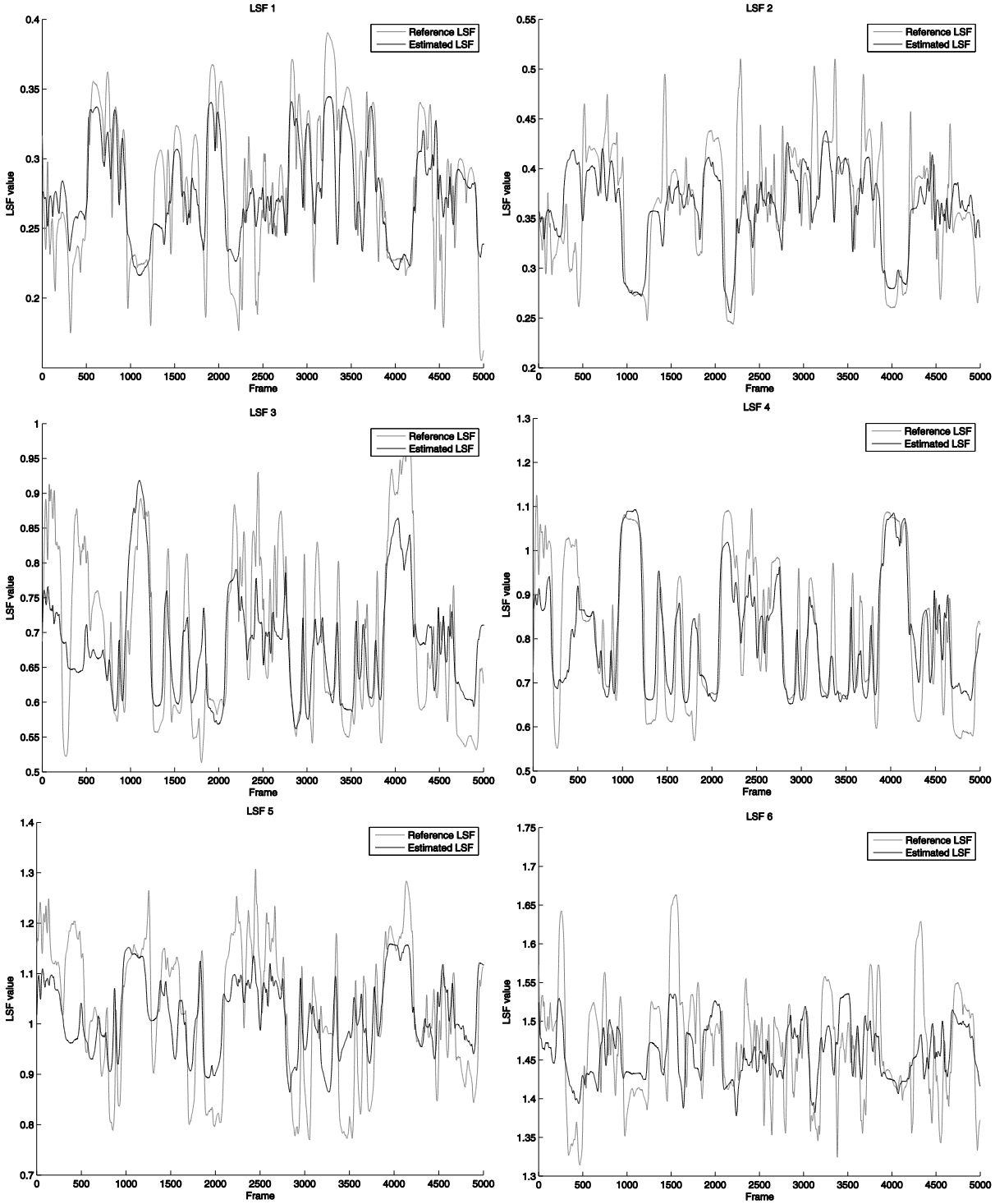


Figure 64 - Comparaison entre les six premiers LSF de référence et les LSF estimés par l'autoencodeur multimodal sur la base de chants traditionnels.

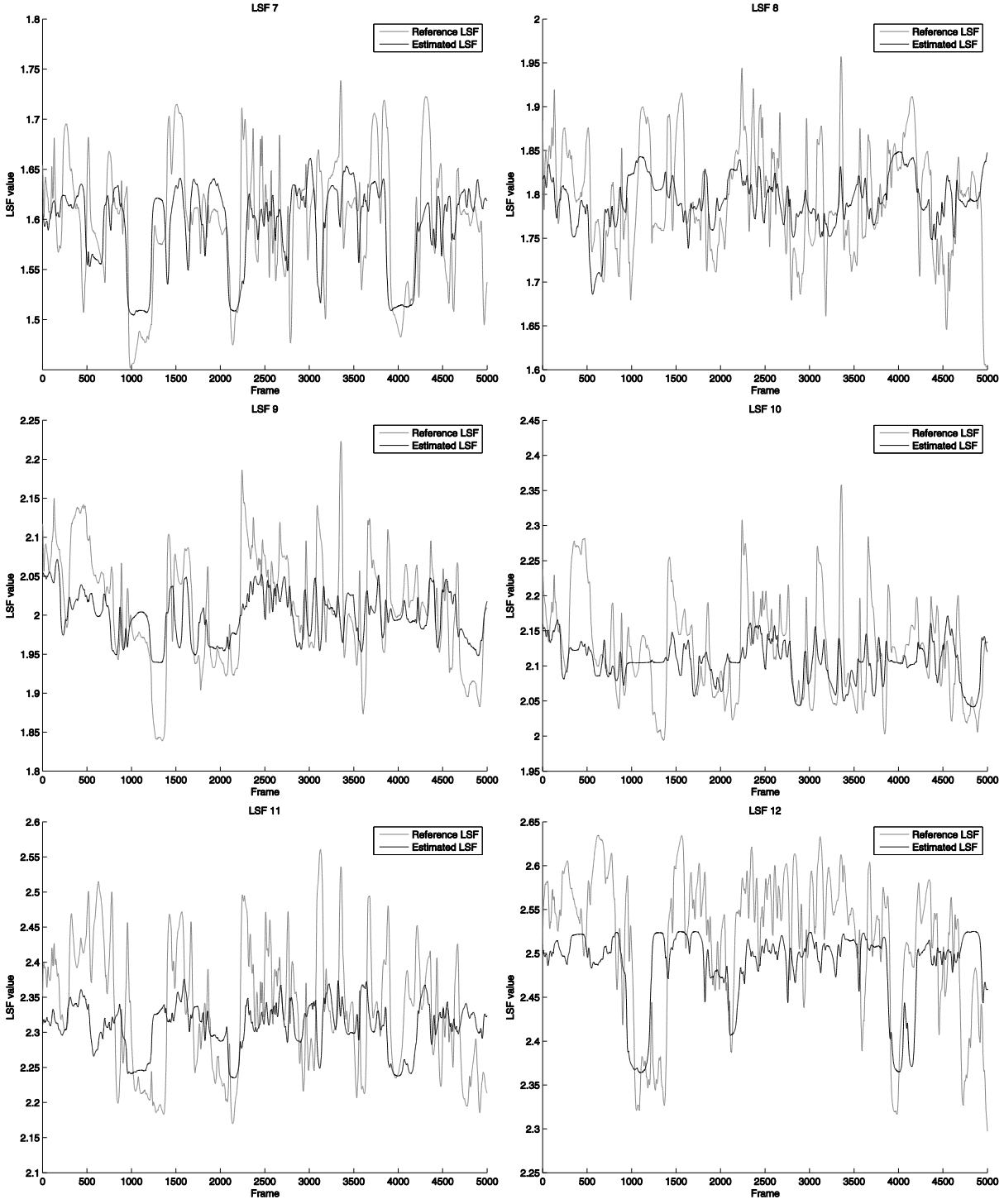


Figure 65 - Comparaison entre les six derniers LSF de référence et les LSF estimés par l'autoencodeur multimodal sur la base de chants traditionnels.

Le Tableau 16 regroupe les résultats de distorsion spectrale et de score d'opinion moyen différentiel sur la base de test.

Tableau 16 – Scores de distorsion spectrale et notes d'opinions moyennes différentielles obtenus pour les deux méthodes testées en comparaison avec une distorsion de 1 dB, transparente à l'oreille.

Méthode	**Distorsion Spectrale (dB)**	ΔMOS
Distorsion "Transparente"	1	0
Autoencodeur Multimodal	4.3	-1.9
EigenLips et *EigenTongues*	5.2	-2.3

3.7 Evaluation perceptive

Une fois les modèles de prédiction d'une part des paramètres de source et d'autre part des paramètres du filtre développés, nous pouvons utiliser ces modèles en synthèse. Nous avons développé un module d'analyse articulatoire permettant de synthétiser un son à partir d'un couple d'images des articulateurs que sont la langue et les lèvres. Ce module permet de traduire une articulation donnée en son. Cet outil permet donc de comparer les différents sons produits par la modification de la position de la langue ou l'aperture des lèvres. Il inclut en outre la possibilité de personnaliser la qualité de la voix en jouant sur les paramètres de source. Nous avons par ailleurs développé un deuxième modèle de synthèse vocale, plus complet, qui combine les prédictions des propriétés du filtre du conduit vocal à l'estimation des paramètres de source spécifiques à un locuteur, illustré Figure 61.

Nous avons mis en place un test perceptif en ligne permettant à des auditeurs d'écouter et évaluer la qualité des différents extraits de voix chantée produits par la méthode de synthèse vocale complète, sans discrimination de compétences musicales ou linguistiques. Le test, intitulé « *Corsican rare singing synthesis: naturalness and comprehensibility assessment* », était proposé en langue anglaise uniquement. Le texte prononcé était donné en début de test. Ensuite, il a été demandé aux sujets d'évaluer douze extraits audio issus de notre base de test ; les extraits audio ont été présentés dans un ordre aléatoire. Pour chacun d'eux, les sujets devaient noter la naturalité du son ainsi que la compréhensibilité. Dans le questionnaire, nous

avons défini la naturalité comme la qualité d'un son à correspondre aux standards d'un auditeur en termes de prosodie, intonation, rythme et accents. La compréhensibilité quant à elle est définie comme le degré avec lequel un sujet auquel on fournit des informations additionnelles à propos de ce que le locuteur prononce reconnaît les phonèmes prononcés dans l'extrait audio. Ces deux critères sont évalués par une note comprise entre 1 et 5, le score de 1 correspondant à une faible qualité et 5 une bonne qualité. Il a également été demandé aux participants d'indiquer s'ils avaient au moins trois ans d'expérience dans les domaines liés à l'audio et à la voix ou non.

Il apparaît qu'un peu plus de la moitié (54,5 %) des 83 répondants au questionnaire peuvent être considérés comme des experts en audio. Nous avons donc un équilibre relatif entre les sujets experts et les sujets non-experts. Les scores donnés par ces 83 sujets sont rassemblés dans le Tableau 17.

La Figure 66 présente le score de naturalité moyen pour chaque type de synthèse, ainsi que leurs écarts statistiques (test de rang de signe de Wilcoxon, voir Tableau 18). Pour chaque type de source, les LSF d'origine obtiennent de meilleurs résultats que les LSF prédits. La méthode de synthèse utilisant l'autoencodeur multimodal obtient systématiquement de meilleurs résultats que la synthèse utilisant *EigenLips* et *EigenTongues*. Le bruit utilisé comme signal de source obtient naturellement de moins bons scores que les autres signaux de sources. Un résultat remarquable est que l'autoencodeur obtient des résultats de naturalité stables quel que soit le signal de source utilisé, mis à part le signal de bruit. Les écarts de naturalité entre les LSF d'origine et les LSF prédits en utilisant l'autoencodeur sont marginaux pour les sources dEGG et bruit blanc, non significatifs pour l'ODGD, et significatifs pour les résidus. En comparaison, les LSF estimés en utilisant les EigenLips et EigenTongues sont très significativement moins naturels qu'avec les deux autres méthodes, quel que soit le signal de source employé.

Tableau 17 - Scores de naturalité et d'intelligibilité en fonction du type de signal d'excitation et de l'origine des valeurs des LSF.

