# Synthèse de la parole à partir d'imagerie ultrasonore et optique de l'appareil vocal

École Supérieure de Chimie Physique Électronique de Lyon

Spécialité Electronique - Télécommunications - Informatique

Projet de fin d'étude

**Thomas Hueber**

# Table des matières

# Table des figures

# Remerciements

Je tiens à remercier les membres du Laboratoire d'Electronique de l'École Supérieure de Physique et Chimie Industrielles de la Ville de Paris ainsi que son directeur Monsieur Gérard Dreyfus. Ils m'ont apporté leur soutien et leur aide au cours de ces six mois de travail. Merci à Pierre Roussel, Yacine Oussar et Rémi Dubois.

Je tiens également à remercier les membres du Laboratoire Traitement et Communication de l'Information de Télécom Paris et plus spécialement Monsieur Gérard Chollet, co-encadrant de mon stage, Madame Isabelle Bloch et Monsieur Guido Aversano pour leur conseils avisés et leurs appuis.

Enfin, je souhaite remercier tout particulièrement mon maître de stage Monsieur Bruce Denby, pour sa disponibilité, sa confiance et son écoute.

# Introduction Générale

La voix occupe une position privilégiée dans l'ensemble des vecteurs d'informations de la société humaine. Le traitement de la parole, qui a connu une expansion fulgurante depuis les années 1960, reste aujourd'hui un axe de recherche fondamental en télécommunications. La parole peut être analysée de bien des façons. Les acousticiens analysent le signal perçu par un microphone (étude de la fréquence fondamentale, de l'énergie, du spectre ...). En revanche, les phonéticiens s'intéressent moins au signal qu'à la manière dont il est produit par le système articulatoire. La cinéradiographie, l'imagerie ultrasonore et plus récemment l'imagerie par résonance magnétique (IRM), sont utilisés depuis quelques années pour l'étude de l'appareil vocal. En visualisant les articulateurs du conduit vocal, ces techniques performantes d'imagerie permettent de mieux appréhender le processus mécanique de production de la parole. Longtemps disjointes, ces deux approches peuvent aujourd'hui se rejoindre et évoluer vers une analyse mutimodale audio-visuelle de la parole. La description conjointe des configurations successives du conduit vocal et des variations acoustiques du signal de parole émis, la modélisation de la relation qui lie les mouvements du système articulatoire et le signal audio, sont les principales motivations du projet *Ouisper*, dans lequel s'inscrit mon stage de fin d'étude.

**Le projet *Ouisper* vise à réaliser un dispositif capable de reconstituer un signal de parole intelligible à partir de la saisie des mouvements de certains articulateurs par imagerie optique et ultrasonore** Un tel dispositif pourrait avoir des applications importantes en médecine (prothèses pour patients privés de l'usage de la parole) et en télécommunications (« téléphone silencieux »)

Des études préliminaires ont été initiées par mon maitre de stage, Monsieur Bruce Denby, professeur à l'université Pierre et Marie Curie (Paris VI), chercheur à l'École Supérieure de Physique et de Chimie Industrielles (ESPCI) de la Ville de Paris et initiateur du projet. Pour mener à bien cette étude multi-disciplinaire, Monsieur Denby a instauré une collaboration avec des équipes aux compétences complémentaires :

- Le Laboratoire d'Électronique de l'ESPCI, dirigé par Monsieur Gérard Dreyfus, spécialisé dans les techniques d'intelligence artificielle et les problèmes de modélisation non-linéaire
- Le Laboratoire Traitement et Communication de l'Information de l'École Nationale Supérieure des Télécommunications (ENST), dirigé par Monsieur Henri Maitre qui bénéficie notamment de la grande expérience dans le traitement de la parole de Monsieur Gérard Chollet, co-encadrant de mon stage
- Enfin, le Vocal Tract Visualization Lab (VTVL) de l'université de Maryland (Baltimore - USA) dirigé par Madame Maureen Stone, qui est un pionnier de l'imagerie de l'appareil vocal.

Le projet *Ouisper* est à ce jour soutenu par l'Agence Nationale de la Recherche (ANR) et la Délégation Générale de l'Armement (DGA). Ainsi, j'ai eu la chance d'intégrer ces différentes équipes, travaillant autour de ce projet novateur et ambitieux. Mon travail s'est organisé autour de l'analyse des images ultrasonores et optiques de l'appareil vocal, le traitement de la parole et la modélisation visio-acoustique. Ce document, qui décrit les résultats des six mois de recherche qui ont constitué mon stage, tente de refléter la pluridisciplinarité du sujet proposé.

Le premier chapitre présente succinctement l'architecture et le fonctionnement de l'appareil vocal, les principales pathologies dont il peut faire l'objet ainsi que les techniques d'imageries qui permettent de l'étudier. Ce premier chapitre décrit l'environnement dans lequel s'inscrit le projet ***Ouisper***. Le deuxième chapitre présente de manière approfondie ce projet et le sujet du stage. Le troisième chapitre expose les techniques de traitement des images mises en oeuvre dans le cadre de ce stage. Le chapitre 4 est consacré aux méthodes d'analyse-synthèse du signal de parole. Enfin les techniques de modélisation visio-acoustique, fondées sur l'utilisation des techniques d'apprentissage artificiel, feront l'objet du dernier chapitre.

# Chapitre 1

# Généralités sur le principe de fonctionnement de l'appareil vocal - Techniques d'imagerie

## 1.1 Introduction

Ce chapitre est consacré à une présentation succincte de l'appareil vocal, de son architecture et de son fonctionnement. Cette présentation, non-exhaustive, a pour but de présenter le cadre dans lequel s'inscrit le projet ***Ouisper***.

## 1.2 Production de la parole

### 1.2.1 Architecture de l'appareil vocal

La figure 1.1 représente une coupe sagittale de l'appareil vocal ou phonatoire [1].

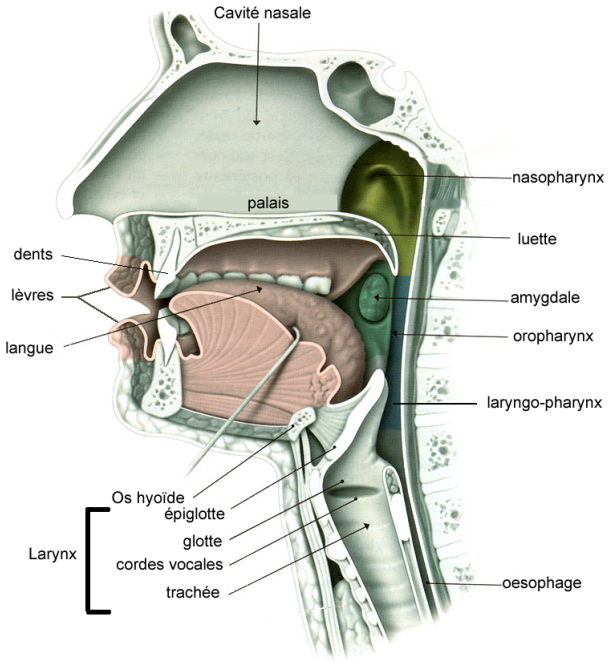


FIG. 1.1 – L'appareil vocal

La section suivante explicite le fonctionnement de l'appareil vocal, ainsi que le rôle joué par ses différents constituants lors du processus de production de la parole.

[1] Source : Le Cerveau( *http ://www.lecerveau.mcgill.ca/*)

### 1.2.2 Fonctionnement de l'appareil vocal

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive. L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le **larynx** où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée. Les **cordes vocales** sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée **glotte**. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés (ou sourds). Les sons voisés (ou sonores) résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, les force à s'ouvrir, ce qui fait tomber la pression et permet aux cordes vocales de se refermer ; pour la plupart des sons, des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des **cavités pharyngienne** (laryngo, naso, oro-pharyngienne), buccale et labiale. Lorsque la **luette** est en position basse, la **cavité nasale** vient s'y ajouter en dérivation.

### 1.2.3 Phonétique articulatoire

La phonétique articulatoire a pour but de relier la configuration de l'appareil vocal à la production de certains phonèmes [Dutoit, 2003]. L'ensemble des phonèmes peut être défini en première approximation comme l'ensemble des sons utilisés lors de l'utilisation d'une langue. Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal. Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la bouche se réduit alors à une modification du timbre vocalique. Si, au contraire, le passage se rétrécit par endroit, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. La bouche est dans ce cas un organe de production à part entière. Les semi-voyelles, quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes. Enfin, la langue a un rôle prépondérant dans la production des liquides. Sa hauteur détermine la hauteur du pharynx : plus la langue est basse, plus le pharynx est court. Elle détermine aussi le lieu d'articulation, région de rétrécissement maximal du canal buccal, ainsi que l'aperture, écartement des organes au point d'articulation. L'articulation de [l] ressemble à celle d'une voyelle, mais la position de la langue conduit à une fermeture partielle du conduit vocal.

## 1.3 Pathologies de l'appareil vocal - Traitements possibles

En raison des applications médicales attendues du projet *Ouisper*, les principales pathologies de l'appareil vocal auxquelles il pourrait porter remède, ainsi que les solutions actuellement mises en œuvre, sont décrites succinctement dans cette section, à partir des ressources fournies par l'Association Française des Mutilés de la Voix [2] et le Groupe Coopérateur Multidisciplinaire en Oncologie (GERCOR [3]).

### 1.3.1 Le cancer du Larynx

Le cancer du larynx peut naître dans n'importe quelle région du larynx. La tumeur prend la forme d'une ulcération anormale de la corde vocale. La maladie évolue par l'extension de la tumeur vers l'autre corde. Parfois, le cancer provoque l'immobilisation des cordes vocales. Le cancer le plus dangereux est celui qui touche l'étage situé au-dessus des cordes vocales, c'est-à-dire la partie sus-glottique. Il est alors nécessaire d'intervenir vite car ce cancer, à la différence du cancer de la glotte, peut s'étendre rapidement aux ganglions lymphatiques avoisinants, lesquels favorisent la prolifération du cancer dans le reste du corps. Généralement, le cancer du larynx ne survient qu'après une longue exposition aux facteurs cancérigènes. La fumée est la cause principale du cancer du larynx ; cette cause est aggravée par la consommation conjointe d'alcool. La respiration de matières cancérigènes telles que l'amiante ou autres poussières cause davantage ce type de cancer. Le traitement du cancer du larynx repose, dans la plupart des cas, sur la chirurgie. Le patient subit l'une des trois interventions suivantes :

- une cordectomie (ablation d'une seule corde vocale) pour une lésion peu étendue,
- une chirurgie reconstructive permettant la conservation de la voix,
- une ablation du larynx : c'est la laryngectomie totale ou pharyngo-laryngectomie si le pharynx est atteint.

[2] Association Française des Mutilés de la Voix - *http ://www.mutiles-voix.com*
[3] GERCOR - http ://www.canceronet.com/

La laryngectomie totale est l'ablation du larynx et éventuellement de ses annexes. L'opération nécessite l'exécution d'une trachéostomie [4] pour maintenir le passage de l'air vers les poumons et de sutures pharyngées destinées à rétablir l'étanchéité de la voie digestive. Ces interventions répondent aux contraintes urgentes imposées par un cancer. Elles ne garantissent pas, pour la plupart d'entre elles, un fonctionnement correct de l'appareil phonatoire. Pour preuve, après laryngectomie totale, le patient ne peut plus parler en voix laryngée, celle utilisant l'air respiratoire.[5] Après une trachéostomie, la phonation primaire n'est rendue possible que lors de l'obstruction du trachéostome ; elle nécessite de plus une certaine rééducation. Certaines solutions existantes tentent néanmoins de préserver la capacité de vocalisation du patient opéré ; elles font l'objet du paragraphe suivant.

### 1.3.2 Traitements existants

Il est nécessaire de distinguer les prothèses internes, implantées lors d'une opération chirurgicale, des dispositifs externes. En effet, une variante de la laryngectomie totale est pratiquée depuis le début des années 90. Elle consiste à conserver une fistule, c'est-à-dire une communication entre la voie digestive et la voie respiratoire. Les différentes techniques visent à équiper la fistule d'une prothèse, appellée prothèse phonatoire ou implant. Elle assure l'étanchéité au passage des aliments, tout en permettant le passage de l'air pulmonaire expiré pour faire vibrer la partie haute de l'oesophage ; c'est la voix trachéo-oesophagienne [6]. Différents types de prothèses ont été expérimentés. Mais il n'existe pas encore à ce jour de prothèse capable d'éviter le risque potentiel de fausse route, c'est-à-dire de passage de liquide dans les voies respiratoires.

Un patient trachéostomisé peut utiliser un dispositif externe nommé **electrolarynx**. Il s'agit d'une prothèse électrique externe, qui, posée contre la gorge, produit une vibration qui est ensuite modulée par la bouche. Il existe aussi des prothèses externes pneumatiques : l'air est prélevé du stome et est amené par un tuyau dans la bouche ; l'extrémité de ce tuyau vibre, ce qui fournit le son initial qui est ensuite modulé par la bouche.

## 1.4 Imagerie ultrasonore de l'appareil vocal

Depuis quelques années, les ultrasons, la cinéradiographie et l'IRM sont utilisés pour l'étude de l'appareil vocal. Ces deux dernières techniques, bien que fournissant des images de grande qualité, ne sont pas inoffensives pour le patient, dans le cadre d'une utilisation fréquente et prolongée. En revanche, l'imagerie ultrasonore ne présente pas de tels inconvénients, au prix certes, d'une qualité d'acquisition plus faible. C'est pour cette raison que cette technique est privilégiée dans le cadre de cette étude. Les sections suivantes explicitent le fonctionnement et les spécificités de l'imagerie ultrasonore de l'appareil vocal.

### 1.4.1 Principe et généralités sur l'imagerie ultrasonore

Les généralités présentées ici, sont issues de sources multiples. Nous noterons néanmoins, le cours de Madame Elsa Angelini [7] et du Docteur Paul Barthez [8].

#### L'onde ultrasonore

Les sons sont émis par des corps animés d'un mouvement vibratoire et se propagent sous forme d'ondes mécaniques susceptibles de subir des réflexions (échos), des réfractions, et des interférences. L'onde sonore en déplacement dans un milieu donné est caractérisée par sa fréquence (f) et sa longueur d'onde (l). Ces 2 grandeurs permettent de déterminer la vitesse de propagation des sons (v) dans le milieu $v = f * l$

Les ultrasons sont des sons dont la fréquence est supérieure à 20 000 Hz ; ils sont inaudibles pour l'oreille humaine. En échographie, les ultrasons utilisés ont une fréquence qui varie entre 2 et 40 MHz. L'onde ultrasonore est émise par la vibration d'un cristal piézoélectrique [9]. L'onde ultrasonore produite présente les mêmes propriétés de transmission qu'une onde sonore mais avec une longueur d'onde plus petite. Ceci lui permet de résoudre des objets de plus petite taille, augmentant ainsi la résolution spatiale de l'observation.

[4] Mise en place d'un trou au milieu du cou, appelé trachéostome ou stome,permettant de respirer sans perturber l'alimentation

[5] On estime la proportion de personnes ayant subi une laryngectomie à 1 sur 3000 en France à ce jour soit 20000 patients environ.

[6] En absence de fistule, l'air pour faire vibrer les cordes vocales peut être fourni par des éructations (rots) contrôlées ; c'est la technique de la voix oesophagienne

[7] Madame Elsa Angelini est professeur à l'ENST - Le document est disponible ici : http ://www.tsi.enst.fr/ angelini/

[8] Monsieur Paul Barthez, est docteur vétérinaire, il enseigne l'imagerie médicale à l'École Nationale Vétérinaire de Lyon - Le document est disponible sur le site http ://www.vet-lyon.fr/

[9] Un élément piezoelectrique convertit l'électricité en vibration mécanique et réciproquement. Une sonde ultrasonore est composée de plusieurs cristaux piézo-électriques.

**Formation de l'image ultrasonore**

Les paramètres déterminants dans la propagation des sons dans les différents milieux sont la densité ou masse volumique (d) et la vitesse de propagation des sons (v). L'impédance acoustique (Z) est définie par le produit de ces 2 caractéristiques du milieu : $Z = d * v$. La vitesse de propagation dans les différents milieux dépend beaucoup de leur dureté. L'impédance acoustique dépend donc également de la dureté des milieux. Une **interface** est constituée de la juxtaposition de deux milieux d'impédances acoustiques différentes. Dans l'organisme, les grandes différences d'impédance acoustique se rencontrent d'une part entre les tissus mous et l'air, d'autre part entre les tissus mous et les tissus durs (os, calculs, corps étrangers). L'interaction d'un onde ultrasonore et d'une interface donne lieu à quatre phénomènes principaux :

- La réflexion
- La réfraction
- La diffusion ou réflexion diffuse
- L'atténuation

Lorsque un faisceau d'ultrasons arrive sur une interface placée à angle droit par rapport à sa direction initiale, une partie est **réfléchie** et repart dans le sens opposé, et l'autre partie traverse l'interface et continue sa route sans changer de direction. Un écho est un son qui est réfléchi et qui est réceptionné après un temps de latence, correspondant à son temps de propagation dans le milieu concerné. La proportion d'ultrasons réfléchis est directement proportionnelle à la différence d'impédance acoustique entre les deux milieux. De telles interfaces, très réfléchissantes (dites « échogènes ») existent lorsque les tissus mous organiques sont en contact avec de l'air (poumon, tube digestif) ou des structures minéralisées (os, calculs).