Signal d'excitation	**Origine des LSF**	**Score de compréhensibilité (μ±σ)**	**Score de naturalité (μ±σ)**
Résidus LPC	Signal acoustique	3,8±1,0	3,6±1,1
Résidus LPC	Autoencodeur multimodal	2,9±1,0	3,2±1,0
Résidus LPC	*EigenLips* et *EigenTongues*	2,1±1,0	2,6±1,0
ODGD	Signal acoustique	3,1±1,1	3,0±1,0
ODGD	Autoencodeur multimodal	3,0±1,2	3,0±1,0
ODGD	*EigenLips* et *EigenTongues*	1,7±1,0	2,1±1,0
dEGG	Signal acoustique	3,2±1,1	3,0±1,1
dEGG	Autoencodeur multimodal	2,6±1,1	2,8±1,1
dEGG	*EigenLips* et *EigenTongues*	1,8±0,8	2,2±0,9
Bruit blanc	Signal acoustique	2,0±1,0	1,3±0,5
Bruit blanc	Autoencodeur multimodal	1,6±0,8	1,2±0,4
Bruit blanc	*EigenLips* et *EigenTongues*	1,1±0,3	1,0±0,3

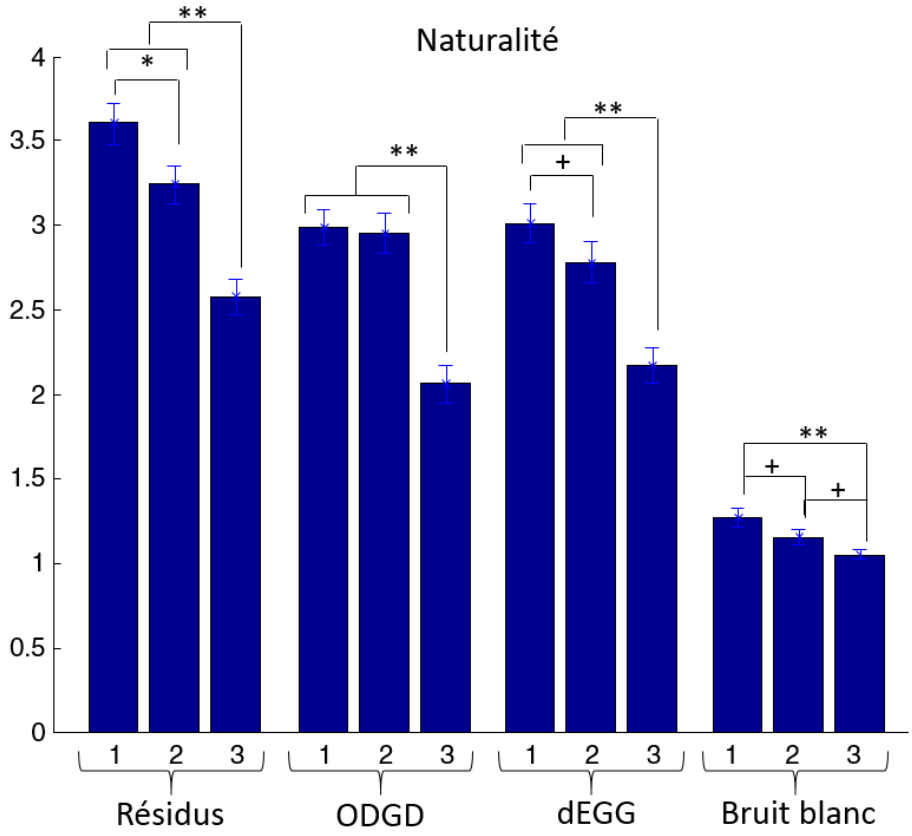


Figure 66 - Evaluation de la naturalité en fonction du type de source et de l'origine du calcul des LSF. La valeur 1 représente les LSF calculés à partir du signal acoustique, la valeur 2 représente les LSF estimés en utilisant l'autoencodeur profond et la valeur 3 représente les LSF estimés en utilisant les EigenLips *et* EigenTongues. *L'erreur type de la moyenne est représentée sur le haut de chaque barre du diagramme. +écarts marginalement significatifs, * écarts significatifs, **écarts très significatifs, voir aussi Tableau 18.*

Tableau 18 – Test de rang de signes de Wilcoxon pour la naturalité (test apparié). Les conditions suivantes sont comparées deux à deux : (1) LSF d'origine, (2) LSF prédits par autoencodeur multimodal et (3) LSF prédits par EigenLips + EigenTongues. *Les seuils marginalement significatifs (+, $p = 3{,}3.10^{-2}$), significatifs (*, $p = 1{,}7.10^{-2}$) et très significatifs (**, $p = 3{,}3.10^{-3}$) tiennent compte d'une correction de Šidák.*

Source	**(1) *vs.* (2)**	**(1) *vs.* (3)**	**(2) vs. (3)**
Résidus LPC	$4{,}4.10^{-3}{}_{*}$	$2{,}85.10^{-11}{}_{**}$	$1{,}83.10^{-6}{}_{**}$
ODGD	$7{,}1.10^{-1}$	$4{,}10.10^{-9}{}_{**}$	$1{,}54.10^{-8}{}_{**}$
dEGG	$3{,}1.10^{-2}{}_{+}$	$1{,}50.10^{-9}{}_{**}$	$3{,}05.10^{-7}{}_{**}$
Bruit	$3{,}1.10^{-2}{}_{+}$	$4{,}67.10^{-4}{}_{**}$	$2{,}0.10^{-2}{}_{+}$

La Figure 67 montre les résultats de compréhensibilité, ainsi que leurs écarts statistiques d'après le test de rang de signe de Wilcoxon (voir Tableau 19). Il s'agit d'un test de rangs sur échantillons appariés ; il permet de comparer deux mesures d'une variable quantitative

effectuées sur les mêmes sujets, même si la variable quantitative ne suit pas une distribution normale ou qu'il n'y a pas égalité des variances dans les deux groupes, contrairement au test de Student. D'après ce test, ces écarts d'évaluation entre les différentes origines des LSF sont davantage marqués en compréhensibilité qu'en naturalité. Par ailleurs, il est à noter que la synthèse utilisant l'onde de débit glottique dérivée présente une compréhensibilité plus homogène que les autres sources, et en particulier qu'elle tend à diminuer l'écart de score entre LSF d'origine et LSF prédits avec l'autoencodeur multimodal. Cette synthèse ne présente pas d'écarts significatifs comparée à celle obtenue avec les LSF d'origine ($p > 0.10$), alors que les synthèses obtenues avec les trois autres sources (résidus, dEGG et bruit blanc) sont toutes très significativement moins naturelles pour les LSF prédits avec l'autoencodeur multimodal. En comparaison, les LSF estimés en utilisant les EigenLips et EigenTongues sont très significativement moins compréhensibles qu'avec les deux autres méthodes, quel que soit le signal de source employé.

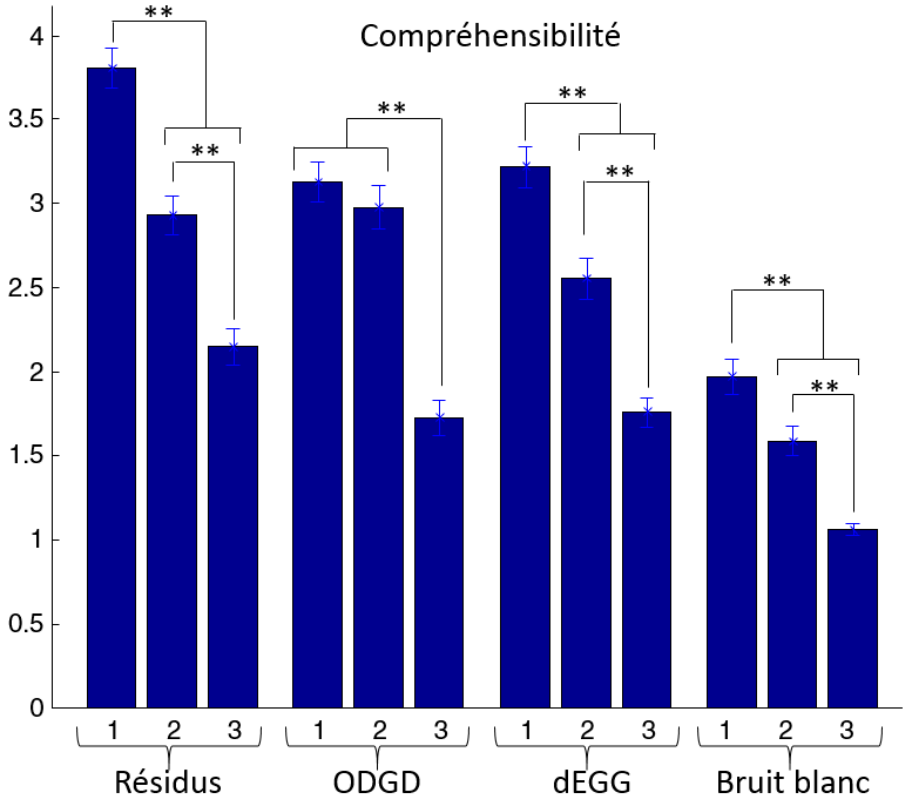


Figure 67 - Evaluation de la compréhensibilité en fonction du type de source et de l'origine du calcul des LSF. La valeur 1 représente les LSF calculés à partir du signal acoustique, la valeur 2 représente les LSF estimés en utilisant l'autoencodeur profond et la valeur 3 représente les LSF estimés en utilisant les EigenLips *et* EigenTongues*. . L'erreur type de la moyenne est représentée sur le haut de chaque barre du diagramme. **écarts très significatifs, voir aussi Tableau 19.*

Tableau 19 – Test de rang de signes de Wilcoxon pour la compréhensibilité (test apparié). Les conditions suivantes sont comparées deux à deux : (1) LSF d'origine, (2) LSF prédits par autoencodeur multimodal et (3) LSF prédits par EigenLips + EigenTongues. *Les seuils non significatifs ($p >> 3{,}3.10^{-2}$) et très significatifs (**, $p = 3{,}3.10^{-3}$) tiennent compte d'une correction de Šidák.*

Source	(1) *vs.* (2)	(1) *vs.* (3)	(2) vs. (3)
Résidus LPC	$5{,}0.10^{-9}$**	$1{,}05.10^{-13}$**	$5{,}71.10^{-7}$**
ODGD	$1{,}4.10^{-1}$	$3{,}12.10^{-11}$**	$9{,}72.10^{-11}$**
dEGG	$1{,}30.10^{-7}$**	$4{,}19.10^{-13}$**	$2{,}29.10^{-8}$**
Bruit	$1{,}20.10^{-4}$**	$7{,}25.10^{-10}$**	$1{,}40.10^{-7}$**

3.8 Discussion

Nous avons présenté dans ce chapitre un modèle permettant d'effectuer une synthèse vocale à partir d'images de la langue et des lèvres et les signaux glottiques. Cette synthèse vise des objectifs de compréhension dans le contexte de la préservation des techniques de chant. Dans [129], l'auteur discute de l'utilisation de technologies de l'information afin d'apporter un retour visuel à l'utilisateur qui souhaite améliorer ses techniques de chant. Parmi les retours proposés, la plupart sont des descripteurs acoustiques de la production vocale. Dans [129], l'auteur montre que la fréquence centrale des formants dépend du mouvement des articulateurs mais aussi de l'entraînement du chanteur et de sa façon de projeter ou non sa voix à la manière d'un chanteur lyrique. Puisque les LSF sont directement liés aux coefficients de prédiction LPC, qui ont un lien avec la position des formants, l'utilisation des LSF comme marqueurs de l'articulation du chanteur semble justifiée. Les éléments que l'auteur estime utiles pour l'amélioration du geste vocal sont la qualité vocale, les consonnes, la qualité et la durée des voyelles, l'élargissement du larynx, le chanter legato ou staccato, le registre, les résonances, la précision de la justesse, la position du larynx, la position de la langue pour la production des voyelles, la position de la mâchoire pour la hauteur, l'alignement entre le cou et la tête, la posture générale et la respiration. Nous utilisons pour notre part des données électroglottographiques qui contiennent les informations de qualité vocale, de rythme et de hauteur. Nous utilisons également des images permettant d'accéder aux informations du mouvement des articulateurs.

Nous avons développé dans un premier temps une méthode qui démontre une application possible de notre modèle pour des utilisations pédagogiques. Ce système permet de

synthétiser à partir d'un couple d'images de la langue et les lèvres le son correspondant. Il permet donc d'illustrer le rapport entre les positions des lèvres et de la langue d'une part, et le son produit d'autre part.

Nous avons également proposé un modèle permettant une synthèse vocale complète d'un extrait de chant en combinant des images des articulateurs. La synthèse vocale, et en particulier la synthèse de voix chantée, pose plusieurs difficultés [130]. Dans les approches classiques, il est difficile de synthétiser un extrait de voix chantée en gérant le texte, la hauteur des notes et le rythme. Notre approche de synthèse propose une méthode permettant de fournir au système les informations concernant le rythme, le texte et la hauteur de façon automatique, sans utiliser de partitions ou codage de la musique. Les performances de notre modèle de synthèse ont été validées sur deux bases de données : une base de voyelles isolées, ainsi qu'une base avec silences et fricatives, sur des chants entiers. Dans les deux cas notre méthode s'est avérée plus fiable qu'un modèle linéaire, ce qui semble confirmer que l'apprentissage profond semble bien adapté pour extraire les informations multimodales de la voix chantée. On s'attendait effectivement à ce qu'un modèle non linéaire reflète mieux le fonctionnement du conduit vocal (rappelons que le conduit vocal est modélisé avec un filtre d'ordre 12 dans le modèle LPC).

Dans [131], l'objectif est de proposer une méthode de synthèse vocale à partir de données acoustiques pour le doublage en imitant la personnalité vocale d'un locuteur. Il s'agit d'utiliser la voix d'un doubleur professionnel et d'en modifier les propriétés afin de reconstituer artificiellement la voix d'un autre locuteur. Ces modifications concernent uniquement le domaine acoustique et aucune technique d'imagerie n'est utilisée. A la fin de cette étude, un test perceptif mené sur 11 sujets, 5 experts dans le domaine du traitement audio et 6 non-experts. Ces tests perceptifs montrent des évaluations plutôt positives de la part des sujets. Dans les évaluations perceptives que nous avons menées, réalisées sur un nombre bien plus significatif de sujets, les extraits audio ont été évalués avec des scores un peu plus bas. Cette différence peut provenir du nombre de répondants à l'étude, de la langue, du type de phonation et également du type de synthèse, purement acoustique ou bien visuo-acoustique.

La prédiction des LSF nous a permis d'effectuer une synthèse vocale de la voix chantée. Les LSF sont cependant des descripteurs de bas niveau de la voix. Nous pourrions envisager d'étendre ce modèle pour détecter des représentations de plus haut niveau, telles que par exemple la détection de phonèmes pour une synthèse plus précise.

Conclusion générale et perspectives

Ce travail propose différentes méthodes d'utilisation de données visuelles et acoustiques pour la modélisation des interprétations vocales et le développement d'outils pour la compréhension du geste vocal. La collecte et l'utilisation de données articulatoires permettent d'apporter des informations techniques à une personne désireuse d'améliorer sa technique vocale. Néanmoins, les données brutes ainsi collectées n'étant pas très lisibles directement, il est important d'être en mesure de proposer un contenu plus pertinent à un utilisateur non expert des technologies d'imagerie. Ce travail de thèse combine des notions d'imagerie biomédicale, de traitement du signal, d'apprentissage statistique, d'acoustique et de phonétique. Afin de proposer un premier décodage des informations articulatoires, nous présentons une technique d'extraction automatique du contour de la langue utilisant un modèle d'apprentissage statistique, dont les performances restent stables sur plusieurs minutes d'enregistrement. Nous proposons ensuite une méthode de synthèse vocale utilisant des données articulatoires et glottiques seulement. Cette méthode a pour but d'étudier l'importance de l'articulation. Nous proposons des approches de développement d'outils utilisant l'apprentissage statistique afin de modéliser le lien entre l'articulation et la production vocale. L'utilisation d'un casque embarquant plusieurs sortes de capteurs nous a permis de constituer des bases de données de voix parlée et de voix chantée. En parole comme en chant, nous avons collecté des données correspondant à des voyelles isolées soutenues, des associations consonnes-voyelles, ainsi que des textes de chants traditionnels ou rares.

Nous avions fait l'hypothèse qu'un apprentissage sur une base de données conséquente fournirait une information suffisante pour modéliser la position de la langue et extraire les paramètres du conduit vocal. Ces données nous ont permis de développer une méthode d'extraction automatique du contour de la langue sur des images échographiques. L'apprentissage profond nous a permis d'extraire des descripteurs à partir d'images échographiques, en se fondant sur une base d'apprentissage dont les contours ont été extraits automatiquement. Notre méthode, qui obtient des performances comparables à des étiquetages manuels et aux outils proposés dans la littérature, a plusieurs avantages. En effet, notre système, dont les performances dépendent seulement de la qualité des données d'entrée et de l'efficacité de l'apprentissage automatique, ne requiert pas d'initialisation manuelle. De plus, notre méthode permet d'extraire le contour d'un grand nombre d'images quelle que soit

la longueur de la séquence d'images : elle permet d'extraire le contour sur des séquences d'images aussi bien que sur des images sélectionnées de façon isolée.

Dans un second temps, nous avons tenté de reconstruire un modèle acoustique du conduit vocal. Notre approche consiste à combiner des informations articulatoires issues des images de la langue et des lèvres avec des informations glottiques. Nous avons mis en œuvre ce modèle pour tester les possibilités de synthèse vocale articulatoire. Pour cela, il est important d'extraire des descripteurs permettant d'établir un lien entre les données articulatoires et le signal acoustique. Nous utilisons des outils d'apprentissage statistique qui permettent de repérer des informations dans les images entières. Les informations articulatoires permettent d'estimer les paramètres d'un filtre et les informations glottiques permettent de construire un signal d'excitation. Nous opérons une distinction entre les trames non voisées, que nous excitons par un bruit, et les trames voisées. Cette distinction nous permet de synthétiser l'ensemble des phonèmes, en voix parlée comme en voix chantée. Nous pouvons utiliser comme signal d'excitation un signal purement synthétique, en variant la fréquence fondamentale et en imposant les autres paramètres de source. Nous pouvons également extraire des informations du signal électroglottographique qui nous permettent de synthétiser un signal d'onde de débit glottique. La combinaison du signal d'excitation et du filtre nous permet effectivement de synthétiser des extraits de voix chantée.

Nous extrayons des descripteurs des couples langue-lèvres qui permettent de prédire l'allure du spectre du filtre du conduit vocal. Nous avons utilisé pour ce faire une méthode linéaire fondée sur les principes de l'analyse en composantes principales, ainsi qu'une méthode non linéaire impliquant un autoencodeur profond et multimodal. Nous avions fait l'hypothèse que la méthode non linéaire exploiterait mieux la relation complexe entre les images des articulateurs et le signal glottique que la méthode linéaire. Nous avons effectivement obtenu de meilleures performances avec l'apprentissage profond qu'avec la méthode linéaire : la distorsion spectrale obtenue en validation sur une base de voyelles isolées passe de 3,0 dB pour la méthode linéaire à 2,2 dB pour notre méthode (un résultat proche de la distorsion transparente). En conditions réelles, avec silences et fricatives, sur des chants entiers, la distorsion spectrale obtenue en validation passe de 5,2 dB pour la méthode linéaire à 4,3 dB pour notre méthode. La naturalité et l'intelligibilité des signaux acoustiques reconstruits ont en outre été évalués par un test perceptif qui confirme ces résultats.

Notre étude a démontré la faisabilité d'une modélisation multimodale des mécanismes du chant. Il reste cependant des pistes qui pourraient être explorées pour améliorer ce modèle. Par exemple, dans notre approche, nous construisons un modèle de filtre en utilisant uniquement les données issues de la langue et des lèvres. L'ajout de la composante nasale permettrait peut-être d'améliorer les performances de prédiction du filtre vocal. Pour cela, il serait intéressant de passer d'une architecture bimodale à une architecture à trois modalités en combinant les images de la langue, des lèvres, ainsi qu'une carte temps-fréquence construite à partir de l'enregistrement effectué par le capteur piézoélectrique fixé sur le nez du chanteur. Notre méthode pourrait être étendue à une telle architecture en suivant le même principe que pour l'architecture bimodale proposée.

4 Références

[1] J. L. Preston, N. Brick et N. Landi, «Ultrasound biofeedback treatment for persisting childhood apraxia of speech,» *American Journal of Speech-Language Pathology,* vol. 22, n° %14, pp. 627-643, 2013.

[2] P. Bacsfalvi et B. M. Bernhardt, «Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography,» *Clinical Linguistics and Phonetics,* vol. 25, n° %111-12, pp. 1034-1043, 2011.

[3] J. Cleland, J. M. Scobbie et A. A. Wrench, «Using ultrasound visual biofeedback to treat persistent primary speech sound disorders,» *Clinical Linguistics and Phonetics,* vol. 29, n° %18-10, pp. 575-597, 2015.

[4] D. Massaro et J. Light, «Using visible speech for training perception and production of speech for hard of hearing individuals,» *Journal of Speech, Language, and Hearing Research,* vol. 47, pp. 304-320, 2004.

[5] O. Engwall, "Augmented Reality Talking Heads as a Support for Speech Perception and Production," in *Augmented Reality - Some Emerging Application Areas*, Rijeka, Intech, 2011, pp. 89-116.

[6] S. Fagel and C. Clemens, "An articulation model for audiovisual speech synthesis—Determination, adjustment, evaluation," *Speech Communication,* vol. 44, no. 1, pp. 141-154, 2004.

[7] D. Fabre, T. Huber and P. Badin, "Automatic animation of an articulatory tongue model from ultrasound images," in *Proceedings of Interspeech*, Singapore, 2014.

[8] S. Lamesch, «Mécanismes laryngés et voyelles en voix chantée, Dynamique vocale, phonétogrammes de paramètres glottiques et spectraux, transitions de mécanismes,» Thèse de doctorat de l'université Pierre et Marie Curie, 2010.

[9] H. V. Carter et H. Gray, Anatomy of the Human Body, Lea and Febiger: Philadelphia, 1918.

[10] F. Legent, L. Perlemutier et C. Vandebrouck, Cahiers d'anatomie O.R.L, Masson, 1975.

[11] E. Bianco, «Notes pour le cours de voix chantée - "Souffle ou résonnance",» CNSM, 2002. [En ligne]. Available: http://www.revoice.fr/Pages/COURSVOIX.aspx.

[12] S. Maeda, "The role of the sinus cavities in the production of nasal vowels," in

Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82, Paris, 1982.

[13] G. Lindsay, «Speech Talk,» 2013. [En ligne]. Available: http://englishspeechservices.com/blog/the-vowel-space/.

[14] I. P. Association, «IPA Home,» [En ligne]. Available: https://www.internationalphoneticassociation.org/redirected_home.

[15] J. Laver, Principles in phonetics, Cambridge University Press, 1994.

[16] C. d'Alessandro, "Voice Source Parameters and Prosodic Analysis," in *Methods in Empirical Prosody Research*, (Language, Context and Cognition; Vol. 3), Stefan Sudhoff; Denisa Lenertova; Roland Meyer; Sandra Pappert; Petra Augurzky; Ina Mleinek; Nicole Richter; Johannes Schliesser. Mouton de Gruyter, 2006, pp. 63-87.

[17] N. Henrich, Thèse de doctorat de l'Université Paris 6, 2001.

[18] F. Le Huche et A. Attali, La voix. Anatomie et physiologie des organes de la voix et de la parole, Paris: Masson, 1991.

[19] «i-Treasures : capturing the intagible,» [En ligne]. Available: http://i-treasures.eu/.

[20] L. Crevier-Buchman, A. Amelot, S. K. Al Kork, M. Adda-Decker, N. Audibert, P. Chawah, B. Denby, T. Fux, A. Jaumard-Hakoun, P. Roussel, M. Stone, J. Vaissière, K. Xu et C. Pillot-Loiseau, «Acoustic Data Analysis from Multi-Sensor Capture in Rare Singing: Cantu in Paghjella Case Study,» *International Journal of Heritage in the Digital Era*, vol. 4, n° %11, pp. 121-132, 2015.

[21] L. Bailly, N. Henrich, X. Pelorson and J. Gilbert, "Vocal folds and ventricular bands in interaction: comparison between in-vivo measurements and theoretical predictions," in *155th Meeting of Acoustical Society of America, Acoustics'08*, Paris, 2008.

[22] N. Henrich, B. Lortat-Jacob, M. Castellengo, L. Bailly and X. Pelorson, "Period-doubling occurences in singing: the "bassu" case in traditional Sardinian "A Tenore" singing," in *International Conference on Voice Physiology and Biomechanics*, Tokyo, 2006.