Lorsque le faisceau ultrasonore arrive sur une surface réflective avec un angle oblique, une partie du faisceau est réfléchi avec un angle de réflexion égal à l'angle incident. La partie transmise est déviée avec un angle qui dépend de la vitesse de propagation des deux milieux concernés. Il s'agit du phénomène de **réfraction**. En échographie, l'interaction du faisceau ultrasonore avec une surface oblique lisse, entraîne une disparition du signal. Aucun son ne revient directement sur la sonde après la réflexion oblique et le faisceau change de direction après la réfraction. Ce phénomène est à l'origine d'un artefact fréquent, appelé **ombre acoustique**.

Lorsque le faisceau ultrasonore arrive sur une surface irrégulière, la réflexion observée est qualifiée de diffuse. Les ondes ainsi rétrodiffusées de manière chaotique sont à l'origine d'un bruit multiplicatif nommé *speckle*[10].

L'intensité ultrasonore détectée par un récepteur diminue avec sa distance à la source et avec la présence d'interfaces sur le trajet du faisceau. C'est notamment le cas pour le faisceau ultrasonore utilisé en échographie, dont l'intensité diminue avec la profondeur d'exploration. Cette atténuation des ultrasons est due aux multiples interactions mentionnées précédemment (réflexion, diffusion, réfraction). L'atténuation des ultrasons dépend des milieux traversés, mais aussi des caractéristiques de l'onde ultrasonore et en particulier de sa fréquence : plus elle est élevée, plus l'atténuation est importante.

| Fréquence des ultrasons | Profondeur d'exploration maximale |
|---|---|
| 2.5-3.5 MHz | 15 cm |
| 5 MHz | 10 cm |
| 7.5 MHz | 5-6 cm |
| 10-12 MHz | 2-3 cm |

L'image ultrasonore est reconstituée à partir des échos recueillis par la sonde. Le temps entre l'émission et la réception d'une onde est converti en distance, en se basant sur la vitesse du son dans l'eau. La cartographie des distances ainsi déterminée permet de former une image. Deux modes de formation de l'image sont couramment utilisés : le mode B et le mode M. Ce dernier ne sera pas exposé dans cette section, seul le mode B (brillance), plus commun y est décrit brièvement.

Le mode B représente l'intensité du signal reçu par la brillance d'un point sur l'écran. Plus le point est brillant, plus la réflexion des ultrasons a été importante, donc, plus l'écho est intense. Ce mode permet d'obtenir une image bidimensionnelle qui représente une coupe de la structure explorée.

**Qualité de l'image ultrasonore**

La qualité de l'image dépend premièrement de la résolution spatiale. Cette dernière est fonction de la fréquence des ultrasons. Plus la fréquence est élevée, plus la longueur d'onde est petite et plus la résolution spatiale est bonne. En revanche, la fréquence d'émission de la sonde a une influence sur l'atténuation des ultrasons, donc sur la profondeur d'exploration. Plus le signal émis par la sonde est de basse fréquence (3,5 - 5 MHz), plus la profondeur d'exploration est importante, mais moins bonne est la qualité de l'image. **La qualité d'une image ultrasonore est donc un compromis entre résolution spatiale et profondeur d'exploration**. En outre,

[10] *speckle* peut être traduit en français par chatoiement. Dans la suite de ce document, nous utiliserons exclusivement le terme *speckle*.

la qualité de l'image ultrasonore dépend également du niveau de *speckle* qui l'entache.La section suivante décrit les spécificités de l'imagerie ultrasonore de l'appareil vocal.

### 1.4.2 L'imagerie ultrasonore de l'appareil vocal

Le *Vocal Tract Visualisation Lab* (VTVL) de l'université de Maryland (Baltimore - USA) dirigé par Madame Maureen Stone, est spécialisé dans l'**imagerie ultrasonore de l'appareil vocal**. Le projet dans lequel s'inscrit cette étude fait l'objet d'une étroite collaboration avec son équipe. Les sections suivantes décrivent les techniques d'imagerie de l'appareil vocal ; la discussion proposée s'appuie sur [Stone, 2003].

#### Protocole d'acquisition

Dans le cas très spécifique de l'imagerie ultrasonore de l'appareil vocal, le transducteur est placé sous la mâchoire. Ce positionnement de la sonde permet la visualisation de la langue, du palais, ainsi que d'un certain nombre de structures décrites ultérieurement. Le laboratoire VTVL a mis au point un système de fixation nommé HATS [11].

Ce système permet un contact quasi-parfait entre la sonde et la mâchoire, durant le processus d'élocution. Ce contact est primordial. En effet, une perte du contact acoustique entraîne systématiquement une perte de l'image. Ce système est présenté à la figure 1.2.

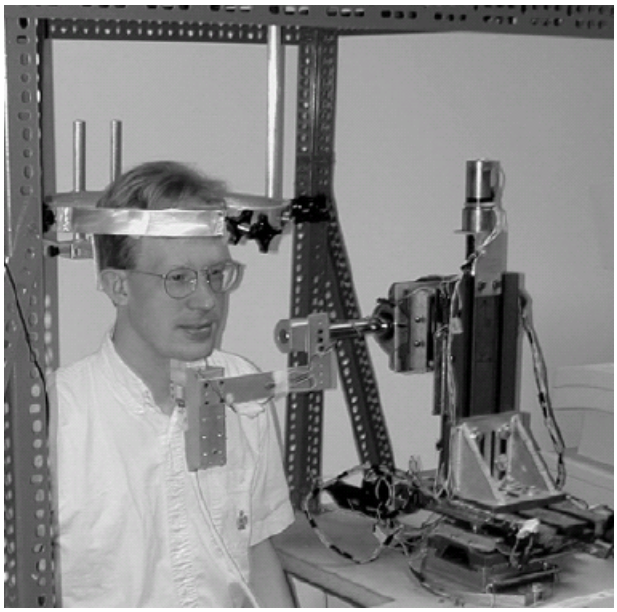

FIG. 1.2 – Système HATS - Head and Transducer

**La position de la sonde s'adapte au mouvement de la mâchoire**, sans perte de contact [12]. Cette technique innovante est un atout majeur du système HATS, système idéal pour des recherches en phonétiques articulatoires. De plus, un système d'imagerie optique, couplé au système ultrasonore permet de visualiser la tête du locuteur. Une acquisition classique effectuée au VTVL est reproduite sur la figure 1.3.

Dans le cadre de recherche en traitement de la parole, l'acquisition d'images ultrasonores de l'appareil vocal s'accompagne de l'enregistrement audio, synchrone avec la vidéo, du texte prononcé par le locuteur. La section suivante propose une série de remarques aidant à la compréhension de l'image ultrasonore de l'appareil vocal.

#### Interprétation des images ultrasonores de l'appareil vocal

Différentes structures sont observables lors de l'imagerie ultrasonore du conduit vocal en mouvement.

**La langue :** La figure 1.4 représente une acquisition ultrasonore d'un locuteur. La zone blanche, parfois discontinue, est la langue. **Le contour inférieur de cette zone est la surface supérieure de la langue**. En comparant cette image à celle de la figure 1.3, il est frappant de constater à quel point l'image ultrasonore de cette structure peut varier d'un locuteur à l'autre.

[11] Head and Transducer Support System

[12] Un **gel de contact** est utilisé afin de combler la fine couche d'air qui subsiste entre la sonde et la peau

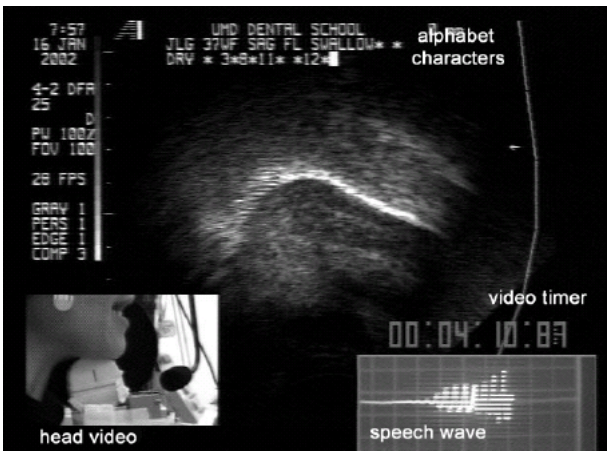


FIG. 1.3 – Imagerie ultrasonore de l'appareil vocal à l'aide du système HATS - Vocal Tract Visualization Lab

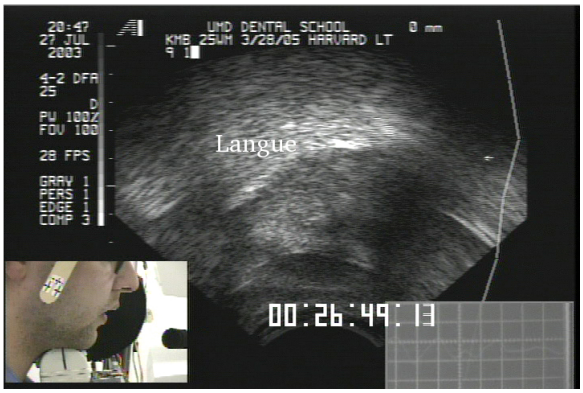


FIG. 1.4 – Imagerie ultrasonore - Mise en évidence de la langue

La langue apparaît ainsi comme un organe difficile à imager. Plusieurs éléments permettent d'expliquer ce phénomène. Tout d'abord, la langue est un organe complexe, composé majoritairement de graisse. Cette dernière contribue à la réfraction de l'onde ultrasonore et diminue l'échogénéicité de la langue. En outre, le caractère irrégulier de sa surface accentue le phénomène de réflexion diffuse, explicité à la section 1.4.1. En revanche, la salive environnante semble limiter son influence en « lissant »la surface de la langue. Une bouche sèche produit une mauvaise image de la langue.

De plus, la qualité de l'image varie d'un sujet à un autre. Un sujet mince possède une langue plus fine, avec moins de graisse ; l'image ultrasonore de sa langue est meilleure que celle obtenue avec un sujet plus gros. Les enfants et les femmes produisent également de meilleures acquisitions.

Enfin, la visibilité de la langue, lors du processus d'élocution, reste très variable. Un angle de plus 50 degrés entre la langue et le plan perpendiculaire au faisceau ultrasonore produit de très mauvaises images. Ce phénomène peut par exemple être observé pour le phonème /i/. A l'inverse, la réalisation du phonème /a/ positionne la langue perpendiculairement au faisceau ultrasonore donc l'image obtenue est bonne.

**Le palais :** Certains positionnements de la langue permettent la visualisation de l'os du palais, comme l'illustre la figure 1.5. L'os du palais est notamment visible lors de la déglutition. Le dos de la langue vient alors coller cette

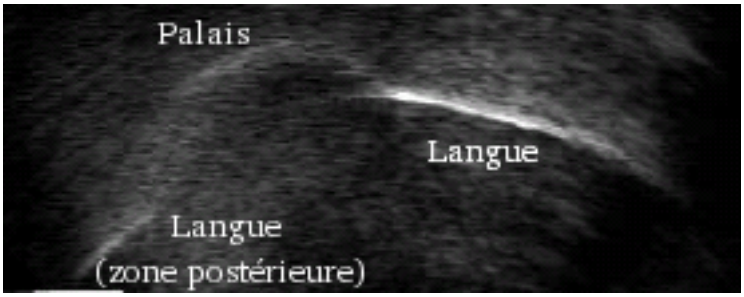


FIG. 1.5 – Imagerie ultrasonore - Mise en évidence du palais

structure, faisant disparaître ainsi la fine couche d'air qui les sépare normalement. Le coefficient de transmission d'une interface air-tissu étant très faible, cette couche d'air atténue fortement l'onde ultrasonore, qui ne peut donc pas atteindre, en temps normal, l'os du palais.

**La mâchoire :** Dans une image ultrasonore, la mâchoire n'est pas directement observable. Responsable de la réfraction quasi-totale de l'onde ultrasonore, elle est à l'origine d'une ombre acoustique. Cette dernière peut obscurcir l'apex (pointe de la langue). Généralement, un centimètre de la langue reste non observable. La figure 1.6 met en évidence ce phénomène.

**L'os hyoïde :** Tout comme la mâchoire, l'os hyoïde réfracte l'onde ultrasonore et engendre une ombre acoustique dans l'image. Cette ombre est plus fine que celle due à la mâchoire. A quelques rares occasions, une partie de l'onde réfractée peut être récupérée par le transducteur. L'os hyoïde peut alors apparaître comme un tache brillante aux contours flous. Cette ombre peut obscurcir la partie postérieure de la langue comme le montre la figure 1.6.

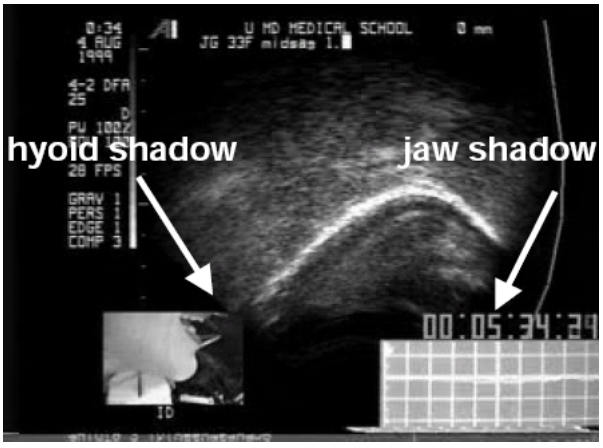


FIG. 1.6 – Imagerie ultrasonore - Ombre acoustique de la mâchoire et de l'os hyoïde.

## 1.5 Conclusion

Le système d'imagerie de l'appareil vocal développé au VTVL est un outil performant pour l'étude et la compréhension du mécanisme de production de la parole. L'acquisition synchrone d'images de la langue et des lèvres avec le signal de parole permet d'envisager une modélisation de la relation visio-acoustique, c'est-à-dire de la relation entre l'image du conduit vocal observée et le signal acoustique émis. C'est l'objectif du projet *Ouisper* que nous abordons maintenant.

# Chapitre 2

# Projet *Ouisper*

## 2.1 Introduction

Ce chapitre est consacré à une présentation du projet *Ouisper*, à ses objectifs, aux applications envisagées, à leurs solutions concurrentes existantes, et enfin, aux difficultés sous-jacentes à la résolution d'un tel problème. Afin de situer mon apport au déroulement du projet, je décrirai les études préliminaires menées par mon maître de stage Monsieur Bruce Denby.

## 2.2 Présentation générale du projet *Ouisper*

### 2.2.1 Objectif

Le projet *Ouisper* vise à réaliser un prototype de dispositif permettant de reconstituer un signal de parole intelligible, en temps réel, à partir de la saisie des mouvements de certains articulateurs du locuteur par imagerie ultrasonore et optique, sans activation des cordes vocales. Le système est destiné, à terme, à être léger et portatif, et servira en tant qu'**interface silencieuse de communication verbale**.

### 2.2.2 Applications envisagées

Un tel dispositif a été imaginé pour de possibles applications médicales. Les principaux bénéficiaires d'un système de synthèse de la parole à partir de la saisie du mouvement de l'appareil vocal, pourraient être les personnes ayant subi les interventions chirurgicales mentionnées au chapitre précédent. En outre, un dispositif permettant de communiquer à l'aide des mouvements habituels du conduit vocal, mais **sans nécessité de vocaliser la parole**, peut aussi ouvrir de multiples perspectives dans les domaines des télécommunications. Une interface de communication silencieuse est aujourd'hui une demande du milieu militaire et policier, pour des personnels opérant dans des situations où le silence est requis. Il en va de même dans le monde des télécommunications civiles qui pourrait également profiter d'un système de saisie silencieuse de données, utilisable, à terme, comme un téléphone silencieux. En effet, si nous sommes capables de synthétiser de la parole à partir d'un mouvement silencieux, nous pourrons alors communiquer à distance en toute confidentialité, sans polluer l'espace sonore.