[23] B. Lortat-Jacob, «Chants de passion Au coeur d'une confréerie de Sardaigne,» 1998.

[24] M. Proctor, E. Bresch, D. Byrd, K. Nayak et S. Narayanan, «Paralinguistic mechanisms of production in human “beatboxing”: A real-time magnetic resonance imaging study,» *Journal of the Acoustical Society of America*, vol. 133, n° %12, pp. 1043-1054, 2013.

[25] T. De Torcy, A. Clouet, C. Pillot-Loiseau, J. Vaissière, D. Brasnu et L. Crevier-Buchman, «A video-fiberscopic study of laryngopharyngeal behaviour in the human beatbox,» *Logopedics Phoniatrics Vocolog,* vol. 39, n° %11, pp. 38-48, 2013.

[26] E. Moulines et F. Charpentier, «Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,» *Speech Communication,* vol. 9, pp. 453-467, 1990.

[27] R. McAuley et T. F. Quatieri, «Speech analysis/synthesis based on a sinusoidal representation,» *IEEE Transactions on Speech and Signal Processing,* vol. 34, n° %14, pp. 744-754, 1986.

[28] B. Kedem, «Benjamin Kedem. Spectral analysis and discrimination by zero-crossings,» *Proceedings of the IEEE,* vol. 74, n° %111, p. :1477–1493, 1986.

[29] C. Roads, The Computer Music Tutorial, Cambridge: MIT Press, 1996.

[30] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette and P. Depalle, "Features extraction and temporal segmentation of acoustic signals," in *International Computer Music Conference*, Ann Arbor, 1998.

[31] L. Rabiner, «On the use of autocorrelation analysis for pitch detection,» *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 25, n° %11, pp. 24-33, 1977.

[32] M. Ross, H. Shaffer, A. Cohen, R. Freudberg et H. Manley, «Average magnitude difference function pitch extractor,» *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 22, n° %15, pp. 353-362, 1974.

[33] M. R. P. Thomas et P. A. Naylor, «The SIGMA Algorithm: A Glottal Activity Detector for Electroglottographic Signals,» *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 17, n° %18, pp. 1557-1566, 2009.

[34] M. R. Every et J. E. Szymanski, «Separation of synchronous pitched notes by spectral filtering of harmonics,» *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, n° %15, pp. 1845-1856, 2006.

[35] T. Hézard, «Production de la voix : exploration, modèles et analyse/synthèse.,» Université Pierre et Marie Curie, Paris, 2013.

[36] G. Fant, Acoustic Theory of Speech Production, Hague: Mouton, 1960.

[37] G. Kang and L. Fransen, "Application of Line-Spectrum Pairs to Low-Bit-Rate Speech

Encoders," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP85*, Tampa, 1985.

[38] B. Doval, C. D'Alessandro et N. Henrich, «The Spectrum of Glottal Flow Models,» *Acta Acustica united with Acustica,* vol. 92, n° %16, pp. 1026-1046, 2006.

[39] G. Fant, J. Liljencrants et Q. Lin, «A four-parameter model of glottal flow,» *STL-QPSR,* vol. 26, n° %14, pp. 1-13, 1985.

[40] P. Alku, «Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering,» *Speech Communication,* vol. 11, n° %12-3, pp. 109-118, 1992.

[41] J. Sundberg, «The kth synthesis of singing,» *Advances in cognitive Psychology,* vol. 2, n° %12-3, pp. 131-143, 2006.

[42] T. Dubuisson, «Glottal Source Estimation and Automatic Detection of Dysphonic Speakers,» University of Mons, 2011.

[43] D. H. Klatt et L. C. Klatt, «Analysis, synthesis, and perception of voice quality variations among female and male talkers,» *The Journal of the Acoustical Society of America,* vol. 87, n° %12, pp. 820-857, 1990.

[44] R. Veldhuis, «A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation,» *The Journal of the Acoustical Society of America,* vol. 103, n° %11, pp. 566-571, 1998.

[45] A. E. Rosenberg, «Effect of Glottal Pulse Shape on the Quality of Natural Vowels,» *The Journal of the Acoustical Society of America,* vol. 49, n° %12B, pp. 583-590, 1971.

[46] B. Doval, C. D'Alessandro and N. Henrich, "The voice source as a causal/anticausal linear filter," in *Voice Quality : Functions, Analysis and Synthesis VOQUAL'03*, Geneva, 2003.

[47] L. Feugère, «Synthèse par règles de la voix chantée contrôlée par le geste et applications musicales,» Université Pierre et Marie Curie, Paris, 2013.

[48] D. H. Klatt, «Review of text-to-speech conversion for english,» *Journal of the Acoustical Society of America,* vol. 82, n° %13, pp. 737-793, 1987.

[49] X. Rodet, Y. Potard et J.-B. Barrière, «The chant project : From the synthesis of the singing voice to synthesis in general,» *Computer Music Journal,* vol. 8, n° %13, pp. 15-31, 1984.

[50] C. Hamon, E. Moulines and F. Charpentier, "A diphone synthesis system based on

time-domain prosodic modifications of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, 1989.

[51] H. Zen, K. Tokuda et A. W. Black, «Statistical parametric speech synthesis,» *Speech Communication*, vol. 51, n° %111, pp. 1039-1064, 2009.

[52] S. King, "A reading list of recent advances in speech synthesis," in *International Congress of Phonetic Science*, Glasgow, 2015.

[53] I. Titze, «The human vocal cords : a mathematical model,» *Phonetica*, vol. 28, pp. 129-170, 1973.

[54] C. H. Coker, "Speech synthesis with a parametric articulatory model," in *Proc. Speech. Symp*, Kyoto, 1968.

[55] P. Birkholz, "Articulatory synthesis of singing," in *Interspeech*, Antwerp, 2007.

[56] E. Saltzman, «Task dynamic coordination of the speech articulators : A preliminary model,» *Experimental Brain Research*, vol. 15, pp. 129-144, 1986.

[57] H. Kawahara, I. Masuda-Katsuse et A. de Cheveigné, «Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds.,» *Speech Communications*, vol. 27, n° %13-4, p. 187–207, 1999.

[58] J. LaRoche, Y. Stylianou and E. Moulines, "Hnm: a simple, efficient harmonic+noise model for speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, 1993.

[59] T. Saitou, M. Goto, M. Unoki and M. Akagi, "Vocal Conversion from Speaking Voice to Singing Voice Using STRAIGHT," in *Proceedings of Interspeech*, Antwerp, 2007.

[60] A. Roebel and J. Fineberg, "Speech to chant transformation with the phase vocoder," in *Proceedings of Interspeech*, Antwerp, 2007.

[61] H. Kenmochi and H. Ohshita , "VOCALOID – Commercial singing synthesizer based on sample concatenation," in *Proceedings of Interspeech*, Antwerp, 2007.

[62] S. Ternström and J. Sundberg, "Formant-based synthesis of singing," in *Proceedings of Interspeech*, Antwerp, 2007.

[63] N. D'Alessandro, B. Doval, C. d'Alessandro, S. Le Beux, P. Woodruff, Y. Fabre and T. Dutoit, "RAMCESS: Realtime and Accurate Musical Control of Expression in Singing Synthesis," *Journal on Multimodal User Interfaces*, vol. 1, no. 1, pp. 31-39, 2007.

[64] L. Feugère, S. Le Beux and C. d'Alessandro, "Chorus Digitalis : polyphonic gestural singing," in *Proceedings of the 1st International Workshop on Performative Speech and Singing Synthesis*, Vancouver, 2011.

[65] P. Birkholz, "Articulatory Synthesis of Singing," in *Proceedings of Interspeech*, Antwerp, 2007.

[66] J. Bonada, M. Umbert and M. Blaauw, "Expressive Singing Synthesis Based on Unit Selection for the Singing Synthesis Challenge 2016," in *Proceedings of Interspeech*, San Francisco, 2016.

[67] M. Umbert, J. Bonada, M. Goto and J. Sundberg, "Expression Control in Singing Voice Synthesis: Features, approaches, evaluation, and challenges," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 55-73, 2015.

[68] G. Dreyfus, Apprentissage statistique, Paris: Eyrolles, 2008.

[69] Y. Bennani et P. Gallinari, «Neural networks for discrimination and modelization of speakers,» *Speech communication*, vol. 17, pp. 159-175, 1995.

[70] L. Deng et D. Yu, Deep Learning: Methods and Applications, Now Publishers, 2014.

[71] Y. Bengio, A. Courville et P. Vincent, «Representation Learning: A Review and New Perspectives,» *EEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, n° %18, pp. 1798-828, 2013.

[72] G. E. Hinton et S. Osindero, «A fast learning algorithm for deep belief nets,» *Neural Computation*, vol. 18, 2006.

[73] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent et s. Bengio, «Why Does Unsupervised Pre-training Help Deep Learning,» *Journal of Machine Learning Research*, vol. 11, pp. 625--660, 2010.

[74] D. Yu et L. Deng, «Deep Learning and Its Applications to Signal and Information Processing,» *IEEE Signal Processing Magazine*, vol. 28, n° %11, pp. 245-254, 2011.

[75] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks," in *Twenty-First Annual Conference on Neural Information Processing Systems*, Vancouver, 2007.

[76] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in Boltzmann machines," in *Parallel distributed processing: explorations in the microstructure of cognition*, Cambridge, MIT Press Cambridge, 1986, pp. 282-317.

[77] D. H. Ackley, G. E. Hinton et T. J. Sejnowski, «A Learning Algorithm for Boltzmann Machines,» *Cognitive Science,* vol. 9, n° %11, p. 147–169, 1985.

[78] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Clearwater, 2009.

[79] G. E. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," in *Neural Networks: Tricks of the Trade: Second Edition*, Toronto, Springer Berlin Heidelberg, 2012, pp. 599-619.

[80] Y. Bengio, «Learning Deep Architectures for AI,» *Foundations and Trends in Machine Learning,* vol. 2, n° %11, pp. 1-127, 2009.

[81] Y. Bengio et O. Delalleau, «Justifying and Generalizing Contrastive Divergence,» *Neural Computation,* vol. 21, n° %16, pp. 1601-1621, 2009.

[82] P. Baldi, "Autoencoders, Unsupervised Learning, and Deep Architectures," in *Unsupervised and Transfer Learning - Workshop held at ICML 2011*, Bellevue, 2011.

[83] L. Arnold, S. Rebecchi, S. Chevallier and H. Paugam-Moisy, "An introduction to deep-learning," in *Advances in Computational Intelligence and Machine Learning, ESANN'2011*, 2011.

[84] A. Marchal et C. Cavé, L'imagerie médicale pour l'étude de la parole, Traité IC2, série Cognition et traitement de l'information, 2009.

[85] M. Stone, «A Guide to Analysing Tongue Motion from Ultrasound Images,» *Clinical Linguistics and Phonetics,* vol. 19, n° %16-7, pp. 455-502, 2005.

[86] T. Hueber, G. Chollet, B. Denby, M. Stone and L. Zouari, "Ouisper: Corpus Based Synthesis Driven by Articulatory Data," in *International Congress on Phonetic Science (ICPhS)*, Saarbrücken, 2007.

[87] T. Hueber, G. Chollet and B. Denby, "Ultraspeech, a portable system for acquisition of high-speed ultrasound, video and acoustic speech data," in *Ultrafest V*, New Haven, 2010.

[88] T. Hueber, G. Chollet, B. Denby and M. Stone, "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," in *Proceedings of the International Seminar on Speech Production*, Strasbourg, 2008.

[89] B. Denby, Y. Oussar, G. Dreyfus and M. Stone, "Prospects for a Silent Speech Interface Using Ultrasound Imaging," in *IEEE International Conference on Acoustics, Speech,*

and Signal Processing, Toulouse, France, 2006.

[90] B. Denby, J. Cai, T. Hueber, P. Roussel, G. Dreyfus, L. Crevier-Buchman, C. Pillot-Loiseau, G. Chollet, S. Manitsaris and M. Stone, "Towards a Practical Silent Speech Interface Based on Vocal Tract Imaging," in *International Seminar on Speech Production 2011*, Montreal, 2011.

[91] T. Hueber, E.-L. Benaroya, B. Denby and G. Chollet, "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-based Silent Speech Interface," in *Interspeech 2011*, Florence, 2011.