### 2.2.3 Solutions existantes et concurrentes

Il n'existe, à notre connaissance, qu'un seul système de communication verbale silencieuse. Il s'agit du système fondé sur l'électromyographie et l'electropalatographie [Jorgensen *et al.*, 2003] développé par la DARPA[1] aux Etats-Unis. Les mesures de l'activité électrique des muscles activant le larynx et la zone sublinguale sont utilisées comme entrée d'un système de synthèse vocale. Peu d'informations sont disponibles sur ce système, destiné exclusivement aux applications militaires.

Pour le domaine médical, les techniques chirurgicales de reconstruction de l'appareil vocal nécessite un effort important de la part du patient, lors de la vocalisation. L'utilisation de prothèses externes(cf. 1.3.1), est aujourd'hui privilégiée, elles permettent la production d'une parole intelligible. Néanmoins, la voix de l'utilisateur manque de clarté et d'intelligibilité. Le dispositif proposé dans le cadre du projet *Ouisper*, apparaît comme une alternative intéressante.

[1] The Defense Advanced Research Projects Agency

En outre, ce type de problème commence à mobiliser la communauté scientifique du traitement de la parole. On notera notamment le projet « La Voix du Silence »qui associe des équipes des laboratoires ICP[2] et LIS [3]. L'objectif poursuivi est de retrouver la structure des messages émis dans les traces infimes des activités motrices recueillies par électromyographie laryngée ainsi que dans les sons murmurés transmis par vibration osseuse via la stéthoscopie maxillo-faciale. Ces données sont obtenues grâce à un dispositif mis au point au *Nara Institute of Science and Technology*, nommé NAM. Cette technique consiste à placer une capsule stéthoscopique en contact avec la peau juste en dessous de l'oreille. Cette approche diffère néanmoins de celle adoptée dans le projet *Ouisper*, puisqu'elle nécessite une activité de phonation, même infime.

## 2.3 Architecture du système d'analyse-synthèse *Ouisper*

### 2.3.1 Schéma général de fonctionnement

La figure 2.1 illustre la mise en œuvre d'un système de synthèse de la parole à partir de la saisie du mouvement de certains articulateurs par imagerie ultrasonore et optique. Dans cette configuration, une sonde ultrasonore fournit une image d'une partie de la cavité buccale, principalement de la langue. D'autres structures également visibles grâce à cette sonde sont décrites à la section 1.4.2. Une caméra optique conventionnelle renseigne sur le mouvement des lèvres. Le traitement synchrone des images ultrasonores et optiques pilote un synthétiseur de parole.

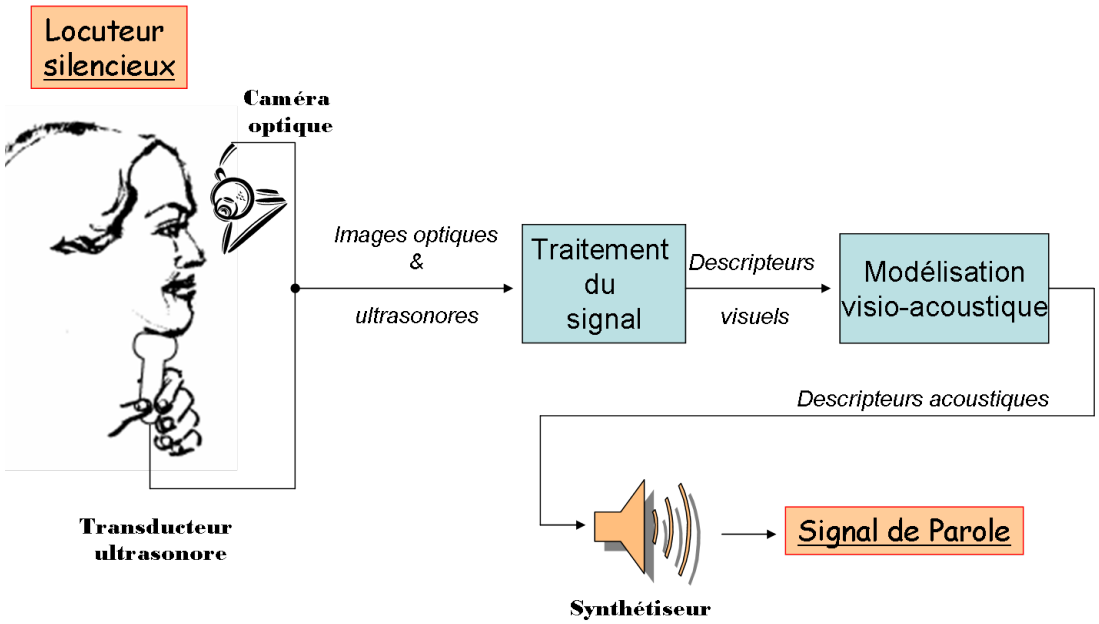


FIG. 2.1 – Schéma général de fonctionnement du système

La conception d'un tel dispositif nécessite une recherche approfondie dans les domaines suivants :
- Acquisition synchrone d'image ultrasonore et optique de l'appareil vocal
- Traitement et codage de l'information extraite des images.
- Modélisation de la relation "Image - Parole"
- Codage et synthèse d'un signal de parole intelligible

L'acquisition des données s'appuie sur les travaux du VTVL décrit au chapitre précédent.

D'autre part, la faisabilité du projet n'est envisageable qu'après avoir posé certaines restrictions et limites. Cette discussion est l'objet du paragraphe suivant.

### 2.3.2 Commentaires

L'étude du fonctionnement de l'appareil vocal, présentée au chapitre précédent, fait apparaître le rôle distinct de la glotte par rapport à celui des différents résonateurs (langue, lèvres, fosses nasales). L'activité de la glotte est directement liée à la prosodie[4] de la voix. Une voix monotone est difficilement intelligible. Cependant, le

[2]Institut de Communication Parlée - Grenoble
[3]Laboratoire des Images et des Signaux - Grenoble
[4]On appelle prosodie, l'intonation et l'évolution des différents accents toniques dans la voix. Le débit de parole, la fréquence fondamentale, l'énergie et le contexte syntaxique définissent la prosodie.

dispositif envisagé est un système de communication silencieux, qui ne nécessite pas de vocalisation, et *a fortiori* aucune activité glottale. Aussi, reconstituer un signal de parole intelligible, sans ce type d'information, est un véritable problème.

D'autre part, les techniques d'imagerie ultrasonore et optique de l'appareil vocal, décrites au chapitre précédent, fournissent une information sur le mouvement de la langue et sur celui des lèvres. Or, d'autres résonateurs jouent un rôle dans la production de la parole. Ce manque d'information constitue une difficulté supplémentaire. Enfin, la réalisation concrète d'un tel dispositif nécessite l'intégration d'un système d'imagerie dans un boîtier portatif.

## 2.4 Contexte du stage

Afin d'appréhender au mieux les objectifs de mon stage de Master Recherche, les études préliminaires menées par Monsieur Denby sont brièvement présentées dans la section suivante.

### 2.4.1 Résultats préliminaires

Dans [Schweitzer *et al.*, 2003], Schweitzer et Denby, propose un système d'étude de la langue par imagerie ultrasonore. Par la suite, Denby et Stone définissent dans [Denby *et al.*, 2006] [Denby et Stone, 2004] l'architecture d'un système de synthèse de la parole à partir d'images ultrasonores et optiques de l'appareil vocal.

Dans [Denby *et al.*, 2006], un algorithme d'extraction de contour a été employé afin de décrire la position de la langue dans chaque image ultrasonore et celle des lèvres dans chaque image optique. Le contenu spectral du signal de parole enregistré, est décrit de manière synchrone à la vidéo, à l'aide des coefficients LSF (cf. 4.2.3). Une modélisation de la relation liant la position de la langue et des lèvres aux coefficients LSF est effectuée à l'aide d'un réseau de neurones. Deux minutes d'acquisition ont été utilisées pour cette modélisation que nous appellerons **modélisation visio-acoustique**. Le modèle ainsi constitué est capable de prédire une description acoustique de la parole à partir des images de l'appareil vocal. Un signal de parole est reconstitué à partir de cette prédiction.

Cette première approche produit des résultats prometteurs. Elle a d'ailleurs été saluée dans les conférences ICASSP 2004 et 2006. Certes l'intelligibilité de la parole synthétisée n'est pas suffisante, mais certains sons semblent correctement modélisés. En outre, une base de données de deux minutes d'acquisition a été utilisée pour cette approche. Une si petite base ne contient que trop peu de contextes linguistiques différents pour espérer une bonne modélisation. Aussi, des acquisitions plus longues ont été effectuées au VTVL, sur un autre locuteur. L'étude présentée dans ce document s'appuie sur [Denby *et al.*, 2006] pour traiter ces nouvelles données.

### 2.4.2 Objectifs du stage

Les objectifs définis en début de stage ont été les suivants :

- Conception d'une base de données informatique à partir des séquences audio et vidéo fournies par le VTVL (durée : une heure)
- Mise au point de nouveaux algorithmes de traitements des images, capables de traiter les nouvelles données.
- Mise en œuvre d'autres méthodes de description du signal de parole

Ces objectifs se sont affinés au cours de ces six mois de travail, la recherche s'effectuant sur l'ensemble du projet et non sur un aspect particulier (image, parole, modélisation).

## 2.5 Conclusion

En s'inscrivant dans la continuité de [Denby *et al.*, 2006] et de [Denby et Stone, 2004], cette étude tente d'apporter des solutions différentes aussi bien dans le traitement d'image que dans celui du traitement de la parole, dans le but d'exploiter une grande quantité de nouvelles données. Le chapitre suivant présente les différentes méthodes développées pour l'analyse des images.

# Chapitre 3

# Traitement des images ultrasonores et optiques de l'appareil vocal

## 3.1 Introduction

Une nouvelle base de données est constituée à partir des enregistrements fournis par le du système d'acquisition multimodal développé au VTVL. L'organisation de ces données est l'objet de la première section de ce chapitre. Les sections suivantes sont consacrées au traitement et à la description des images ultrasonores et optiques.

## 3.2 Constitution de la base de données *Ouisper*

### 3.2.1 Le corpus IEEE/Harvard

L'acquisition de données à l'aide du système développé au VTVL nécessite le choix d'un corpus de texte, prononcé par un locuteur lors de la saisie des mouvements de la langue et des lèvres par imagerie ultrasonore et optique. Le choix de ce corpus est important, car ce dernier doit présenter une couverture suffisante de l'espace phonétique. En effet, la finesse de la modélisation du lien visio-acoustique par apprentissage artificiel dépend de la représentativité des configurations articulatoires étudiées.

Le corpus de texte IEEE/Harvard [IEEE, 1969] a été choisi pour la constitution de cette base de données. Ce corpus est constitué de 72 listes de 10 phrases en langue anglaise. La grande majorité de ces 720 phrases présente la même structure grammaticale, environ le même nombre de syllabes ainsi que le même nombre de mots. Ces caractéristiques globales contraignent le locuteur à utiliser la même intonation ou prosodie lors de la prononciation de ces 720 phrases. Ce point est important dans le cadre de notre étude. Comme nous l'avons indiqué à la section 2.3.2, en l'absence d'activité glottale, la gestion de la prosodie semble délicate. Aussi, l'étude d'exemples présentant la même prosodie, peut permettre de limiter l'influence de cette dernière dans la modélisation visio-acoustique. De plus, les phrases du corpus IEEE/Harvard sont dites phonétiquement équilibrées. La proportion de chacun des phonèmes présents dans chacune des phrases est cohérente avec celle observée dans un corpus beaucoup plus important, donc dans la langue anglaise en général. Ce dernier point est également un atout pour une exploration profonde de l'espace des configurations articulatoires possibles de l'appareil vocal.

### 3.2.2 Contenu de la base de données *Ouisper*

Un locuteur masculin d'une trentaine d'années a été choisi pour prononcer le corpus IEEE/Harvard. Le système d'acquisition du VTVL a été utilisé afin de saisir les mouvements de son appareil vocal lors de l'élocution. Une heure d'acquisition a été effectuée. Le tableau ci-dessous décrit ses caractéristiques.

| | |
|---|---|
| Format du flux vidéo | DV |
| Format du flux audio | PCM WAV |
| Résolution et fréquence d'échantillonnage du flux audio | 44.1 kHz - 16 bits |
| Encapsulation | Quicktime |
| Résolution | $640 \times 480$ |
| Entrelacement | Non |
| Fréquence d'échantillonage | 29.97 fps |

L'illustration de la figure 1.4 est une trame de la vidéo ainsi obtenue. Un travail de préparation des données vidéo brutes est indispensable. Ce pré-traitement consiste à :

- Supprimer les segments qui ne correspondent pas au processus d'élocution (déglutition, toux, etc...)
- Segmenter la vidéo en 720 fichiers afin d'obtenir un fichier vidéo par phrase.
- Décomposer la vidéo en une série d'images au format JPEG.
- Extraire les pistes audio afin d'obtenir un fichier audio par phrase.
- Choisir un système de fichiers cohérent, permettant de définir rapidement des sous-bases d'apprentissage, de validation et de test pour la modélisation visio-acoustique.

Le traitement d'une heure de vidéo selon ce schéma est long et fastidieux. J'ai donc été amené à développer une méthode de traitement automatique, qui pourra être réutilisée lors d'acquisitions ultérieures. L'annotation des segments à conserver a été effectuée grâce au logiciel multi-plateforme *Wavesurfer*[1] développé par le *Royal Institute of Technologie* (KTH) de Stockholm. La segmentation vidéo et la conversion en une série d'images ont été effectuées grâce au logiciel libre scriptable *Transcode* [2]. Les différents traitements ainsi que le classement des données ont été réalisés à l'aide d'un programme *Perl*.

Le tableau ci-dessous fournit des informations quantitatives sur la base de données ainsi constituée.

| | |
|---|---|
| Nombre d'images | 72473 |
| Temps de parole | 40 minutes et 18 secondes |
| Espace disque | 18 Go environ |

En vue d'une modélisation de la relation visio-acoustique, il est indispensable d'extraire des images ultrasonores et optiques de l'appareil vocal une série de descripteurs pertinents, témoignant de l'état du conduit vocal lors de la phonation. Cette étape est l'objet des sections suivantes.

## 3.3 Traitement des images ultrasonores

### 3.3.1 Pré-traitement

Dans le cadre de l'imagerie ultrasonore, il est nécessaire de séparer le pré-traitement de l'image (corrections géométriques, filtrage) de leur description. Deux techniques successives de pré-traitement font l'objet des paragraphes suivants.

#### Dépliement

La forme de l'image échographique dépend du type de sonde utilisée. On distingue deux catégories de sondes échographiques.

- Les sondes pour lesquelles chaque cristal ou groupe de cristaux émet des ultrasons toujours dans la **même direction**. L'image est formée par la juxtaposition des lignes formées par l'excitation successive de chaque cristal ou groupe de cristaux. Ces sondes sont dites linéaires.
- Les sondes pour lesquelles les ultrasons sont envoyés dans **une direction différente** à chaque impulsion au moyen d'un balayage mécanique ou électronique. Ces sondes sont dites à balayage.

Les sondes linéaires et linéaires courbes sont des sondes constituées de plusieurs cristaux alignés en rangée linéaire ou légèrement courbées. Les sondes linéaires présentent l'avantage d'utiliser des ultrasons ayant tous la même direction. Ce dernier élément est déterminant pour l'exploration des structures anisotropes, comme les tendons, pour lesquelles l'échogénicité est fortement influencée par l'orientation des ultrasons. En revanche, ces sondes ne sont pas adaptées à l'imagerie ultrasonore de l'appareil vocal en raison de la forme courbe de la surface inférieure du menton, sur laquelle est placée la sonde.

Les sondes à balayage sont des sondes constituées d'un ou de plusieurs cristaux pour lesquelles la direction du faisceau ultrasonore varie afin de balayer la zone à explorer. Ce balayage n'est pas directement visible sur l'image reconstituée, car il est trop rapide pour être perçu et l'opérateur a l'impression d'une image en temps réel. **L'image reconstituée à partir d'une sonde à balayage a une forme de cône**. Il existe deux grands types de balayage : le balayage mécanique et le balayage électronique.

Une sonde à balayage mécanique est composée d'un ou de plusieurs cristaux montés sur un support qui oscille lors de l'acquisition. Le cristal baigne en général dans un liquide pour permettre son mouvement qui est souvent perceptible lorsqu'on place sa main sur la sonde.

Les sondes à balayage électronique sont constituées de plusieurs cristaux arrangés en ligne ou en anneau. Des interférences entre les faisceaux ultrasonores des cristaux élémentaires peuvent faire changer la direction générale du faisceau. Ces interférences peuvent être utilisées avantageusement pour orienter le faisceau ultrasonore résultant, dans une direction donnée, en décalant très légèrement la mise en charge des différents cristaux

[1] Wavesurfer : http ://www.speech.kth.se/wavesurfer
[2] *Transcode* : http ://www.transcoding.org

de la sonde. Le changement de direction du faisceau ultrasonore est obtenu en modifiant le décalage de la mise en charge des différents cristaux par un décalage de phase de l'impulsion électrique.