[92] S. K. Al Kork, A. Jaumard-Hakoun, M. Adda-Decker, A. Amelot, L. Buchman, P. Chawah, G. Dreyfus, T. Fux, C. Pillot-Loiseau, P. Roussel, M. Stone, K. Xu and B. Denby, "A Multi-Sensor Helmet to Capture Rare Singing, an Intangible Cultural Heritage Study," in *Proceedings of 10th International Seminar on Speech*, Cologne, 2014.

[93] P. Chawah, T. Fux, M. Adda-Decker, A. Amelot, N. Audibert, B. Denby, G. Dreyfus, A. Jaumard-Hakoun, C. Pillot-Loiseau, P. Roussel, M. Stone, K. Xu and L. Buchman, "An educational platform to capture, visualize and analyze rare singing," in *ISCA, INTERSPEECH 2014: Show & Tell Contribution*, Singapore, 2014.

[94] M. Stone et E. Davis, «A head and transducer support system for making ultrasound images of tongue/jaw movement,» *The Journal of the Acoustical Society of America*, vol. 98, n° %16, pp. 3107-3112, 1995.

[95] K. Xu, Y. Yang, A. Jaumard-Hakoun, M. Adda-Decker, A. Amelot, S. K. Al Kork, L. Crevier-Buchman, P. Chawah, G. Dreyfus, T. Fux, C. Pillot-Loiseau, P. Roussel, M. Stone and B. Denby, "3D tongue motion visualization based on ultrasound image sequences," in *Interspeech 2014*, Singapore, 2014.

[96] K. Xu, Y. Yang, A. Jaumard-Hakoun, G. Dreyfus, P. Roussel, M. Stone and B. Denby, "Development of a 3D Tongue Motion Visualization Platform Based on Ultrasound Image Sequence," in *Proceeding of 18th International Congress on Phonetic Sciences (ICPhS 15)*, Glasgow, 2015.

[97] A. A. Wrench and P. Balch, "Towards a 3D Tongue model for parameterising ultrasound data," in *Proceeding of 18th International Congress on Phonetic Sciences (ICPhS 15)*, 2015, 2015.

[98] I. Fasel and J. Berry, "Deep Belief Networks for Real-Time Extraction of Tongue

Contours from Ultrasound During Speech," in *2010 20th International Conference on Pattern Recognition*, 2010.

[99] Y. S. Akgul, C. Kambhamettu et M. Stone, «Automatic extraction and tracking of the tongue contours,» *IEEE Transactions on Medical Imaging*, vol. 18, n° %110, pp. 1035-1045, 1999.

[100] M. Li, R. Kambhamettu et M. Stone, «Automatic Contour Tracking in Ultrasound Images».

[101] M. Aron, A. Roussos, M. Berger, E. Kerrien and P. Maragos, "Multimodality acquisition of articulatory data and processing," in *European Conference on Signal Processing*, Lausanne, 2008.

[102] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active appearance models," in *European conference on computer vision*, Freiburg, 1998.

[103] A. Roussos, A. Katsamanis and P. Maragos, "Tongue tracking in ultrasound images with active appearance models," in *In IEEE International Conference on Image Processing*, Cairo, 2009.

[104] K. Xu, Y. Yang, M. Stone, A. Jaumard-Hakoun, C. Leboullenger, G. Dreyfus, P. Roussel et B. Denby, «Robust contour tracking in ultrasound tongue image,» *Clinical Linguistics and Phonetics*, vol. 1, n° %11, pp. 1-31, 2016.

[105] K. Xu, T. G. Csapo, P. Roussel et B. Denby, «A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization,» *The Journal of the Acoustical Society of America*, vol. 139, n° %15, pp. EL154-EL160, 2016.

[106] A. Jaumard-Hakoun, S. K. Al Kork, M. Adda-Decker, A. Amelot, L. Crevier Buchman, G. Dreyfus, T. Fux, P. Roussel, C. Pillot-Loiseau, M. Stone and B. Denby, "Capturing, Analyzing, and Transmitting Intangible Cultural Heritage with the i-Treasures Project," in *Ultrafest VI*, Edinburgh, 2013.

[107] J. Cai, T. Hueber, S. Manitsaris, P. Roussel, L. Crevier-Buchman, M. Stone, C. Pillot-Loiseau, G. Chollet, G. Dreyfus and B. Denby, "Vocal Tract Imaging System for Post-Laryngectomy Voice Replacement," in *International IEEE Instrumentation and Measurement Technology Conference*, Minneapolis, MN, 2013.

[108] A. Jaumard-Hakoun, K. Xu, P. Roussel-Ragot, G. Dreyfus, M. Stone and B. Denby,

"Tongue contour extraction from ultrasound images based on deep neural network," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, 2015.

[109] M. Stone, "Investigating speech articulation," in *The Handbook of Phonetic Sciences*, Chichester, John Wiley & Sons, 2010, pp. 9-38.

[110] T. Hueber, Thèse de doctorat de l'Université Pierre et Marie Curie, Paris, 2009.

[111] S. Manitsaris, F. Xavier, B. Denby, G. Dreyfus and P. Roussel, "An Open Source Speech Synthesis Module for a Visual-Speech Recognition System," in *Acoustics 2012*, Nantes, 2012.

[112] J. Cai, T. Hueber, B. Denby, E.-L. Benaroya, G. Chollet, P. Roussel, G. Dreyfus and L. Crevier-Buchman, "A Visual Speech Recognition System for an Ultrasound-Based Silent Speech Interface," in *International Congress on Phonetic Science*, Hong Kong, 2011.

[113] T. Hueber, E. Benaroya, G. Chollet, B. Denby, G. Dreyfus and M. Stone, "Visuo-Phonetic Decoding using Multi-Stream and Context-Dependent Models for an Ultrasound-based Silent Speech Interface," in *Interspeech*, Brighton, 2009.

[114] T. Hueber, G. Chollet, B. Denby, G. Dreyfus and M. Stone, "Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips," in *Interspeech*, Anvers, 2007.

[115] T. Hueber, G. Chollet, B. Denby, G. Dreyfus and M. Stone, "Phone Recognition from Ultrasound and Optical Video Sequences for a Silent Speech Interface," in *Interspeech*, Brisbane, 2008.

[116] J. D. Markel et A. H. Gray, Linear Prediction of Speech, Berlin: Springer Verlag, 1976.

[117] T. Parsons, "Linear Systems and Transforms," in *Voice and speech processing*, New York, McGraw-Hill, 1986, p. 51.

[118] L. R. Rabiner et B. Gold, Theory and application of digital signal processing, Prentice-Hall: Englewood Cliffs, 1975.

[119] T. Parsons, "Pitch and formant estimation," in *Voice and speech processing*, New York, McGraw-Hill, 1986, pp. 197-198.

[120] K. Al-Naimi, S. Villette and A. Kondoz, "Improved LSF estimation through anti-aliasing filtering," in *IEEE Workshop on Speech Coding Proceedings*, Tsukuba, 2002.

[121] M. Turk et A. Pentland, «Eigenfaces for Recognition,» *J. Cognitive Neuroscience*, vol.

3, n° %11, pp. 71-86, 1991.

[122] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel and M. Stone, "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," in *Proceedings of ICASSP*, Honolulu, USA, 2007.

[123] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio et P.-A. Manzagol, «Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion,» *Journal of Machine Learning Research,* vol. 11, pp. 3371-3408, 2010.

[124] P. Vincent, H. Larochelle, Y. Bengio and P. A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," in *In Proceedings of the 25th International Conference on Machine Learning*, Helsinki, 2008.

[125] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A. Y. Ng, "Multimodal Deep Learning," in *ICML 2011*, Bellevue, Washington, USA, 2011.

[126] N. Srivastava et R. Salakhutdinov, «Multimodal Learning with Deep Boltzmann Machines,» *The Journal of Machine Learning Research,* vol. 15, n° %11, pp. 2949-2980, 2014.

[127] H. Stoppiglia, G. Dreyfus, R. Dubois et Y. Oussar, «Ranking a Random Feature for Variable and Feature Selection,» *Journal of Machine Learning Research,* vol. 3, pp. 1399-1414, 2003.

[128] S. Rein, F. Fitzek et M. Reisslein, «Voice Quality Evaluation in Wirelesss Packet Communication Systems: A Tutorial and Performance Results for ROHC,» *IEEE Wireless Communications*, pp. 60-67, 2005.

[129] D. M. Howard, «Technology For Real-Time Visual Feedback In Singing Lessons,» *Research Studies in Music Education,* vol. 24, n° %11, pp. 40-57, 2005.

[130] X. Rodet, "Synthesis and processing of the singing voice," in *Proceedings of the 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio, MPCA-2002*, Leuven, 2002.

[131] F. Fontana and D. L. Gonzalez, "Advanced LPC techniques of voice regeneration for "Virtual Dubbling"," in *Forum Acusticum*, Budapest, 2005.

[132] A. Dognon, Les ultrasons et leurs applications, Presses universitaires de France, 1953.

[133] Z. Tüske, P. Golik, R. Schlüter and H. Ney, "Acoustic Modeling with Deep Neural

Networks Using Raw Time Signal for LVCSR," in *Interspeech*, Singapore, 2014.

[134] D. d'Alessandro, N. d'Alessandro, B. Doval and S. Le Beux, "Comparing Time- and Spectral-Domain Voice Source Models for Gestural Controlled Voice Instruments," in *Proc. of the International Conference on Voice Physiology and Biomechanics*, Tokyo, 2006.

[135] A. Ferencz, J. Kim, Y.-B. Lee and J.-W. Lee, "Automatic pitch marking and reconstruction of glottal closure instants from noisy and deformed electro-glotto-graph signals," in *8th International Conference on Spoken Language Processing*, Jeju Island, 2004.

TONGUE CONTOUR EXTRACTION FROM ULTRASOUND IMAGES BASED ON DEEP NEURAL NETWORK

Aurore Jaumard-Hakoun[1,2], Kele Xu[1,2*], Pierre Roussel-Ragot[2*], Gérard Dreyfus[2], Maureen Stone[3], Bruce Denby[1,2*]*

[1]Université Pierre et Marie Curie, Paris, France
[2] SIGnal processing and MAchine learning Lab, ESPCI ParisTech, PSL Research University, Paris, France
[3]Vocal Tract Visualization Lab, University of Maryland Dental School, Baltimore, USA
*Present affiliation: Institut Langevin, ESPCI ParisTech, PSL Research University, Paris, France
aurore.hakoun@espci.fr, denby@ieee.org

ABSTRACT

Studying tongue motion during speech using ultrasound is a standard procedure, however automatic ultrasound image labelling remains a challenge, as standard tongue shape extraction methods typically require human intervention. This article presents a method based on deep neural networks to automatically extract tongue contours from speech ultrasound images. We use a deep autoencoder trained to learn the relationship between an image and its related contour, so that the model is able to automatically reconstruct contours from the ultrasound image alone. We use an automatic labelling algorithm instead of time-consuming hand-labelling during the training process. We afterwards estimate the performances of both automatic labelling and contour extraction as compared to hand-labelling. Observed results show quality scores comparable to the state of the art.

Keywords: Tongue shape, Medical imaging, Machine learning, Ultrasound

1. INTRODUCTION

Although ultrasound (US) provides a non-invasive and easy to implement tongue imaging method, the presence of multiplicative (Rayleigh) noise makes contour extraction with standard image processing techniques a challenge. Currently, most tongue contour extraction algorithms augment raw image data with a priori knowledge based on the physics of tongue movement. Simple examples require that contours found in a given frame be spatially "smooth" or forbid abrupt changes in contour shape between consecutive frames.

In [1], it has been shown that a deep neural network architecture is able to learn the contour extraction task when trained on hand-labelled contours. In this case, the smoothness criterion arises naturally because hand labelling is guided by a priori knowledge of the class of forms that a human tongue can assume. Hand labelling, however, is time consuming, and, furthermore, does not provide an obvious means of including the second constraint, i.e., that contours extracted from frames nearby in time must be "similar".

In this article, we repeat the procedure of [1], but replace hand-labelled training data with contours extracted by an automatic algorithm that uses block-matching to enforce a crude frame-to-frame similarity condition. This approach allows training data to be obtained in a rapid and relatively painless way, and provides a means of testing whether the deep neural network architecture, which processes only one image at a time, is able nonetheless to embed a priori knowledge corresponding to this additional constraint.

2. METHODS

2.1. Deep Neural Networks and autoencoders

2.1.1. Restricted Boltzmann Machines

The model of Deep Neural Networks proposed in [2] is based on the stacking of Restricted Boltzmann Machines (RBMs). A Restricted Boltzmann Machine is a neural network composed of a layer with visible units and a layer with hidden units, connected through directional links (weights), which are symmetric. The probability of activation of a hidden unit depends on the weighted activations of the units in the visible layer (and vice-versa, since the connections are symmetric).

2.1.2. Deep architectures

Training a deep neural network uses a supervised learning strategy based on the stacking of RBMs trained layer per layer from bottom to top. Using deep networks has several advantages. First of all, deep learning (DL) algorithms provide data-driven feature extraction in which the output of each layer gives a representation of input data. Moreover, DL is able to deal with large sets of data. Deep neural networks often give very good results, which explains why they

are currently popular in many signal processing applications [3].

2.2. Training strategy

2.2.1. Learning the relationship between US and contour

Our method is divided into two phases. In the first phase, the network, acting as an autoencoder, is trained to reproduce its input vector. This vector is the concatenation of an ultrasound image and a binary image that represents the contour of the tongue, both reduced to 33 x 30 pixels, resulting in 1980 components, plus one constant input (bias). In the second phase, the network is asked to learn to reconstruct the tongue contour from the ultrasound image only. If we use a network trained on both contour and ultrasound image inputs, it is not obvious that the network will be able to produce a contour image if it lacks one of the inputs. The method described in [1] proposes to estimate the contours using ultrasound images only, under the hypothesis that the representation learned by a network trained on the two kinds of images embeds the relationship between these two kinds of data. The architecture used is called an autoencoder (see [4] [5] [6] and [7] for details). This type of network is trained to find an internal representation (code) of the input data so that it can be precisely reconstructed from this internal representation only.

In our case, if we are able to build an encoder that can generate a hidden coding like the one produced by the combination of ultrasound and contour images, but using ultrasound data only, then the decoder should be able to decode hidden information to reconstruct both ultrasound and contour data. This encoder is obtained in a "translational" manner [1] from the original encoder: the first RBM is replaced by what is called a translational RBM (tRBM, see Figure 1).

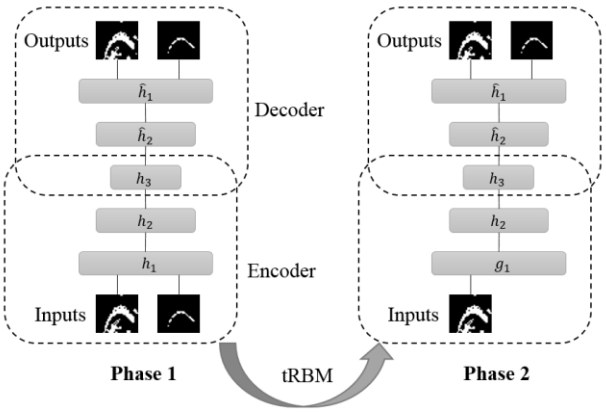


Figure 1: The two phases of learning. In the first step, the network learns the relationship between US images and contour. In the second phase, it is able to use this relationship to reconstruct the contour.

In the second phase, we learn only the parameters of the first layer, the others remaining unchanged. In other words, tRBM is trained to produce the same hidden features as the features extracted from the original RBM but without contour inputs. Then, if we use the original autoencoder but replace the first RBM by the tRBM, we can reconstruct a contour image that matches the tongue shape for each reduced ultrasound image of the test database.

2.2.2. Initial labeling

For both training and test, we used data from recording sessions described in [8]. We use in our training database an initial contour automatically extracted with an image processing algorithm developed in the Max/MSP software environment that outputs a tongue surface for each ultrasound image. Each ultrasound image is first pre-processed in order to select a region of interest of the relevant portion. On these images, the contour detection is done columnwise, from left to right. For each column, from top to bottom, every white pixel followed by a black pixel is considered a candidate contour point. This implies that several pixels can be selected as candidates. Since only one pixel per column is retained, a decision is made as to which candidate indeed belongs to the contour. This is done by comparing the current image to the previous one, the idea being that if a pixel was part of the previous contour, it or one of its neighbours must belong to the next one. If however no point from the previous contour matches one of the candidates, the selection is based on the neighbouring columns. Using this procedure, we pick up a set of (x, y) coordinates corresponding to the tongue surface contour in each image. These coordinates are then used as the ground truth (referred to as Ref) for the training set of the autoencoder.

3. CHOICE OF NETWORK ARCHITECTURE

Each example from the training dataset is presented to the network as an array containing the normalized intensities of the two binary images (1980 pixels + 1 bias). Several hyperparameter sets of the structure were explored (defined in sections 3.1-3.4): the number of layers, the number of unit per layer, the number of epochs for training and the size of "mini-batches", which are subsets of training data, usually of 10 to 100 examples. Our choice of parameters was based on the validation error (root mean squared difference between input and reconstruction) on a 17,000 example dataset (15,000 examples for training and 2,000 for validation).

3.1. Deep architectures

Stacking RBMs increases the level of abstraction of the model. However, we must determine the appropriate depth. For this purpose, we tried several architectures with various depths. In our experiments, we fixed 1,000 units per layer, 50 epochs and mini-batches of size 1,000 and tested the performances for a structure with 2, 3 and 4 hidden layers. The lowest validation error was achieved while using 3 hidden layers (see table 1).

Table 1: Influence of the number of layers on the validation error.

Number of hidden layers	Validation error
2	0.39
3	**0.38**
4	0.44

3.2. Network complexity

In classical machine learning models, we should use more training cases than parameters to avoid overfitting [9]. However, it is common to have a large number of hidden units in deep architectures. For our application, we based our choice of hidden units on the performances of several configurations allowing reasonable computing time, shown in table 2.

Table 2: Influence of the number of hidden units on the validation error for the 3 layer model.

Number of hidden units per layer	Validation error
500	0.41
1000	0.38
2000	**0.37**

3.3. Use of mini-batches

The use of mini-batches speeds up the algorithm because a weighted update occurs for each mini-batch instead of each example. However, finding an ideal mini-batch size is not straightforward. According to [9], the training set should be divided into mini-batches of 10 to 100 examples. We decided to test tongue contour reconstruction using several mini-batch sizes: 10, 50 and 100 examples per mini-batch.

Table 3: influence of mini-batch size on the validation error.

Mini-batch size	Validation error
10	0.65
50	0.53
100	**0.38**
200	0.40

Results showed that for a 3 layer network with 1,000 units per layer, 50 epochs and mini-batches of size 10, the error reached 0.65, while it decreased to 0.38 with mini-batches of size 100 and increases above.

3.4 Number of epochs

We used a similar procedure for testing the number of epochs necessary for weight updates. Keeping a reasonable number of epochs is crucial for computation time, and the time vs. performance balance should be considered. We used a 3 hidden layer network with 1,000 hidden units per layer, using mini-batches of size 100, and tested 5, 50 and 250 epoch runs. Using too many epochs degrades the performances. Furthermore, the number of epochs is one of the main bottlenecks for computation time.

Table 4: Influence of the number of training epochs on the training error.

Number of epochs	Validation error
5	0.41
50	**0.38**
250	0.40

4. RESULTS

4.1. Evaluation criteria

During the training stage, we used an autoencoder made of a 3-layer encoder associated with a symmetric decoder, with 2,000 hidden units, mini-batches of size 100 and 50 epochs. The evaluation of the quality of tongue shape reconstruction requires definite criteria and comparison to a reference. Generally speaking, a proper tongue shape is a curve that follows in a realistic manner the lower edge of the bright line appearing on an ultrasound image. It is important to extract the entire visible surface appearing in the ultrasound image, without adding artifacts [10]. In order to evaluate the quality of tongue shapes obtained with the DL method, we trained the network on a 17,000 example database and randomly selected another 50 ultrasound images from the same recording session and same speaker to test the tongue contour extraction. We first compared the contour coordinates obtained with DL to those obtained with manual labelling. However, the set of tongue contour coordinates does not always have the same number of points (see figure 2), so that comparison between coordinates is not straightforward. In [11], a measure is proposed to compare each pixel of a given curve to the nearest pixel (in terms of L_1 distance) on the curve it is compared to. This measure, named Mean Sum of

Distances (MSD) (see eq. (1)), provides an evaluation in pixels of the mean distance from a contour U to a contour V, even if these curves do not share the same coordinates on the x axis or do not have the same number of points. Contours are defined as a set of (x, y) coordinates: U is a set of 2D points $(u_1, \dots, u_n)$ and V is a set of 2D points $(v_1, \dots, v_m)$. *MSD* is defined as followed:

$$MSD(U,V) = \frac{1}{m+n}\left(\sum_{i=1}^{m} \min_j |v_i - u_j| + \sum_{i=1}^{n} \min_j |u_i - v_j|\right). \quad (1)$$

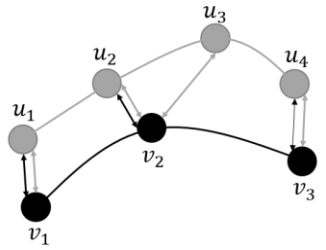


Figure 2: Comparing two tongue contours using MSD allows a comparison between two shapes even if some points are missing.

4.2. Experiments

Some example contours found using DL are shown in Figure 3. It now remains to compare the various methods used and evaluate the results. In addition to the comparison of the coordinates from DL to manual labelling (Hand), of course, we also want to compare automatically labelled ground truth (Ref) computed in sec. 2.2.2, to manual labelling in order to complete our analysis. Results appear in Table 5.

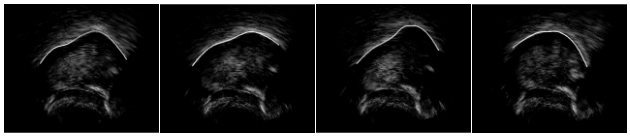


Figure 3: Examples of extracted contours for different tongue shapes.

The results show that the contours obtained using DL, Ref, and Hand labelling are of rather similar quality. This implies that the DL autoencoder, despite being shown only one image at a time, was able to achieve results comparable to an algorithm, Ref, that has access to the preceding image in the sequence. This suggests that the DL architecture has embedded a priori structural information stemming from the similarity constraint imposed in the Ref algorithm. This is an interesting result that may open the way to incorporating additional structural cues, for example from a physical 3D tongue model.

We also wished to compare these MSD values to those reported in the literature. In [11], the labelling from EdgeTrak, which uses Snake method (see [12] and [13]), is compared to two manual inputs from two different experts. To compare MSD values in pixels for different resolutions, we converted these values into millimetres using image resolution. Image size was 112.9 x 89.67 mm. The comparison between an expert 1 and an expert 2 gives a MSD of 0.85 mm (2.9 pixels with the conversion 1 px = 0.295 mm), the comparison between expert 1 and EdgeTrak gives a MSD of 0.67 mm, while the comparison between expert 2 and EdgeTrak gives an MSD of 0.86 mm. In [1], after 5 cross-validations, the average MSD computed on 8640 images is 0.73 mm. Our MSD values, computed with the equivalence 1 px = 0.35 mm, are quite similar to these, which allows us to conclude that the results obtained using DL trained with an automatic algorithm are of good quality.

Table 5: Average values of MSD for the comparison between Hand and Ref; Hand and DL; and Ref and DL.

	Average MSD (mm)
Hand vs. Ref	0.9
Hand vs. DL	1.0
Ref vs. DL	0.8

5. DISCUSSION

The use of a deep autoencoder to automatically extract the contour of the tongue from an ultrasound picture appears to give promising results. The results also show the interest of using automatically extracted contours as ground truth instead of manually labelling large amount of data, which is time consuming. Moreover, since our technique provides performances similar to those of an algorithm that uses temporal information, we can consider that our network was able to learn a new constraint based on its inputs, even if it does not use temporal prior knowledge. The choice of our network structure was adjusted and validated by several performance tests. In the future, providing the algorithm with a variety of learning databases, composed of sentences, words or phonemes pronounced by several speakers and in various modalities (e.g., speech or singing) would be a way to testing the sensitivity of the algorithm to variations in experimental conditions.

7. ACKNOWLEDGEMENTS

This work is funded by the European Commission via the i-Treasures project (FP7-ICT-2011-9-600676-i-Treasures).
We are also grateful to Cécile Abdo for the algorithm she developed on the Max/MSP software environment.