L'avantage majeur des sondes à balayage électronique par rapport aux sondes à balayage mécanique est de pouvoir reconstituer en temps réel et en même temps plusieurs modes (Mode B et TM). **Dans le cadre de ce document, la sonde utilisée est à balayage électronique.**

Le premier traitement de l'image ultrasonore est une transformation géométrique visant à **déplier la région d'intérêt conique** pour permettre une visualisation dans un plan cartésien. Cette étape est motivée par les considérations suivantes :

- La visualisation de l'objet étudié (image cartésienne) est incohérente avec la géométrie du système d'acquisition (faisceau conique).
- Les ombres acoustiques de la mâchoire et de l'os hyoïde délimitent de manière « naturelle »la région d'intérêt, à savoir la zone d'activité de la langue.
- Une éventuelle boîte englobante rectangulaire, entourant le cône ultrasonore, présenterait de larges zones inutiles.

L'image ultrasonore de forme conique est ainsi discrétisée sur une grille dont la forme est représentée par la figure 3.1.

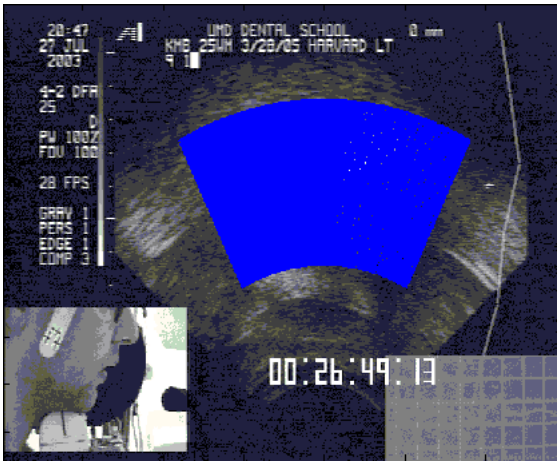


FIG. 3.1 – Dépliement de l'image ultrasonore conique - Grille de discrétisation

Cette grille s'appuie sur les ombres acoustiques de la mâchoire et de l'os hyoïde. Elle est définie par l'ensemble des points de coordonnées $(r, \theta)$, comme le montre la figure 3.2.

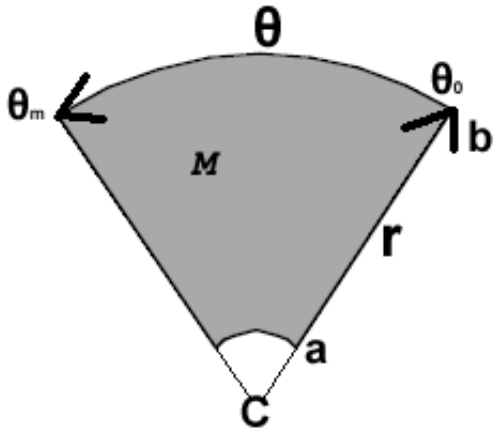


FIG. 3.2 – Détail de la grille de discrétisation

Les coordonnées polaires $(r, \theta)$ de chaque pixel **M** sont définies à partir des coordonnées cartésiennes $(u, v)$ et du point **C** (cf figure 3.2) de coordonnées $(x_0, y_0)$ :

$$\begin{cases} r = \sqrt{(u - x_0)^2 + (v - y_0)^2} \\ \theta = \arctan(\frac{v - y_0}{u - x_0}) \bmod 2\pi \end{cases} \tag{3.1}$$

La discrétisation, à proprement parler, s'effectue en associant au point $\mathbf{M}(r,\theta)$ vérifiant $a \leq r \leq b$ et $\theta_0 \leq \theta \leq \theta_m$ la moyenne arithmétique de l'intensité de l'ensemble des points $\mathbf{E}(\rho,\varphi)$ tels que : $r \leq \rho \leq r+dr$ et $\theta \leq \varphi \leq \theta+d\theta$. La finesse $(dr, d\theta)$ de la grille de discrétisation est un paramètre de la méthode. La figure 3.3 illustre le résultat de cette discrétisation sur une grille de taille $(50 \times 50)$.

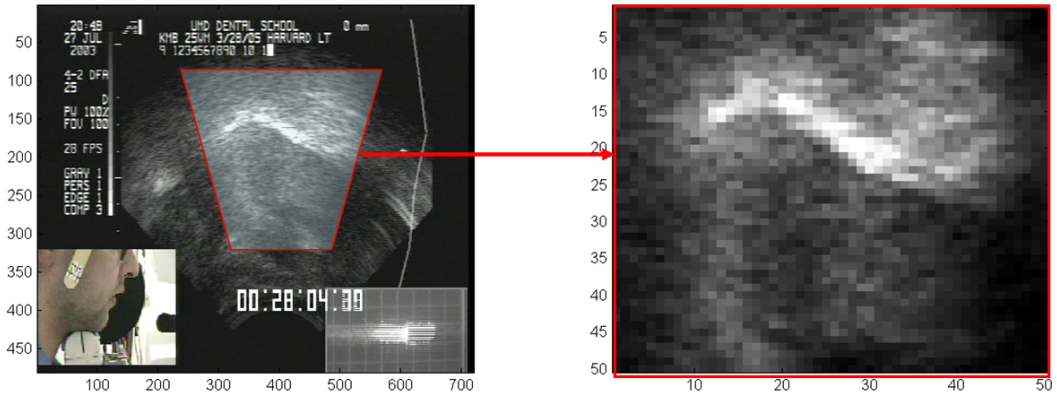


FIG. 3.3 – Discrétisation sur une grille $(50 \times 50)$.

Ainsi, la discrétisation permet de « déployer »l'image ultrasonore conique et de définir une région d'intérêt cohérente avec la géométrie du système d'acquisition. De plus elle réduit considérablement la dimensionnalité de l'espace image.

**Etude du *speckle***

Les images échographiques ont des propriétés particulières, différentes des images optiques. Après avoir étudié leur formation, il est important de prendre en considération le bruit qui les affecte, le *speckle*.

Rappelons, tout d'abord, que deux types de signaux peuvent être perçus par le transducteur ultrasonore :

- Les échos qui proviennent des transitions d'impédance acoustique (cf 1.4.1)
- Les ondes rétrodiffusées qui proviennent des diffuseurs distribués aléatoirement dans le milieu. Ces derniers sont à l'origine du *speckle*.

L'aspect granuleux de l'image ultrasonore, dont la présence du bruit de *speckle* est responsable, doit être atténué afin de faciliter le paramétrage des structures de l'image. La modélisation du bruit de *speckle*, nécessaire au choix des techniques de filtrage, fait l'objet d'une littérature abondante. La thèse de Clovis Tauber [Tauber, 2005], fournit une excellente synthèse des différents modèles existants. Dans le cadre de cette étude, nous proposons d'illustrer deux modèles de *speckle* : le plus simple, le modèle de Rayleigh et un modèle plus général, celui de Rice.

Le modèle de Rayleigh a été introduit initialement dans une étude du *speckle* dans les images acquises par laser [Goodman, 1975]. Il suppose un grand nombre de rétro-diffuseurs à des distances mutuelles petites devant la longueur d'onde du signal. Goodman montre que la densité de probabilité de l'intensité du *speckle* s'écrit :

$$P_X(x) = \frac{x}{\sigma^2}.\exp(-\frac{x^2}{2\sigma^2}) \tag{3.2}$$

Dans ce modèle, la fonction de densité de probabilité de l'intensité du *speckle* suit une distribution de Rayleigh (cf figure 3.4), la variance $\sigma^2$ est l'énergie de rétrodiffusion moyenne, c'est le seul paramètre de ce modèle. Le modèle de Rayleigh modélise le *speckle* pur. En effet, la distribution spatiale des diffuseurs est aléatoire et non corrélée.

Le modèle de Rayleigh échoue lors de la présence d'une composante cohérente due à la présence d'une structure régulière de diffuseurs dans le milieu. Le modèle de Rice, plus élaboré, est décrit dans [V.Dutt et G.Greenleaf, 1995]. Ce modèle intègre la présence d'une composante cohérente dans la répartition des diffuseurs. Dans ce modèle, la densité de probabilité de l'intensité du *speckle* suit une distribution de Rice :

$$P_X(x) = \frac{x}{\sigma^2}.\exp(-\frac{x^2+|v|^2}{2\sigma^2})I_0\frac{x.|v|}{\sigma^2} \tag{3.3}$$

Cette distribution est représentée par la figure 3.5, pour une variance unitaire et différentes valeurs de la constante $v$.

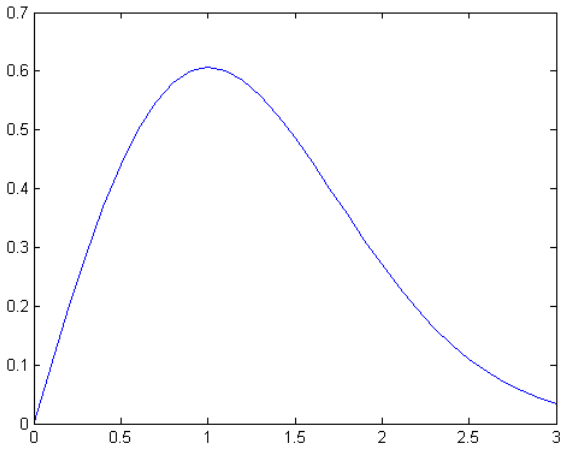


FIG. 3.4 – Fonction de densité de probabilité de Rayleigh

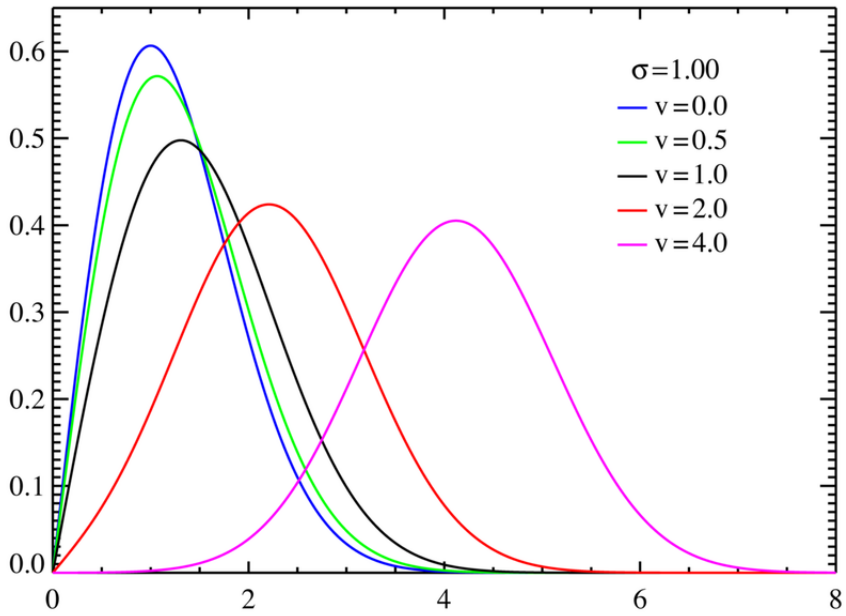


FIG. 3.5 – Fonction de densité de probabilité de Rice

Ce modèle est une généralisation du modèle de Rayleigh [3] ; il est défini par deux paramètres $(v, \sigma)$, il modélise de manière plus réaliste le *speckle*.

Afin d'étudier la nature du bruit présent dans les images ultrasonores **non re-discrétisées** de l'appareil vocal, une zone de *speckle* est isolée. La répartition des intensités des pixels de cette zone est modélisée à l'aide d'une distribution de Rayleigh et d'une distribution de Rice. Cette manipulation est réalisée à l'aide de l'outil *Distribution Fitting Tool* de *Matlab*. Les résultats de cette étude sont tracés à la figure 3.6.

Les diffuseurs n'étant pas répartis de façon complètement aléatoire dans la cavité buccale, le modèle de Rayleigh semble moins performant que celui de Rice. Ce dernier, décrit de manière plus réaliste le *speckle*. des images ultrasonores de l'appareil vocal. Cependant, il est important de renouveler l'expérience sur une image discrétisée sur la grille polaire décrite précédemment. En effet, les différents traitements ultérieurs seront appliqués exclusivement sur ces images. La figure 3.7 permet de conclure sur une éventuelle modification de la nature du *speckle* lors de la discrétisation.

De nouveau, la répartition observée semble suivre une distribution de Rice. Aussi, le *speckle* observé dans les acquisitions ultrasonores de l'appareil vocal peut être correctement décrit par un modèle de Rice, y compris après la discrétisation de l'image. **La discrétisation n'a donc pas modifié la nature du bruit original**. Cette conclusion est utile pour le choix d'un filtre adéquat. Ce dernier devra être dédié au traitement particulier des images échographiques. Son choix est l'objet du paragraphe suivant.

[3] Pour $v = 0$, nous retrouvons l'équation 3.2

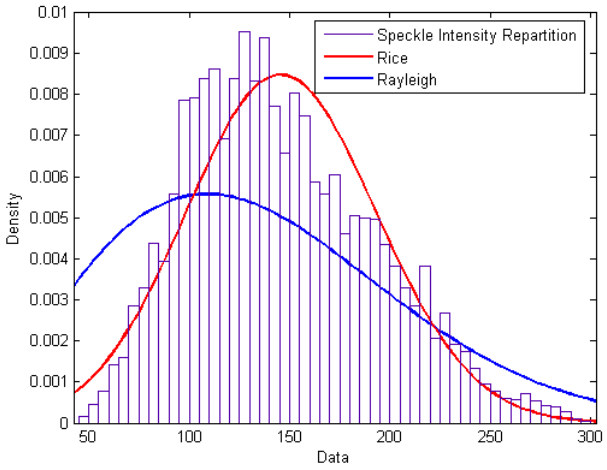


FIG. 3.6 – Modélisation de la répartition de l'intensité d'une zone de *speckle* dans une image **non re-discrétisée**

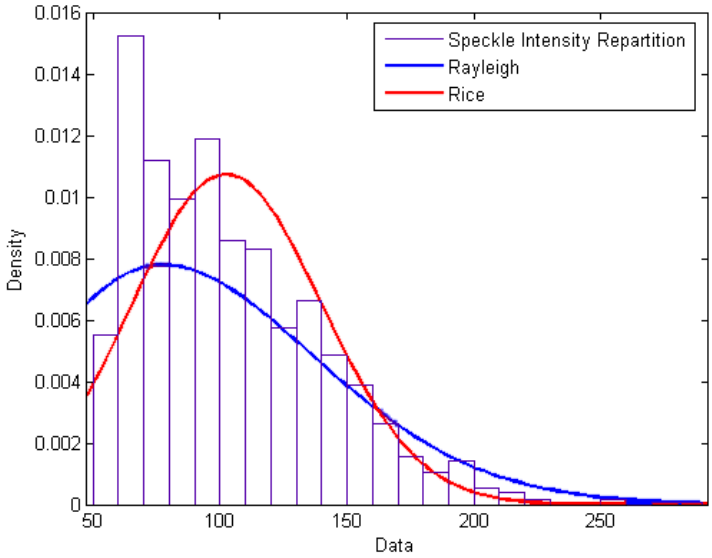


FIG. 3.7 – Modélisation de la répartition de l'intensité d'une zone de *speckle* dans une image **re-discrétisée**

### Filtrage du *speckle*

L'objectif recherché est de réduire le *speckle* dans les **images discrétisées** afin de restaurer l'information utile dans l'image. Pour cela, il est nécessaire de :

- lisser le mieux possible les zones homogènes
- préserver les contours et les structures de l'image
- conserver la texture là où elle est présente

Le filtrage du *speckle* a largement été étudié. Dans le cadre du traitement des images échographiques, nous distinguerons les filtres dits « classiques »(cf [Lee, 1980], [Kuan *et al.*, 1985]), et les filtres dits « de diffusion anisotrope ».

Les filtres classiques, décrits par Lee, utilisent la mesure d'un indice statistique nommé **Coefficient de Variation**. On distingue deux types de coefficients de variation :

- Le coefficient de variation global noté $\Gamma$
- Le coefficient de variation local noté $\gamma$

Le coefficient de variation global est définit par :

$$\Gamma^2 = \frac{var(I_\Omega)}{\bar{I}_\Omega^2} \tag{3.4}$$

avec $\Omega$ l'ensemble des pixels d'une zone homogène, $I$ l'intensité et $\bar{I}_s$ l'intensité moyenne des pixels sur cette zone. Tauber [Tauber, 2005] montre que $\Gamma$ est une caractérisation globale du *speckle* présent dans l'image.