6. REFERENCES

[1] Fasel, I., Berry, J. 2010. Deep Belief Networks for Real-Time Extraction of Tongue Contours from Ultrasound During Speech. *20th International Conference on Pattern Recognition* Istanbul, IEEE, 1493-1496.

[2] Hinton G. E., Osindero, S. 2006, A fast learning algorithm for deep belief nets. *Neural Computation*, 18, 1527-1554.

[3] Yu, D., Deng, L. 2011. Deep Learning and Its Applications to Signal and Information Processing. *IEEE Signal Processing Magazine*, 28, 245-254.

[4] Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P. 2009. Exploring Strategies for Training Deep Neural Networks. *Journal of Machine Learning Research*, 10, 1-40.

[5] Bengio, Y. 2009. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2, 1-127.

[6] Vincent, P. Larochelle, H. Lajoie, I., Bengio, Y., Manzagol, P-A. 2010. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 11, 3371-3408.

[7] Arnold, L., Rebecchi, S., Chevallier, S., Paugam-Moisy, L. 2011. An introduction to deep-learning. *Advances in Computational Intelligence and Machine Learning, ESANN'2011*, Bruges, 477-488.

[8] Cai, J., Hueber, T., Manitsaris, S., Roussel, P., Crevier-Buchman, L., Stone, M., Pillot-Loiseau, C., Chollet, G., Dreyfus, G., Denby, B. 2013. Vocal Tract Imaging System for Post-Laryngectomy Voice Replacement. *International IEEE Instrumentation and Measurement Technology Conference*, Minneapolis, MN.

[9] Hinton, G. E. 2012. A Practical Guide to Training Restricted Boltzmann Machines. In: Montavon, G., Orr, G. B., Müller, K-R. (eds.), *Neural Networks: Tricks of the Trade (2nd ed.)*, Springer, 599-619.

[10] Stone, M. 2005. A Guide to Analysing Tongue Motion from Ultrasound Images. *Clinical Linguistics and Phonetics*, 19, 455-502.

[11] Li, M., Kambhamettu, R., Stone, M. 2005. Automatic Contour Tracking in Ultrasound Images. *Clin. Linguist. Phon*., 19, 545-554.

[12] Kass, M., Witkin, A., Terzopoulos, D. 1988. Snakes: Active contour models. *International Journal of Computer Vision*, 1, 321-331.

[13] Akgul, Y. S., Kambhamettu, C., Stone, M. 1999. Automatic Extraction and Tracking of The Tongue Contours. *IEEE Transactions on Medical Imaging*, 18, 1035-1045.

A Multi-Sensor Helmet to Capture Rare Singing, an Intangible Cultural Heritage Study

S. K. Al Kork[1,2], A. Jaumard-Hakoun[1,2], M. Adda-Decker[3], A. Amelot[3], L. Buchman[3], P. Chawah[3], G. Dreyfus[2], T. Fux[3], C. Pillot-Loiseau[3], P. Roussel[2], M. Stone[4], K. Xu[1,2], B. Denby[1,2]

[1]*Université Pierre Marie Curie, Paris, France,*

[2]*Signal Processing and Machine Learning Lab, ESPCI Paris-Tech, Paris, France,*

[3]*Phonetics and Phonology Laboratory, LPP-CNRS, UMR7018, University Paris3 Sorbonne Nouvelle*

[4]*Vocal Tract Visualization Lab, University of Maryland Dental School, Baltimore, USA*

samer_alkork@hotmail.com, aurore.hakoun@espci.fr, madda@univ-paris3.fr, angelique.amelot@univ-paris3.fr, lise.buchman@numericable.fr, patrick.chawah@gmail.com, gerard.dreyfus@espci.fr, thibaut.fux@univ-paris3.fr, claire.pillot@univ-paris3.fr, pierre.roussel@espci.fr, mstone@umaryland.edu, kelele.xu@gmail.com, denby@ieee.org

Abstract

A portable helmet based system has been developed to capture motor behavior during singing and other oral-motor functions in a non-laboratory experimental environment. The system, based on vocal tract sensing methods developed for speech production and recognition, consists of a lightweight "hyper-helmet" containing an ultrasonic (US) transducer to capture tongue movement, a video camera for the lips, and a microphone, coupled with a further sensor suite including an electroglottograph (EGG), nose-mounted accelerometer, and respiration sensor. The system has been tested on two rare, endangered singing musical styles, Corsican "Cantu in Paghjella", and Byzantine hymns from Mount Athos, Greece. The versatility of the approach is furthermore demonstrated by capturing a contemporary singing style known as "Human Beat Box."

Keywords: portable speech collection system, ultrasound, EGG, accelerometer, data acquisition, intangible cultural heritage, i-Treasures project.

1. Introduction

A major objective of the i-Treasures project is to provide students with innovative multi-media feedback to train specific articulatory strategies for different type of rare singing, considered an endangered Intangible Cultural Heritage. To this end, i-Treasures will carry out vocal tract capture during rare singing performances to enable study of production mechanisms, and to define reliable features for a subsequent animation of these articulatory movements for use in educational scenarios and automatic classification tasks. To accomplish this, it is necessary to build a system that can record the configuration of the vocal tract – including tongue lips, vocal folds and soft palate – in real time, and with sufficient accuracy to establish a link between image features and actual, physiological elements of the vocal tract.

Ultrasound, US, is a popular non-invasive technique for real time imaging of the vocal tract. Examples of portable devices that acquire US images of the tongue and video of the lips, for applications in speech synthesis and silent speech interfaces (Denby and Stone 2004) (Cai, et al. 2011), have been described in the literature. Ultrasound also requires no external magnetic field, and can thus also be readily complemented with other sensors. Here, we present a system based on a helmet containing a US probe, lip camera, and microphone, coupled with a suite of other sensors including electroglottograph (EGG), to measure and record vocal fold contact movement during speech; a piezoelectric accelerometer, for detecting the nasal resonance of speech sounds (Stevens, Kalikow and Willemain 1990); and a respiration sensor belt to determine breathing modalities (Tsui and Hsiao 2013).

The proposed system is advantageous in that 1) it is lighter and easier to wear for long periods than other solutions proposed in the literature (Wrench, Scobbie and Linden 2007); and 2) the combination with other sensors has the potential to greatly enhance our knowledge of rare singing techniques, and allow the extraction of sensorimotor features in order to drive a 3D avatar for learning scenarios.

2. Methods

Figure 1 presents a schematic overview of the modules contained in the capture system. Each sensor first requires specific gain tuning and/or zero calibration protocols, before the streams are simultaneously and synchronously recorded with the RTMaps toolkit, which will be described in section 2.2.

2.1. Helmet design and sensor setup

The helmet allows simultaneous collection of vocal tract and audio signals. As shown in Figure 2, it includes an adjustable platform to hold the US probe in contact with the skin beneath the chin. The probe used is a microconvex 128 element model with handle removed to reduce its size and weight, which captures a 140° image to allow full visualization of tongue movement. The US machine chosen is the Terason T3000, a system which is lightweight and portable yet retains high image quality, and allows data to be directly exported to a PC via Firewire. A video camera (from The Imaging Source) is positioned facing the lips (Figure 2). Since differences in background lighting can affect computer recognition of lip motion, the camera is equipped with a visible-blocking filter and infrared LED ring, as is frequently done for lip image analysis. Finally, a commercial lapel microphone (Audio-Technica Pro 70) is also affixed to the helmet to record sound.

The three non-helmet sensors are directly attached to the body of the singer as indicated in Figure 3. An accelerometer attached with adhesive tape to the nasal bridge of the singer captures nasal bone vibration related to nasal tract airway resistance, which is indicative of nasal resonance during vocal production. Nasality is an important acoustic feature in voice timbre. An EGG (Model EG2-PCX2, Glottal Enterprises Inc.) is strapped to the singer's neck to record a time dependent signal whose peaks are reliable indicators of glottal opening and closing instances (Henrich, et al. 2004). Finally, on the singer's chest, a

respiration sensor or “breathing belt” is affixed to measure breathing modalities during singing.

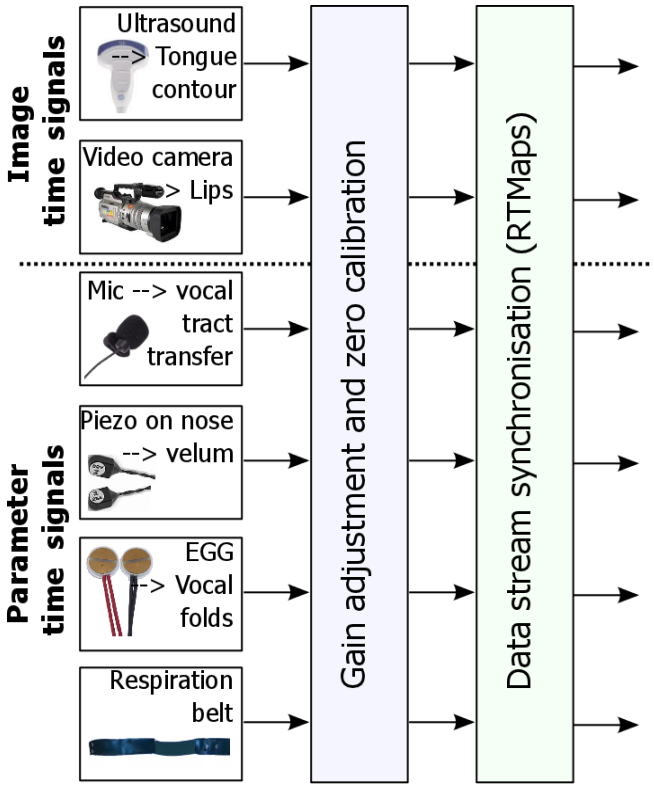


Figure 1 *Overview of the vocal tract capture system for the rare singing sub-use cases. Six sensors (left), capture motion of five physical events: (1) tongue, (2) lips, (3) the acoustic speech wave, (4) nasal resonance, (5) vocal folds, and (6) respiratory muscle. Each instrument is processed (middle) and digitally recorded (right).*

2.2. Data Acquisition and system design

2.2.1. System architecture

The data acquisition system must be able to synchronously record US and video data at sufficiently high frame rates to correctly characterize the movements of the tongue and lips, as well as the acoustic speech signals, the EGG, the accelerometer and the respiratory `waveforms. The acquisition platform was developed using the Real-Time, Multi-sensor, Advanced Prototyping Software (RTMaps®, Intempora Inc, Paris FR).

2.2.2. RTMaps real time user interface

The data acquisition platform has the ability both to record and display data in real time, and the acquired data can be stored locally or transferred over a network. Figure 4 displays a screen shot from the platform. Ultrasound and video images are streamed at a rate of 60 frames per second, then stored in either .bmp or jpeg format. Image size for US and camera are 320 by 240 pixels and 640 by 480 pixels respectively. The EGG, the microphone, the piezoelectric accelerometer and respiration belt are interfaced to a four-input USB sound card (AudiBox44VSL) whose output interfaces to the acquisition system. These four analog input signals are sampled at 44100 Hz with a 16 bit encoding. The sampled analog signals are saved to a .wav format.