Le coefficient de variation local est défini comme une estimation locale du coefficient de variation global :

$$\gamma^2(s) = \frac{1}{|\eta_s|} \sum_{p \in \eta_s} \frac{(I_p - \bar{I}_s)^2}{(\bar{I}_s)^2} \tag{3.5}$$

où $\eta_s$ est le voisinage du pixel central $s$ et $\bar{I}_s$ l'intensité moyenne des pixels sur ce voisinage. Lee démontre que :

- $\gamma \sim \Gamma$ dans une fenêtre contenant des pixels d'une région homogène
- $\gamma \gg \Gamma$ dans une fenêtre contenant des pixels d'un contour

**Le coefficient de variation local est donc un détecteur de contour** dans les images comportant du *speckle*. Le filtre de Lee effectue un lissage classique sur les zones homogènes détectées par mesure statistique du coefficient de variation local et préserve ainsi les contours.

L'autre catégorie de filtres adaptés aux images contenant du *speckle* est celle des **Filtres de diffusion anisotrope**. Introduit par Perona et Malik [Perona et Malik, 1990], cette technique est brièvement décrite ci-après.

Le principe du filtre de diffusion anisotrope est de contrôler le lissage de manière progressive, isotrope dans les zones de réflectivité (intensité des pixels) homogène et anisotrope sur les contours. Il s'agit de produire une image plus lisse avec moins d'extrema locaux.

Le filtre de diffusion anisotrope s'appuie sur l'utilisation des **équations aux dérivées partielles** (EDP). Ce type d'équation, d'application courante dans divers domaines de la physique, est également important en traitement du signal et de l'image.

L'équation de diffusion de la chaleur est un exemple classique d'EDP. La quantité de chaleur u(x,t) en un point x à un instant t est régie, dans un milieu isotrope, par l'équation :

$$\begin{cases} \frac{\partial u}{\partial t} = \Delta u \\ u(x,0) = u_0(x) \end{cases} \tag{3.6}$$

en notant $\Delta u$ le laplacien de $u$.

La solution de cette équation **linéaire** décrit la propagation au cours du temps de la chaleur dans un milieu isotrope. Lorsque cette équation est appliquée à une image $u$, $t$ n'est plus le temps mais un **paramètre d'échelle**. A partir d'une image $u_0$, la résolution de cette équation fournit, à l'échelle $t$, les images simplifiées $u(.,t)$. La simplification d'une image par résolution de cette équation de diffusion, lisse l'ensemble de l'image, le bruit comme les contours des structures homogènes. Pour preuve, remarquons que l'équation de diffusion de la chaleur peut s'écrire :

$$\frac{\partial u}{\partial t} = \Delta u = div(\nabla u) \tag{3.7}$$

Cette équation nécessite une diffusion identique en tout point de l'image (milieu isotrope). L'idée de Perona et Malik est de lisser l'image dans les zones homogènes, et de ne pas faire évoluer l'image le long des contours. Ils proposent l'équation suivante :

$$\frac{\partial u}{\partial t} = div(g(|\nabla u|)\nabla u) \tag{3.8}$$

où $g$ est une fonction décroissante, valant 1 en 0, et tendant vers 0 en l'infini. Les conditions initiales sont simples, $u(.,t=0) = u_0$. Une des fonctions $g$ proposée dans [Perona et Malik, 1990] est :

$$g(s) = \frac{1}{1 + (\lambda s)^2} \tag{3.9}$$

où $\lambda$ est une constante.

Dans la méthode de lissage (ou filtrage) par diffusion anisotrope le gradient de l'image $u(x,t)$ est utilisé comme détecteur de contour. Aussi, lorsque $|\nabla u| \gg \lambda$ (sur un contour) alors $g(|\nabla u|)$ tend vers 0. La méthode se comporte alors comme un filtre passe-tout, l'intensité des pixels concernés n'est pas modifiée. Lorsque $|\nabla u| \ll \lambda$ alors $g(|\nabla u|)$ tend vers 1. Nous retrouvons alors l'équation de la chaleur et la méthode effectue une diffusion isotrope (lissage).

Dans [Yu et Acton, 2002], Yu propose une méthode de filtrage combinant les deux approches. La méthode ne sera pas décrite en détail dans ce document, mais nous pouvons néanmoins observer l'équation aux dérivées partielles non-linéaire introduite par l'auteur :

$$I_s^{t+\triangle t} = I_s^t + \frac{\triangle t}{|\bar{\eta}_s|} div[c(\gamma_s^t)\nabla I_s^t] \tag{3.10}$$

où $\eta_s$ est le voisinage du pixel central $s$, $\gamma(s)$ le coefficient de variation local au pixel $s$, $c$ une fonction assimilable à la fonction $g$ (*cf* équation 3.9) et $\nabla I_s^t$, le gradient de $I$ sur le voisinage $\eta_s$ et $t$ est le paramètre d'échelle.

Yu introduit une approche dynamique du coefficient de variation et définit ainsi un coefficient instantané de variation ($\gamma_s^t$). Grâce à [Yu et Acton, 2002], nous avons implémenté le filtre de diffusion anisotrope. Le code source est fourni en annexe A). Ce filtre est appliqué aux images échographiques de l'appareil vocal, les résultats obtenus sont illustrés par les figures 3.8 à 3.12 [4].

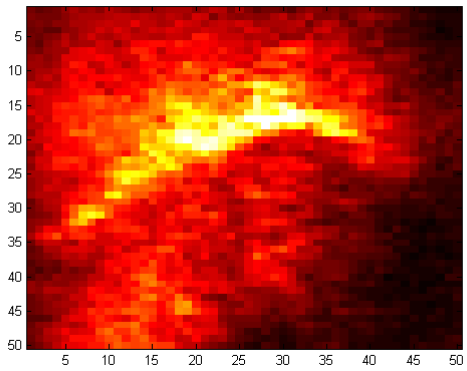


FIG. 3.8 – Image originale après discrétisation

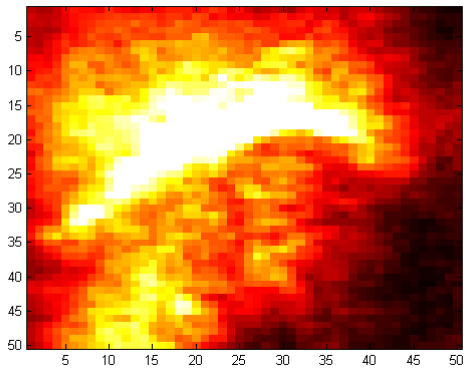


FIG. 3.9 – Filtre de diffusion anisotrope - Paramètre d'échelle $t = 5$

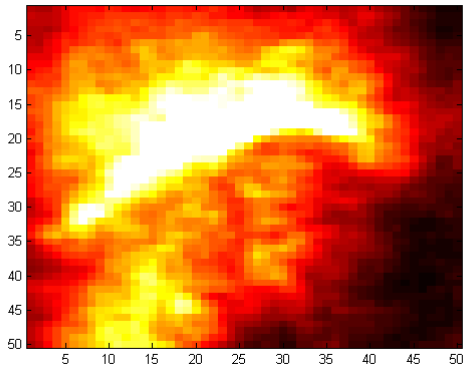


FIG. 3.10 – Filtre de diffusion anisotrope - Paramètre d'échelle $t = 10$

Un profondeur d'échelle de 40 suffit à lisser le *speckle* tout en préservant les contours des structures importantes. Nous pouvons à présent décrire les images ultrasonores.

[4] Une carte de couleur « chaude »a été utilisée à la place des niveaux de gris habituels, afin de permettre une meilleure visualisation des résultats du filtrage.

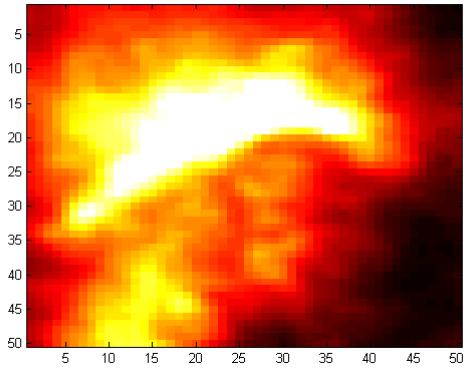


FIG. 3.11 – Filtre de diffusion anisotrope - Paramètre d'échelle $t = 20$

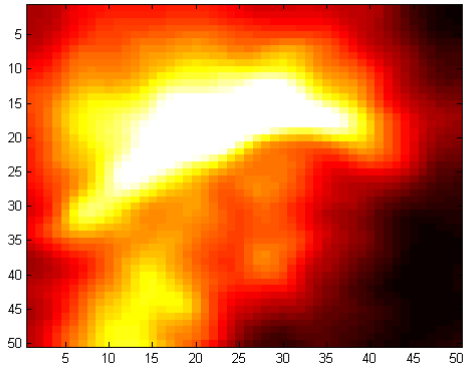


FIG. 3.12 – Filtre de diffusion anisotrope - Paramètre d'échelle $t = 40$

### 3.3.2 Extraction du contour de la langue

L'étude préliminaire (*cf* [Denby *et al.*, 2006], [Denby et Stone, 2004]), décrite dans 2.4.1), propose un algorithme de description de l'image ultrasonore basé sur le **l'extraction du contour de la langue**. L'application de cet algorithme sur les nouvelles données n'a pas donné les résultats escomptés. Les images ultrasonores de la nouvelle base de données se distinguent de celles constituant la base utilisée pour [Denby *et al.*, 2006]. D'après 1.4.2, nous observons que la langue de ce nouveau locuteur semble être plus difficile à imager. De plus, le contraste entre la langue et les autres structures y est plus faible. Une méthode basée sur les **contours actifs** [5], proposée par l'équipe du VTVL [M.Li *et al.*, 2003] a été expérimentée. Cette méthode nécessite une initialisation manuelle du contour de la langue, propriété que nous considérons comme un inconvénient. De plus, les résultats obtenus n'ont pas été satisfaisants. Aussi, un nouvel algorithme a été développé ; il est décrit dans le paragraphe suivant.

L'objectif reste l'extraction du contour de la langue. La méthode est fondée sur :
- une recherche des points candidats au contour
- un filtrage grossier des points candidats
- un affinement progressif du contour de la langue

**Recherche des points candidats au contour** D'après la discussion de la section 1.4.2, le contour recherché est le bord inférieur de la zone échogène principale de l'image. Aussi, les points candidats au contour sont définis comme les extrema du gradient calculé selon la direction verticale, c'est-à-dire parallèlement au faisceau ultrasonore.

Cependant, cette recherche ne permet pas de différencier, les points du contour de la langue de ceux des structures voisines. Un filtrage « grossier », permet d'éliminer les points erronés. Il consiste à ne garder que les points dont l'intensité est supérieure à la moyenne de l'intensité de l'ensemble des candidats. Cette méthode est empirique mais néanmoins efficace ; elle est illustrée à la figure 3.13.

Le schéma de gauche explicite la recherche des points candidats, celui le droite met en évidence les résultats obtenus par ce filtrage « grossier ».

[5] introduite en 1987 par Kass, cette technique est également appelée *snake*

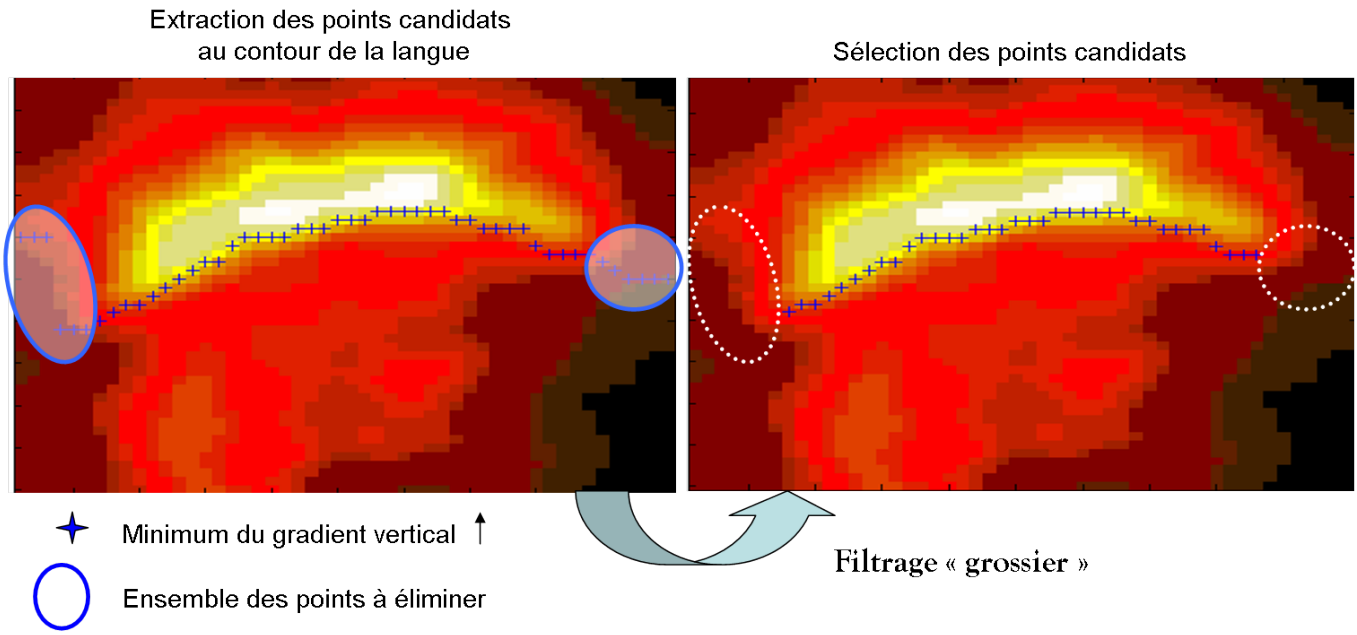


FIG. 3.13 – Extraction et sélection des points candidats au contour de la langue

Toutefois, pour certaines positions de la langue, l'image ultrasonore est très difficile à exploiter. Ce problème a été abordé à la section 1.4.2. Lorsque l'apex se relève pour approcher le palais ou les dents, la langue forme un angle faible avec la direction du faisceau ultrasonore, qui n'est presque pas réfléchi directement. En revanche,le phénomène de réflexion diffuse est amplifié et l'image devient alors très bruitée. Le lissage effectué par le filtre de diffusion anisotrope devient alors très important ; la figure 3.14 met en évidence ce phénomène. La recherche de points candidats au contour de la langue devient alors délicate ; la même figure 3.14 présente les résultats obtenus par cette méthode pour ce type d'image « pathologique ».

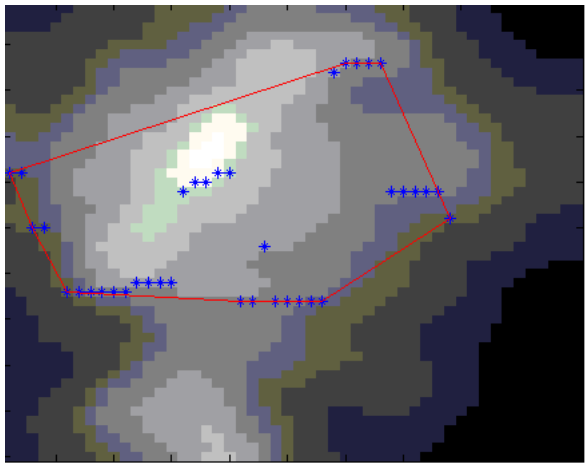

FIG. 3.14 – Cas pathologique pour l'extraction du contour de la langue

Le polygone rouge est l'enveloppe convexe des points candidats restant **après** l'étape de filtrage « grossier ». **Le calcul de l'aire de l'enveloppe convexe permet d'accéder à une bonne mesure de la dispersion des points candidats** : c'est un indicateur de la « mauvaise »qualité de la trame à traiter. Les points visiblement erronés sont appelés ***outliers***. L'algorithme proposé dans ce document tente justement de résoudre ces cas pathologiques.

**Chaînage des points candidats et élimination des points aberrants** **L'algorithme proposé ici vise à approcher le contour inférieur de la langue à l'aide d'une spline**[6]. Ces courbes possèdent les propriétés intéressantes suivantes :

- Si une spline $S$ de degré $n$ est définie sur $[a, b]$ alors $S$ est une fonction de classe $C^{n-1}$ sur cet intervalle ($C^{n-1}$ est l'ensemble des fonctions $n-1$ fois dérivables, et de dérivée $n-1$ continue sur $[a, b]$)

[6] Un fonction spline est une fonction définie par morceaux par des polynômes.

– Si $a = t_0 \leq t_1 \leq ... \leq t_{k-1} = b$ alors sur chaque sous-intervalle $[t_i, t_{i+1}]$ $S$ est un polynôme de degré $n$.
– La spline est parfaitement définie à partir des points de contrôle $(t_i, S(t_i))$.

Ce type de courbe hérite donc des propriétés de **continuité** et de **régularité** des polynômes. De plus, elle se définit à partir de quelques points de contrôle. L'interpolation à l'aide d'une spline est préférée à l'interpolation polynomiale car des résultats similaires sont obtenus avec des polynômes de degrés inférieurs.

L'algorithme vise à résoudre **conjointement** le problème d'interpolation et le problème d'élimination des points aberrants. Il s'agit de chercher **le** sous-ensemble de points candidats, nommé ensemble « super-candidat », pour lesquels l'interpolation spline satisfait les contraintes suivantes :

1. La forme de la spline interpolante est valide.
2. L'erreur quadratique médiane, entre la spline est les points super-candidats, est minimale.

La figure 3.15 met en évidence la première contrainte.

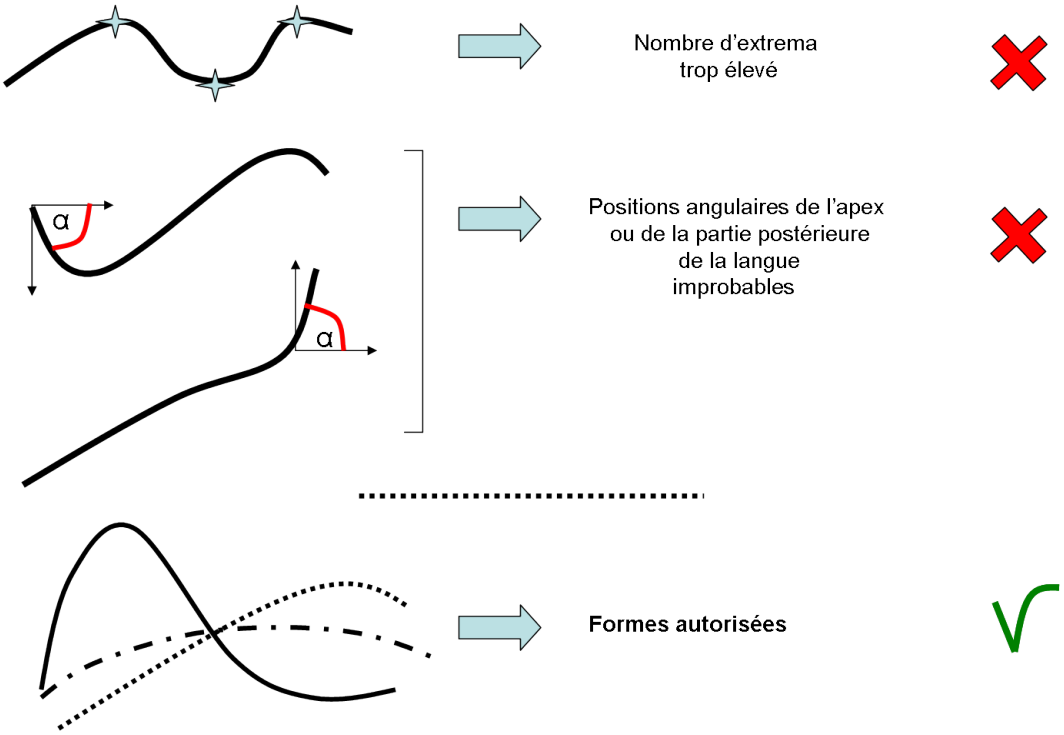


FIG. 3.15 – Formes interdites de la spline interpolante

Une forme interdite est un ensemble de contraintes sur :
– le signe de la dérivée au bornes de l'intervalle de définition de la spline
– le nombre d'extrema de la spline.

La distance $d$ entre une fonction spline $S$ et un ensemble $T$ de $M$ points « super-candidats »notés $[p_1, p_2, ..., p_M]$ de coordonnées $(u_i, v_i)$, est définie par :

$$d^2 = M_{1 \leq i \leq M}(d_i^2) = M_{1 \leq i \leq M}(|S(u_i) - v_i|) \tag{3.11}$$

avec $M$, la fonction médiane.

L'algorithme proposé est le suivant.

1 Effectuer un tirage aléatoire d'un ensemble $T$ de $M$ super-candidats, sous-ensemble de l'ensemble $E$ des $N$ points candidats issu de l'étape de filtrage grossier.
2 Calculer l'interpolation spline des points de $T$, on note $S$ la spline interpolante ainsi déterminée.
3 Si la forme de $S$ est interdite (cf. 3.15), retour à (1).
4 Stocker les couples $(d, S)$ dans l'ensemble $A$ ($d$ est la distance définie précédemment).
5 Répéter $[1, 2, 3, 4]$ $n$ fois.
6 – Si $A = \emptyset$ alors on diminue le nombre de super-candidats ($M = M - 1$), puis retour à (1).
– Sinon, la spline $S'$ de plus faible distance $d$ est choisie comme contour de la langue.

La méthode proposée est heuristique ; ni le temps de convergence, ni l'unicité de la solution ne sont assurés. Dans le cas où aucune spline valide n'est trouvée, le nombre de points super-candidats est diminué. La loi de diminution optimale est $M(t) = M - 1$ : elle garantit un nombre maximal de points utilisés pour l'interpolation spline. Cependant, l'utilisation de cette loi ralentit considérablement la méthode. Aussi, on pourra utiliser une loi du type $M(t) = M_0 \exp(-\frac{t}{\rho})$ avec $\rho$ un paramètre contrôlant la décroissance.

D'autre part, plus le nombre de points super-candidats diminue, (*i.e.* plus le nombre d'*outliers* estimés est important), plus il est préférable de tirer un nombre $n$ important d'ensemble $T$. La encore, il peut être intéressant de choisir une loi de la forme $n(t) = n_0 \exp(\frac{t}{\rho})$. Enfin,cet algorithme est inspiré de l'algorithme des **moindres carrés médians**, ce qui explique l'utilisation de l'erreur quadratique médiane et non moyenne. La valeur médiane permet d'atténuer l'influence perturbatrice des valeurs extrêmes des points aberrants.

Cette méthode fournit de bons résultats, y compris pour les trames « pathologiques ». Les figures 3.16 et 3.17 illustrent les résultats obtenus avec cet algorithme (après discrétisation et filtrage). Les points aberrants écartés sont représentés par des croix bleues. Les points retenus pour l'interpolation spline sont représentés par des astérisques verts.

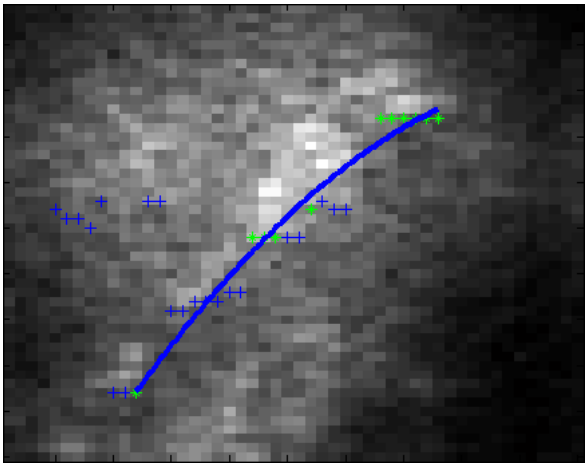

FIG. 3.16 – Interpolation du contour de la langue par une spline d'ordre 4 - Exemple 1

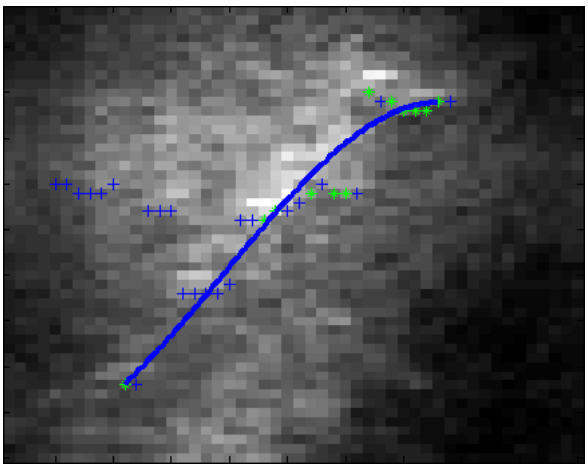

FIG. 3.17 – Interpolation du contour de la langue par une spline d'ordre 4 - Exemple 2

La description de l'image ultrasonore consiste donc en une modélisation du contour de la langue, à l'aide d'une spline d'ordre 4. L'information extraite de cette image est résumée par un intervalle de définition et les 4 coefficients de la spline soit 6 valeurs.

Toutefois, la visualisation des images ultrasonores de l'appareil vocal, lors de l'élocution, peut laisser penser que d'autres informations peuvent être prises en considération. En effet, le mouvement d'autres structures, comme celui de l'os hyoïde et ceux des structures situées sous la surface de la langue (muscle, graisse) peuvent participer à la production du son. Aussi, une description globale de l'image ultrasonore a été envisagée ; elle fait l'objet de la section suivante.

### 3.3.3 Approche *EigenTongues*

Nous proposons ici une description globale des images ultrasonores. Il ne s'agit plus de segmenter et de décrire certaines structures, mais de coder de façon globale et optimale chaque image. Ce type d'approche est notamment utilisé pour la reconnaissance de visages en imagerie optique. L'article de Turk ([Turk et Pentland, 1991]) construit un espace approprié à la description des visages, définit par des vecteurs qu'il appelle *EigenFaces*. Nous cherchons à étendre ce concept aux images de l'appareil vocal, et utilisons la dénomination *EigenTongues*.

Une image $I(x, y)$ peut être représentée par une matrice $N$ par $N$ [7], ou par un vecteur de taille $N^2$. Dans notre cas, le déploiement de l'image ultrasonore par discrétisation sur une grille polaire fournit un vecteur de taille $50 \times 50 = 2500$. Ce vecteur peut être perçu comme un point dans un espace de dimension 2500. Un ensemble de $K$ images peut donc être décrit comme un nuage de $K$ points dans cet espace de grande dimensionnalité.

D'autre part, les images ultrasonores présentent de nombreuses similarités entre elles. Cette propriété est accentuée par le filtrage. Ce dernier permet d'homogénéiser les zones se comportant de façon complètement aléatoire d'une image à l'autre. Aussi, l'ensemble des images est visualisable dans cet espace de dimension 2500 sous la forme d'un nuage relativement compact. L'idée de l'approche *EigenTongue* est de trouver un espace de plus faible dimensionnalité qui décrive correctement ces images. La recherche de cette espace utilise l'**analyse en composante principale** (ACP), dont nous décrivons maintenant la mise en œuvre pratique.
Soit $E$ un ensemble d'apprentissage de $M$ images de taille $N \times N$. Soit $A$ la matrice de taille $N^2$ par $M$ dont chaque colonne est une image de $E$. On définit la matrice de covariance $C$ de $A$ par :

$$C = \frac{1}{M}(AA^T) \tag{3.12}$$

La décomposition en valeurs singulières de cette matrice de taille $N^2$ par $N^2$ fournit $N^2$ valeurs singulières et $N^2$ vecteurs singuliers. Ces derniers forment une base d'un nouvel espace de dimension $N^2$ qui **explique de façon optimale la variance observée** dans les images de l'ensemble d'apprentissage. Ce nouvel espace est appelé *TongueSpace*[8]. Les vecteurs singuliers de taille $N^2$ peuvent être visualisés sous la forme d'images de taille $N \times N$. Ces vecteurs sont appelés *EigenTongues*. Le premier vecteur singulier est orienté selon la direction de plus grande variance observée sur l'ensemble d'apprentissage. Le second est contraint de se situer dans le sous-espace perpendiculaire au premier. A l'intérieur de ce sous espace, il est orienté dans la direction de la variance observée maximum. Le troisième se situe dans le sous-espace perpendiculaire aux deux premiers, il est toujours orienté vers la variance maximum, etc... La figure 3.18 illustre ces trois premières composantes principales ou *EigenTongues*. L'ACP est effectuée sur 3000 images tirées aléatoirement parmi les 72473 images disponibles dans la base de données **Ouisper**.

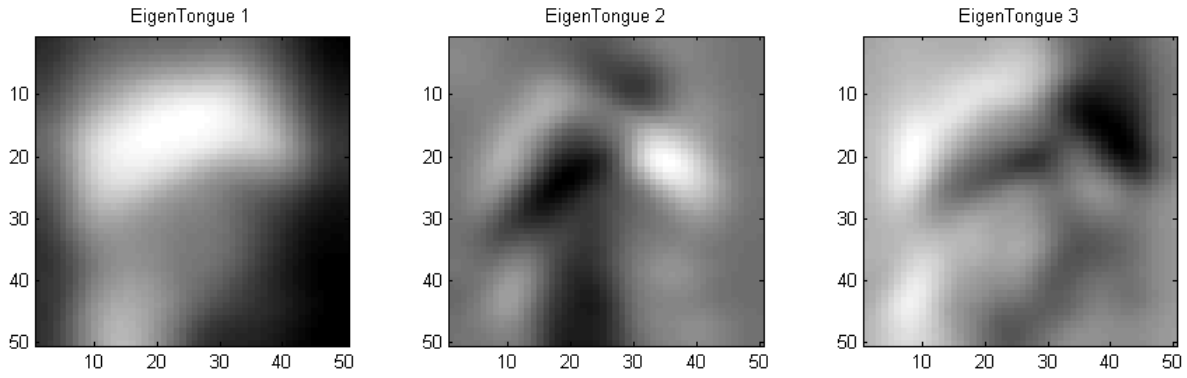


FIG. 3.18 – EigenTongues - ACP effectuée sur 3000 images réduites et filtrées

Le *TongueSpace* peut être perçu comme l'espace des images possibles. Il peut être utilisé pour coder de nouvelles images, en utilisant simplement leurs coordonnées dans cet espace. Une image, n'ayant pas servi à la construction du *TongueSpace*, peut ainsi être décrite comme une combinaison linéaire des *EigenTongues*. De plus, en raison de la similitude générale des nouvelles images avec les images de la base d'apprentissage, il est raisonnable de penser qu'un nombre restreint d'*EigenTongue* peut suffire à décrire correctement ces nouvelles images. Cette supposition est un point clé de la méthode et nous nous proposons de la vérifier par l'expérience. Les figures 3.19 à 3.21 montrent une reconstruction d'une image filtrée à partir des 5,20 et 40 premières *EigenTongues* sur les 2500 existantes. Cette image de test n'a pas été utilisée pour l'ACP initiale.

A la vue des résultats obtenus, il semble être possible de décrire une image à partir de 20 de ces coordonnées dans le *TongueSpace*. Le codage mise en œuvre porte sur l'intégralité de l'image. Il s'agit bien d'une approche globale.

La section suivante est consacrée au traitement des images optiques.