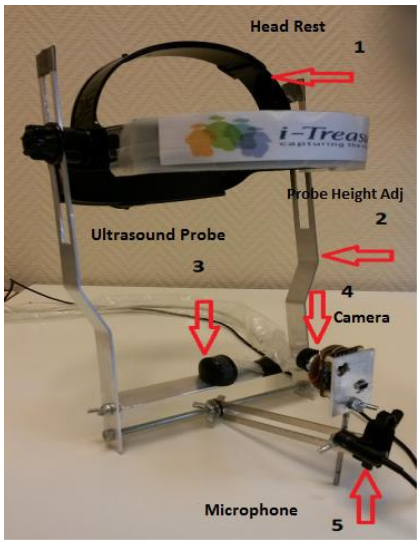


Figure 2 *Multi-sensor Hyper-Helmet: 1) Adjustable headband, 2) Probe height adjustment strut, 3) Adjustable US probe platform, 4) Lip camera with proximity and orientation adjustment, 5) Microphone.*

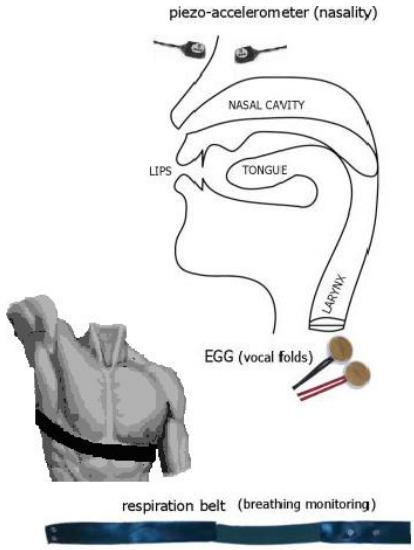


Figure 3 *Schematic of the placement of non-helmet sensors, including the (1) accelerometer, (2) EGG, and (3) respiration belt*

2.2.3. Definition of recording material

In order to study a variety of singing techniques, and to extract features for automatic classification and pedagogical scenarios, we will collect material of varying degrees of complexity: isolated vowels (/i/, /u/, /e/, /o/, /a/), CV syllables (/papapapapa/, /tatatatata/, /kakakakaka/...), sung phrases and entire pieces, where the material is to be produced both in spoken and singing modes. Byzantine chant, Corsican Paghjella, and the contemporary singing style known as “Human Beat Box”, HBB, have been chosen for study. For Byzantine chant, different styles (Mount Athos *vs* Ecumenical Patriarchate of Constantinople styles for example) have been selected. For Corsican Paghjella, we propose to study *versa* (melodies) from three different regions known for their traditional singing styles: Rusio, Sermanu and Tagliu-Isolacciu. The protocol to be used for Sardinian Canto a Tenore is still under discussion. For HBB,

basic material will be recorded as defined in (Proctor, et al. 2013), as well as short HBB phrases and longer performances in different styles, with details still to be defined.

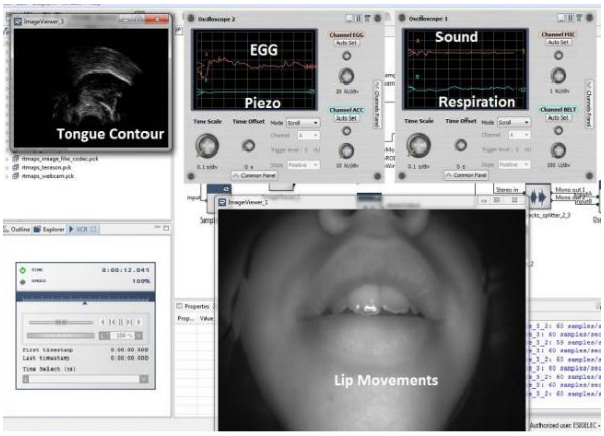


Figure 4 *RTMaps Data Acquisition User Interface, showing simultaneous recording and visualization of ultrasound tongue images, lip video, and the four analog sensors.*

3. Preliminary experimental results

3.1. Sensor synchronization

Sensor synchronization is crucial in our study since extracted sensor features are ultimately to be used to drive a 3D tongue avatar. To check synchrony, an event common to all sensor channels is created using the procedure illustrated in Figure 5. A syringe containing ultrasound gel is struck by a spring-loaded weight, ejecting a gel droplet that strikes the head of the ultrasound probe and shorts together the two sides of the EGG sensor. Vibration induced in the syringe body is detected in the nasality accelerometer, and the acoustic signal of the weight hitting the syringe is recorded by the microphone. Finally, the video camera normally used for the lips is positioned so that it can also capture the droplet as it is ejected. The resulting signals obtained from the 5 sensors are shown in Figure 6

We have calculated the timestamps at which the gel droplet occurred in all sensors. The time stamp at which the droplet was triggered for the microphone, piezo and EGG was 4.060s, 4.070s and 4.070s respectively. The droplet was captured by the camera and Ultrasound at image number 244 in sequence and at calculated time stamps of 4.066s based stream rate of 60 fps (Figure 6). The average latency among all sensors is thus of the order of 10ms.

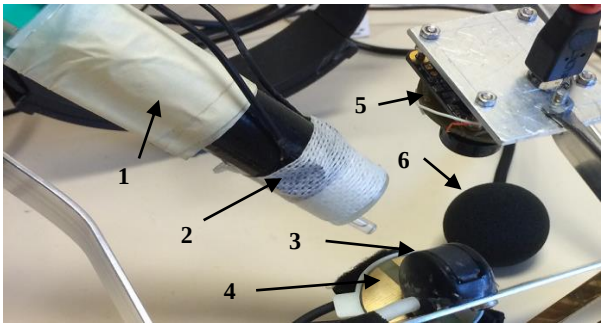


Figure 5 *Sensor synchronization experimental setup test. 1) Syringe, 2) Piezo, 3) EGG sensor, 4) US transducer, 5) Camera, 6) Microphone*

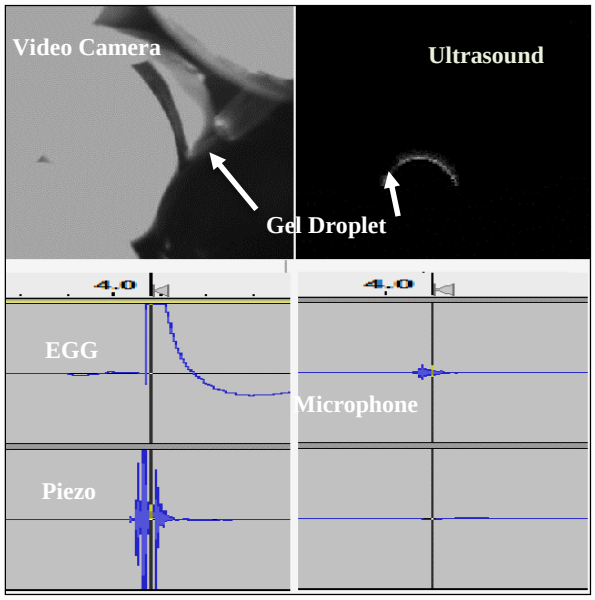


Figure 6 *Top left, arrow indicates gel droplet ejected from tip of syringe; top right, arrow indicates arrival point of droplet on ultrasound probe; bottom left and right, time waveforms of EGG, Piezo, and microphone, signals, see discussion in text*

3.2. Rare-singing data collection

Our data collection activities are focused on preparatory steps dealing with technical, operational and functional requirements. These steps are listed here below and some details are provided for the major steps.

3.2.1. Assessment phase of the hyper-helmet for the different singing types:

The voice capture system has been tested with one expert singer each for the Human Beat Box, HBB (Da Vox) (Figure 7a), the Paghjella (B. Sarocchi) Figure 7b, and the Byzantine (D. Manousis) (Figure 7c) musical styles. Each singer participated in a recording session to validate the helmet with respect to his or her style, and to assist us in specifying an appropriate data collection protocol. A recording session consists of three phases: 1) singer preparation (wearing of the helmet and body sensors), 2) sensor calibration and, 3) data collection proper. The three phases need to be optimized with respect to time delays in sensor setup and ease of use. Figure 6 illustrates the versatility of the helmet, which can be used with any head size and shape and does not impede the singing function.

The Byzantine expert singer (D. Manousis) produced vowels in both singing and speaking mode, before singing a dozen segments of Byzantine chants in both Mount Athos and Ecumenical Patriarchate of Constantinople styles. The Corsican Paghjella singer (B. Sarocchi) first produced spoken and singing voice using isolated vowels and connected CV syllables with major Corsican vowels and consonants, and then performed three Paghjella songs. For the HBB case, we undertook several testing and recording sessions with our expert, (Da Vox). A specific problem for HBB is the difficulty of stabilizing of the ultrasound probe in view of the large range of motion of the jaw in this singing style, as compared to the

other styles. Each of the three singing styles produced about 30 minutes of singing material, which are being used to develop and assess the next steps to be undertaken in the continuing development of our synchronous data collection platform, as well as our data calibration, data display and analysis modules.

Figure 7 *a) HBB expert singer (Da Vox), b) Paghjella expert singer (B. Sarocchi), and c) Byzantine expert singer (D. Manousis)*

3.3. Pilot data

In this section, some samples of the pilot data are presented. Figure 8 shows similar vocalic /o/-like sounds by three singers specialized in different singing styles: Byzantine chant (left), Cantu in Paghjella singer (secunda voice, middle) and HBB singer (right). Initial data of ultrasound tongue images and video camera lips image are displayed for all the above singers as shown Figure 9, Figure 10 and Figure 11.

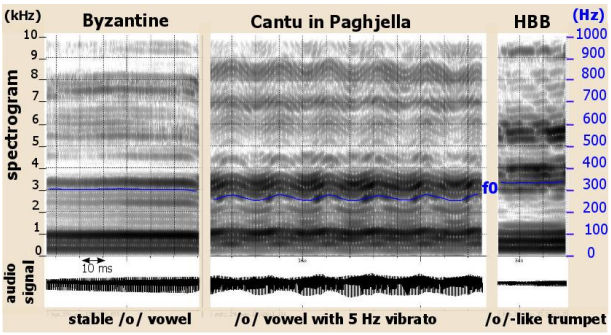


Figure 8 *Three vocalic /o/ samples of different style singing voice (Byzantine,Cantu in Paghjella, HBB). Spectrograms (10kHz band in black on the left) and f0 curves (in blue on the right) are shown in the upper panel, corresponding acoustic waveforms in the lower one.*

Figure *9 Byzantine Singing Case: Left) Ultrasound Tongue image. Right) Video camera lip image*

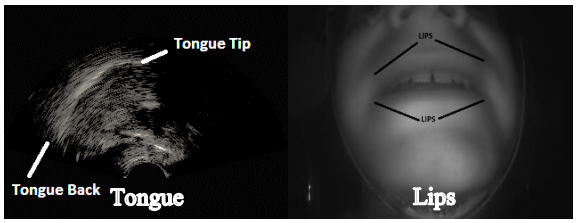


Figure 10 *HBB singing case: Left) Ultrasound Tongue image. Right) Video camera lip image*

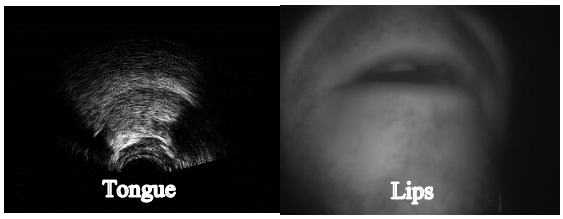


Figure 11 *Corsican singing case: Left) Ultrasound Tongue image. Right) Video camera lip image*

4. Discussion and conclusion

The vocal tract acquisition system for the preservation of rare singing techniques has evaluated on initial data taken on all three singing cases. Results are promising with respect to system reliability, portability and quality of the recorded data. The system still needs to be tested on a larger number of users, and the helmet better adapted to the HBB case due to rapid jaw movements encountered in this singing style. The next steps will involve integrating a feature extraction processing block for all the sensors into the RTMaps platform, as well as provide an easy user friendly interface to drive and control a 3D vocal tract avatar to be developed in future.

5. Acknowledgements

This work is funded by the European Commission via the i-Treasures (FP7-ICT-2011-9-600676-i-Treasures).

6. References

Cai, Jun, Thomas Hueber, Bruce Denby, Elie-Laurent Benaroya, Gérard Chollet, Pierre Roussel, Gérard Dreyfus, and Lise Crevier-Buchman. 2011. "A visual speech recognition system for an ultrasound-based silent speech interface." *In Proc. of ICPhS.* pp. 384-387.

Denby, B., and M. Stone. 2004. "Speech synthesis from real time ultrasound images of the tongue." *In Acoustics, Speech and Signal Processing; ICASSP.* pp 685-I688.

Henrich, N., C. d'Alessandro, M. Castellengo, and B. Doval. 2004. "On the use of the derivative of electroglottographic signals for characterization of nonpathological voice phonation." *Journal of the Acoustical Society of America* pp. 1321-1332.

Proctor, M., E. Bresch, D. Byrd, K. Nayak, and S. Narayanan. 2013. "Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study." *Journal of the Acoustical Society of America (JASA)* 133 (2): pp. 1043-1054.

Stevens, K.N., D.N. Kalikow, and T.R. Willemain. 1990. "A miniature accelerometer for detecting glottal waveforms and nasalization." *Journal of Speech and Hearing Research JSHR* pp. 594-599.

Tsui, W.H., and Tzu-Chien Hsiao. 2013. "Method and System on Detecting Absominals for singing." *Proc. IEEE EMBC.* pp.1-8.

Wrench, A., J. Scobbie, and M. Linden. 2007. "Evaluation of a helmet to hold ultrasound probe." *Ultrafest IV.*