## 3.4 Traitement des images optiques

Les données acquises au VTVL comportent, en plus des images ultrasonores, une vue de profil de la tête du locuteur, acquise par un système d'imagerie optique traditionnel. Dans le cadre du projet *Ouisper*, nous

[7] la méthode est décrite pour une trame carrée, la généralisation à une image rectangulaire est immédiate
[8] Par analogie avec *FaceSpace*

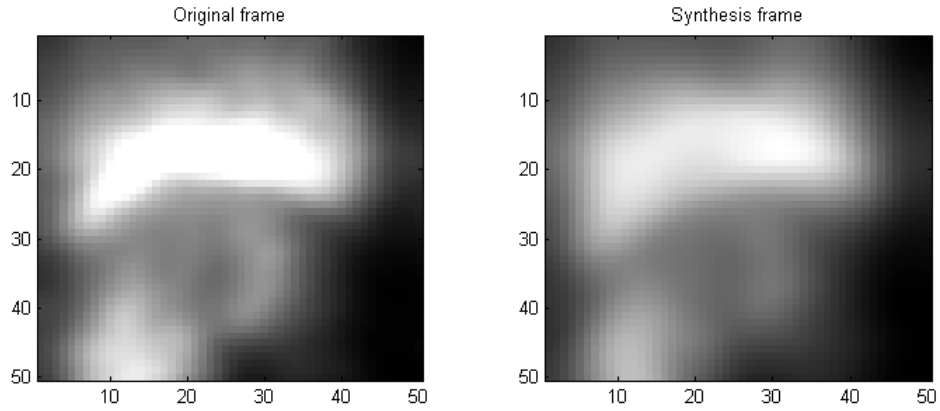


FIG. 3.19 – Projection d'une image de test sur le *TongueSpace* - Reconstruction à partir des 5 premières *EigenTongues*

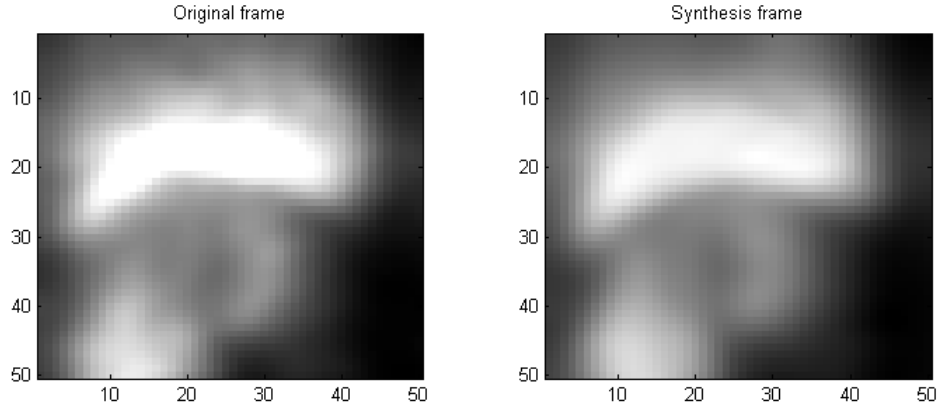


FIG. 3.20 – Projection d'une image de test sur le *TongueSpace* - Reconstruction à partir des 20 premières *EigenTongues*

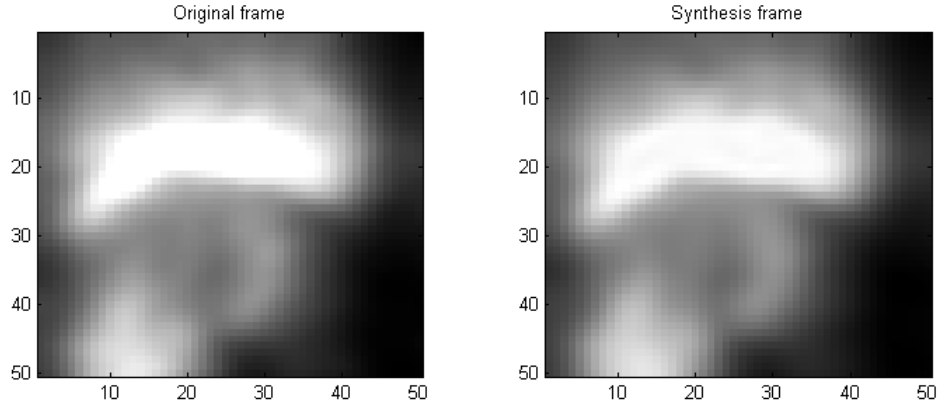


FIG. 3.21 – Projection d'une image de test sur le *TongueSpace* - Reconstruction à partir des 40 premières *EigenTongues*

nous intéressons à la description du mouvement des lèvres. **L'objectif est de positionner les lèvres et la commissure dans le plan image.**

### 3.4.1 Pré-traitement

Le système d'acquisition *HATS* maintient la tête du locuteur lors de l'élocution. Aussi, la région d'intérêt est identique pour toutes les trames vidéo. L'image couleur est convertie en niveaux de gris. Derrière le visage du locuteur, se situe un fond uniforme, ce qui facilite un seuillage de l'image. Enfin, le contour des lèvres est

extrait simplement par une méthode de *Sobel*. Cette chaîne de traitement est illustrée par la figure 3.22. Dans

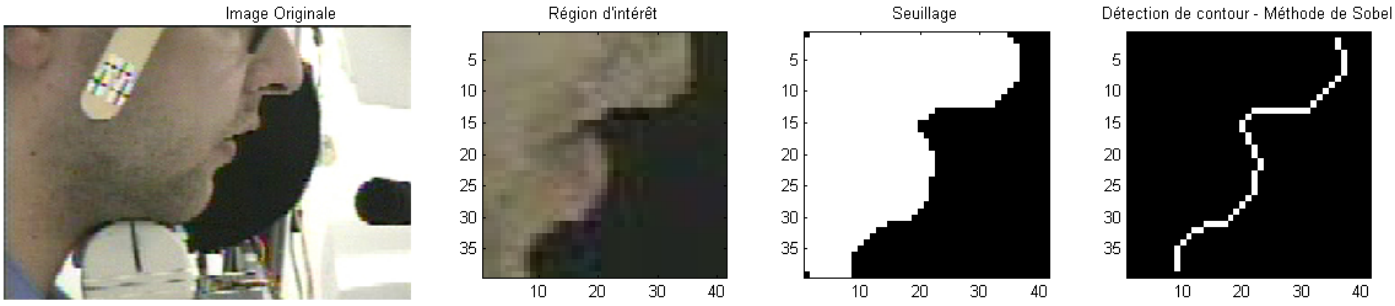


FIG. 3.22 – Vue de profil du visage - Prétraitement

[Denby *et al.*, 2006], la description du contour des lèvres s'effectue en considérant le contour des lèvres comme une fonction. Les positions des lèvres (inférieure et supérieure) et de la commissure sont obtenues en cherchant les extrema de cette fonction. Cependant, cette technique ne peut pas s'appliquer au locuteur de la nouvelle base de données *Ouisper*. En effet, le profil de ses lèvres peut parfois présenter **un point de rebroussement**. La figure 3.23 illustre ce phénomène.

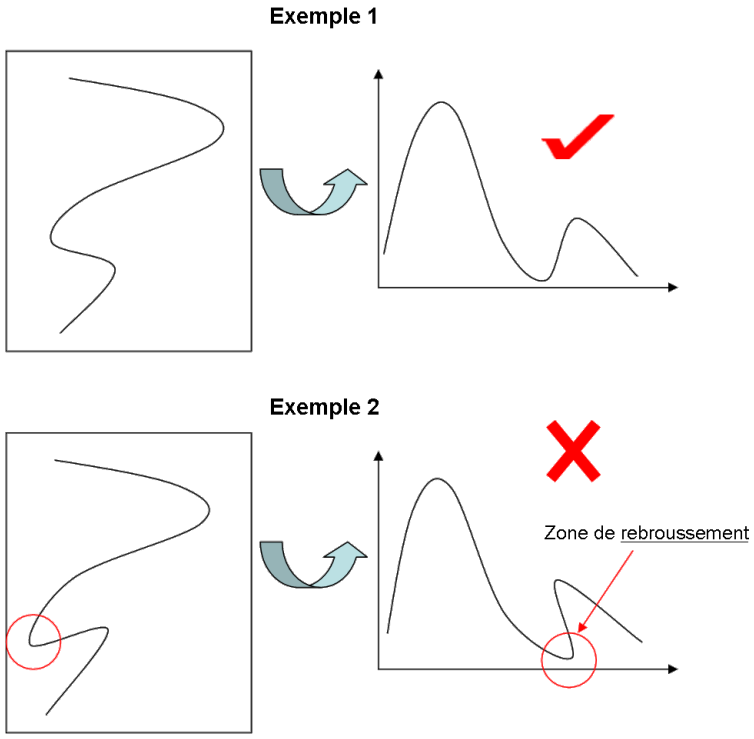


FIG. 3.23 – Interprétation du contour des lèvres comme une fonction

Pour résoudre ce problème, une approche applicable aux différents profils de lèvres a dû être envisagée. Elle fait l'objet de la section suivante.

### 3.4.2 Description du profil des lèvres

Dans [Feldman et Singh, 2005], Feldman reprend une idée de Attneave [Attneave, 1954], selon laquelle la quantité d'information maximale d'un contour, se situe dans les zones de plus forte courbure. Feldman fait référence à la notion de *Turning angle* [9], noté $\alpha$. Le *Turning Angle* est défini par l'angle entre les deux tangentes $\phi$ $\phi + \Delta\phi$ prises en deux points consécutifs d'une courbe $L$ (cf figure 3.24.

Pour une courbe uniformément discrétisée, Feldman démontre que la courbure est environ égale à la valeur du *turning angle*.

D'autre part, il est raisonnable de penser que les lèvres et la commissure sont des lieux de forte courbure. Aussi, le calcul du *Turning Angle*, va nous permettre de les localiser indépendamment de la forme de leur

[9] *Turning angle* peut être traduit par angle de braquage. Nous utiliserons cependant le terme anglo-saxon

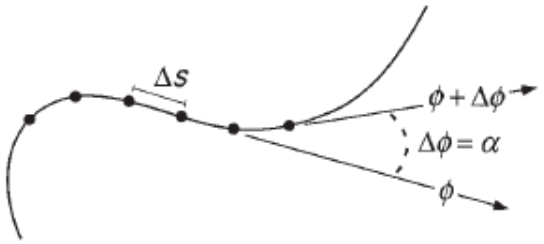


FIG. 3.24 – *Turning angle*

contour. Cette méthode a été implémentée ; la figure 3.25 illustre les résultats obtenus. Dans notre cas, il faut

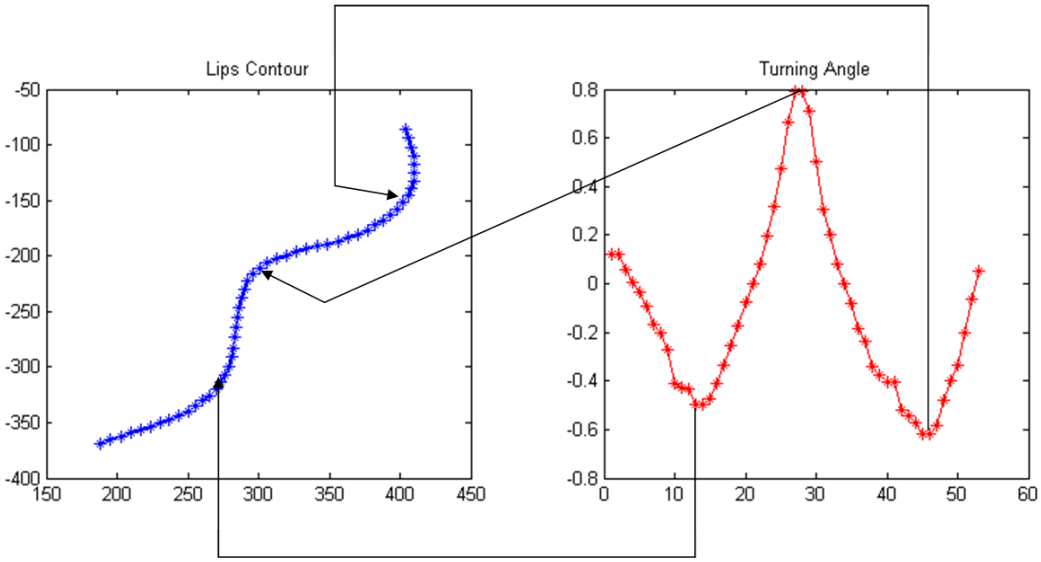


FIG. 3.25 – Localisation des lèvres et de la commissure grâce au *Turning angle*

remarquer que la méthode de calcul du *Turning angle* n'est pas rigoureusement identique à celle décrite dans [Feldman et Singh, 2005]. En effet, les points du contour ne sont pas tous uniformément espacés. Chaque point est porté par un pixel, donc deux points voisins (en 8-connexité), sont espacés de 1 ou $\sqrt{2}$ unités. Une méthode d'interpolation du contour visant à rétablir la 4-connexité devrait être envisagée. Cependant, la qualité des résultats obtenus semble suffisante. De plus, cette méthode fournit non seulement la position des lèvres et de la commissure mais également une information sur leur forme. Cette méthode est donc utilisée pour décrire le profil des lèvres du locuteur lors de l'élocution.

## 3.5 Conclusion

Les méthodes mises en œuvre dans cette étude visent à extraire l'information des images ultrasonores et optiques. L'objectif des différents traitements (discrétisation, filtrage, extraction de contour) est de faciliter l'extraction de descripteurs pertinents pour modéliser la configuration de la langue et des lèvres à chaque instant.

Toutefois, cette extraction de descripteurs s'effectue sans prise en compte de la dynamique temporelle du phénomène. A ce stade de l'étude, aucune approche visant à estimer le mouvement de la langue et des lèvres (Flux optique, *Block Matching*, ...) n'a été mise en œuvre. Ce type de méthode pourra être envisagée par la suite.

L'approche multimodale du processus de production de la parole du projet *Ouisper* nécessite également une description du signal audio. Cette dernière fait l'objet du chapitre suivant.

# Chapitre 4

# Analyse et description du signal de parole

## 4.1 Introduction

L'analyse de la parole peut être envisagée de deux manières. La première consiste à étudier l'évolution du contenu spectral du signal. Ce type d'approche nécessite d'extraire les caractéristiques fréquentielles du signal à intervalle régulier. Nous nommerons cette approche « description acoustique ». Le second type d'analyse, nommée « description segmentale », consiste à découper le signal en segments de taille variable. Ces segments peuvent avoir un sens phonétique. On cherchera par exemple à délimiter les unités linguistiques comme les phones, les diphones etc ... Cependant ce type de segmentation nécessite la connaissance a priori du texte à l'origine du signal. Des méthodes de segmentation automatique ont été développées afin de résoudre ce problème ; c'est le cas du le système ALISP [1].

Les travaux préliminaires au projet Ouisper ([Denby *et al.*, 2006]) utilisent la description acoustique du signal de parole, plus précisément la méthode d'analyse-synthèse LPC[2]. Cette approche est reprise dans le cadre de cette étude. Le codage MFCC[3] a également été expérimenté ; il sera décrit dans ce chapitre. Une des principale perspectives du projet *Ouisper* est l'utilisation de la segmentation ALISP, méthode qui sera brièvement explicitée.

## 4.2 Description acoustique du signal de parole

### 4.2.1 Contraintes imposées par l'échantillonnage vidéo

. L'estimation du contenu spectral du signal de parole nécessite le fenêtrage préalable de ce dernier. Dans le cadre d'une modélisation visio-acoustique, ce fenêtrage doit être **synchrone** avec la vidéo. La figure 4.1 illustre cette contrainte.

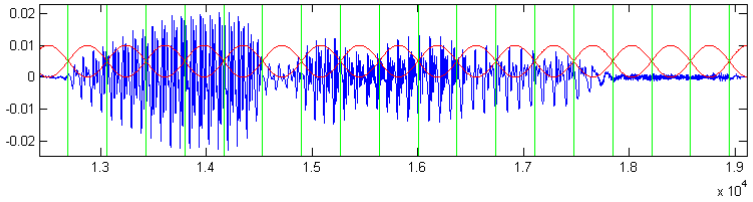


FIG. 4.1 – Fenêtrage du signal audio - Synchronisme audio-visuel - Analyse grossière

Les limites des trames vidéo figurent en vert, le signal de parole en bleu, son fenêtrage en rouge. La vidéo est cadencée à 29.97 images par seconde, soit une image toutes les 33 ms environ. Évaluer le contenu spectral du signal de parole toutes les 33 ms revient à utiliser une fenêtre de 66 ms décalée de 33 ms (C'est le cas à la figure 4.1). Dans notre exemple, une fenêtre de Hamming est utilisée. Ce fenêtrage permet de mettre en relation directe une image du conduit vocal et le spectre du signal audio correspondant. Cependant, le fenêtrage doit respecter les contraintes de stationnarité du signal de parole. Si une voyelle est considérée comme stable sur 100

[1] Système d'analyse-synthèse de la parole développé au Laboratoire Traitement et Communication de L'information de Télécom Paris

[2] LPC signifie *Linear Predictive Coding*

[3] MFCC signifie *Mel Frequency Cepstral Coefficient*

ms environ, une plosive ne l'est que sur 10 ms. Aussi, un fenêtrage plus fin est réalisé ; il est représenté sur la figure 4.2.

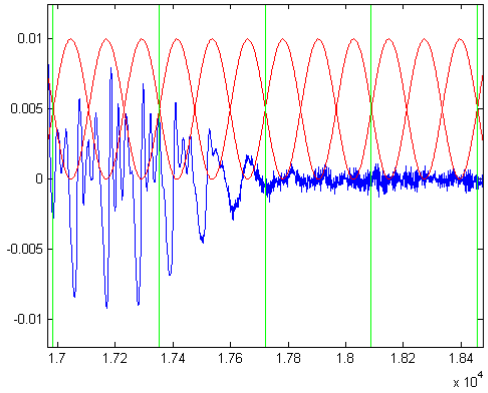


FIG. 4.2 – Fenêtrage du signal audio - Synchronisme audio-vidéo - Analyse fine

Dans ce cas, une image est mise en relation avec 3 trames audio de 22 ms décalées de 11 ms. Ce fenêtrage plus fin permet une meilleure description du signal de parole.

Cependant, ce fenêtrage fin est difficile à mettre en œuvre dans le cadre de la modélisation visio-acoustique. Dans ce document, il ne sera pas utilisé pour cette dernière.

### 4.2.2 Analyse-synthèse LPC

**Considérations physiques** La parole peut être considérée comme le résultat du filtrage d'un signal source, par le conduit vocal. Dans le cas d'un son voisé, le signal source est un train d'impulsions à la fréquence de vibration de la glotte. Cette fréquence est appelée fréquence fondamentale ou *pitch*[4]. Dans celui d'un son non-voisé, l'air turbulent en sortie du larynx peut être modélisé par un bruit blanc gaussien. Le conduit vocal peut être modélisé par une juxtaposition de plusieurs cylindres de même longueur mais de diamètre variable ([Rabiner et Juang, 1993],[G.Fant, 1970]). Ce modèle physique peut être assimilé à un **filtre auto-régressif**. Sa fonction de transfert peut s'écrire :

$$\forall z \in C, H(z) = \frac{1}{a_0 + a_1 z^{-1} + ... + a_N z^{-N}} \tag{4.1}$$

en notant $C$ l'ensemble des nombres complexes.

Les pôles de $H(z)$ sont à l'origine d'un pic dans le module de la réponse fréquentielle $|H(\exp i\omega)|$. Ces pics modélisent les fréquences de résonance du conduit vocal ; ils sont appelés **formants**. La position des formants caractérisent en grande partie la parole produite. En s'appuyant sur cette modélisation, Atal et Hanauer [Atal et Hanauer, 1971], ont introduit la technique d'analyse-synthèse LPC. Ils proposent une manière d'estimer les coefficients du filtre auto-régressif modélisant le conduit vocal.

**Étude du voisement et estimation de la fréquence fondamentale** La première étape de l'analyse est l'estimation du voisement et de la fréquence fondamentale dans les trames voisées. Cette étape est commune à un grand nombre de méthodes d'analyse. Elle est considérée comme un point délicat de la description du signal de parole et constitue encore aujourd'hui un axe de recherche à part entière. Aussi, seul son principe général sera décrit dans ce document. Le lecteur pourra par exemple consulter [Doval, 1994] pour une synthèse approfondie des méthodes disponibles.

Nous décrivons maintenant une méthode simple d'estimation de la fréquence fondamentale fondée sur une approche temporelle. Soit $x[n]$, le signal de parole compris dans une fenêtre de longueur $N$. La fonction d'auto-corrélation de ce signal s'écrit :

$$\phi(n) = \sum_{i=1}^{N} x(i)x(n+i) \tag{4.2}$$

Cette fonction est maximale pur les fréquences harmoniques de la fréquence fondamentale. Estimer la fréquence fondamentale revient à rechercher les maxima de cette fonction, comme le montre la figure 4.3. A droite

[4]Dans ce document, la notion de *pitch* ne sera pas différenciée de la notion de fréquence fondamentale

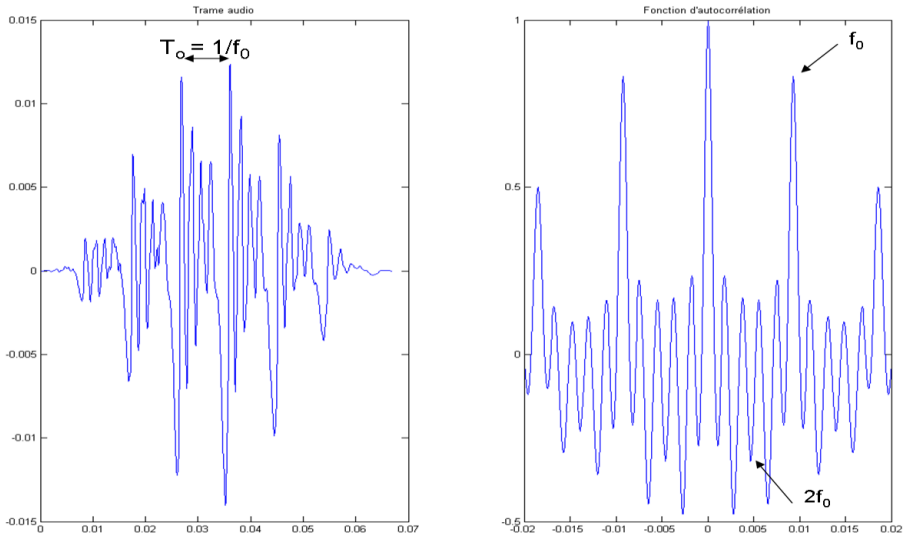


FIG. 4.3 – Méthode temporelle pour l'estimation de la fréquence de parole

figure une trame d'un son voisé, multipliée par une fenêtre de Hamming. Une estimation manuelle de la période fondamentale est effectuée sur la forme d'onde, en évaluant la longueur moyenne entre deux maxima [5] ; elle fournit $T_0 = 0.0093$ secondes. A gauche est représentée la fonction d'autocorrélation : le maximum de la fonction, mise à part l'origine, est atteint au temps 0.0093 seconde, soit une fréquence fondamentale $f_0$ de 107 Hz. Toutefois, cette technique ne permet pas, à elle seule, de différencier la période fondamentale de ses harmoniques. En effet, l'amplitude de la fonction d'autocorrélation à certaines fréquences harmoniques de la fréquence fondamentale peut être supérieure à celle observée à la fréquence fondamentale elle-même. Plusieurs méthodes sont proposées afin de s'assurer de détecter la fréquence fondamentale et non un de ses sous-multiples. On pourra notamment privilégier l'extremum de la fonction d'autocorrélation correspondant à la période fondamentale la plus petite, comme le propose [Cadic, 2003].

La fonction d'autocorrélation peut également servir à déterminer le caractère voisé/non-voisé d'une trame de parole. Nous notons $\alpha$ la valeur de la fonction d'autocorrélation à la fréquence fondamentale estimée. En théorie, si le signal est parfaitement harmonique, alors $\alpha = 1$. En revanche, pour une trame (de longueur infinie) de bruit blanc, $\alpha = 0$. Aussi, un seuillage des valeurs de $\alpha$ peut permettre de décider du caractère voisé/non-voisé du signal de parole. La figure 4.4 illustre une trame de signal non-voisé.

Mise à part l'origine, le maximum de la fonction d'autocorrélation est 0.29. Dans le cas de la figure 4.3, le maximum vaut 0.74. Pour décider du voisement, un seuil à 0.5 pourrait ainsi être fixé.

L'approche proposée est très simple. D'autres techniques, bien plus élaborées, comme l'algorithme YIN [6] [de Cheveigné et Kawahara, 2002], fournit une estimation beaucoup plus fiable de la fréquence fondamentale et du voisement.

Cependant, la fréquence fondamentale n'est pas un descripteur utilisable dans le cadre du projet **Ouisper**. Rappelons que ce dispositif est destiné à une communication silencieuse, prohibant toute activité glottale. Toutefois, son utilisation lors de la synthèse LPC, que nous décrivons maintenant, facilite l'évaluation de la qualité du signal artificiel produit grâce à la modélisation visio-acoustique. C'est pourquoi son estimation reste importante.

**Estimation des coefficients du filtre AR modélisant le conduit vocal** Il s'agit maintenant d'estimer les coefficients du filtre auto-régressif introduit précédemment. Ces derniers minimisent l'erreur quadratique $e$ telle que :

$$e = \sum_{-\infty}^{\infty} |x(n) - y(n)|^2 \tag{4.3}$$

avec $y(n) = \sum_{i=1}^{n} a_i x(n-i)$.
Atal et Hanauer [Atal et Hanauer, 1971], montrent que ces coefficients, notés $(a_k)_{1 \leq k \leq N}$ (si $N$ est l'ordre du

[5] Cette méthode n'est certes pas rigoureuse, elle permet cependant d'expliciter simplement cette première approche de l'estimation de la fréquence fondamentale

[6] L'algorithme YIN, développé à l'IRCAM, est *Open-Source*, sous le *Copyright* CNRS/IRCAM, 2002

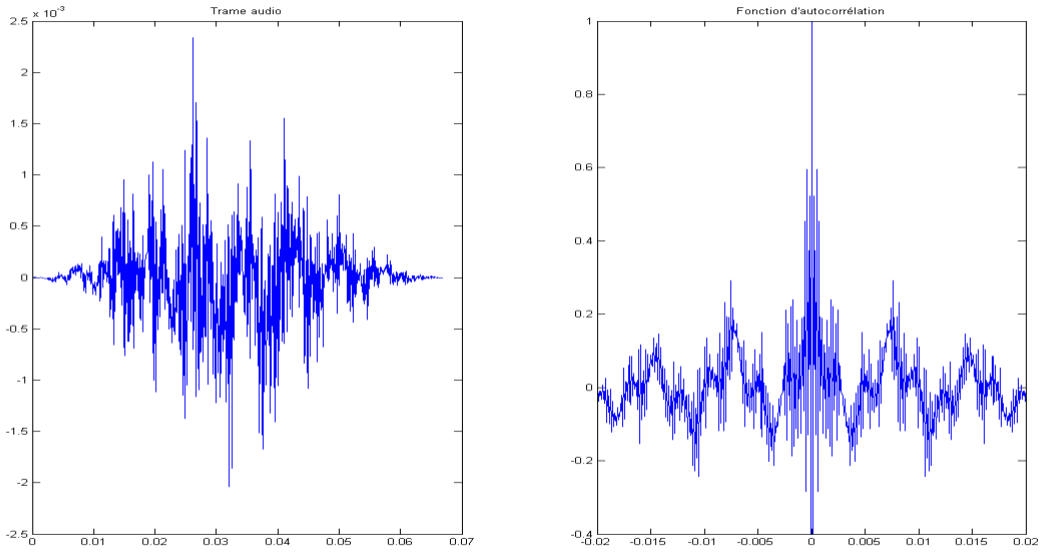


FIG. 4.4 – Fonction d'autocorrélation d'une trame non-voisée

filtre) sont solution du système suivant :

$$\begin{pmatrix} r_x[0] & r_x[1] & ... & r_x[N-1] \\ r_x[-1] & r_x[0] & ... & r_x[N-2] \\ ... & ... & ... & ... \\ r_x[-N+1] & r_x[-N+2] & ... & r_x[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ ... \\ a_N \end{pmatrix} = \begin{pmatrix} r_x[1] \\ r_x[2] \\ ... \\ r_x[N] \end{pmatrix} \tag{4.4}$$

où $r_x[k] = \sum_{n=-infty}^{\infty} x[n-k]x[n]$ est la fonction d'autocorrélation empirique du signal $x[n]$. L'algorithme de **Levinson-Durbin** permet la résolution de ce système. Les détails de cet algorithme ne seront pas exposés ici.

Après avoir estimé le voisement, la fréquence fondamentale (sur les trames voisées) et les coefficients du filtre AR, nous pouvons alors **re-synthétiser** le signal de parole. Pour une trame voisée, le signal source est un train d'impulsions à la fréquence du *pitch* ; dans le cas contraire, il s'agit d'un bruit blanc gaussien. La figure 4.5, résume les étapes d'analyse, et la synthèse de la parole par LPC .

L'ordre d'analyse LPC doit être choisi en fonction de la fréquence d'échantillonnage. D'après [Dutoit, 2003], pour un signal échantillonné à 11025 Hz, l'ordre du filtre AR doit être comprise entre 10 et 20. En vue de la modélisation visio-acoustique, nous cherchons à limiter le nombre de descripteurs du signal de parole. Aussi, nous optons pour un ordre 12 et une fréquence d'échantillonnage de 11025 Hz. La taille des trames d'analyse (filtre AR et *pitch*) est de 22 ms, elles sont décalées toutes les 11 ms.

Ainsi, le signal de parole est décrit par les coefficients d'un filtre modélisant le conduit vocal. Cependant, les coefficients LPC ne sont pas considérés comme des descripteurs fiables pour le codage du signal de parole. On leur reproche leur large dynamique et la possible instabilité du filtre dont l'un des coefficients viendrait à être faiblement modifié. Cependant, il existe d'autres représentations des coefficients LPC qui permettent d'atténuer cet effet. Dans [Denby *et al.*, 2006] et [Denby et Stone, 2004], les *Line Spectrum Frequencies* (LSF) sont utilisées ; elles font l'objet de la section suivante.

### 4.2.3 Représentation des coefficients LPC à l'aide des LSF

#### Principe des Line Spectrum Frequencies - LSF

La représentation des coefficients LPC à l'aide des LSF ne repose plus sur une modélisation physique du conduit vocal. Elle peut à ce titre paraître quelque peu artificielle. La présentation suivante des LSF s'appuie sur l'article de Fang Zheng [Zheng *et al.*, 1998]. L'auteur montre que la fonction de transfert du conduit vocal $H(z)$ décrite à la section 4.2.2, peut s'écrire :

$$h(z) = \frac{P(z) + Q(z)}{2} \tag{4.5}$$

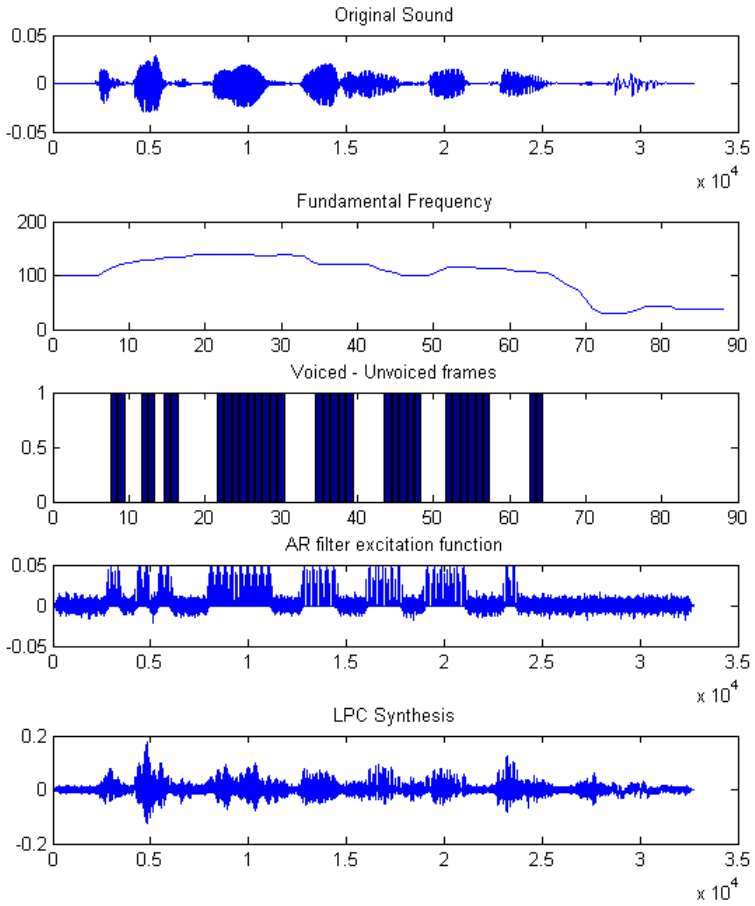


FIG. 4.5 – Analyse- Synthèse LPC : Résultats des différents traitements

avec

$$\begin{cases} P(z) = 1 + \sum_{p=1}^{N}(a_p + a_{N+1-p})z^{-p} + z^{-(N+1)} \\ Q(z) = 1 + \sum_{p=1}^{N}(a_p - a_{N+1-p})z^{-p} + z^{-(N+1)} \end{cases} \tag{4.6}$$

Les polynômes $P$ et $Q$ présentent les propriétés intéressantes suivantes :

- Tous les zéros de $P$ et de $Q$ sont sur le cercle unités. Ils peuvent donc s'écrire $z_i = e^{j\omega_i}$ et chaque $\omega_i$ est appelé *line spectrum frequency* ou paire de lignes en fréquence.
- Les zéros de $P$ sont entrelacés avec ceux de $Q$.
- $(-1)$ est zéro de $P$ et $(1)$ est zéro de $Q$.
- En plus de 1 (resp. $-1$), $P$ (resp $Q$) possède $N/2$ paires de zéros conjugués (si $z_i$ est zéros de P ou de Q, alors $\bar{z}_i$, le complexe conjugué de $z_i$, l'est aussi).

En notant $z_i = e^{j\omega_i}$ les zéros de $P$ et $t_i = e^{j\theta_i}$, les zéros de $Q$, on montre que

$$0 \leq \omega_1 \leq \theta_1 \leq \omega_2 \leq \theta_2 \leq ... \leq \omega_{\frac{N}{2}} \leq \theta_{\frac{N}{2}} \leq \pi \tag{4.7}$$

Après cette transformation, les coefficients du filtre autorégressif de l'approche LPC sont équivalents aux LSF $\omega_i$ et $\theta_i$, considérés comme plus stables. Cette transformation est réversible, et la conversion LSF vers LPC est sans perte. S'il est vrai que cette méthode s'appuie principalement sur un jeu d'écriture mathématique, les coefficients LSF peuvent néanmoins être reliés "visuellement" à la réponse du filtre autorégressif de la modélisation LPC. Dans le domaine fréquentiel, la fonction de transfert de ce filtre peut s'écrire :

$$H(\exp j\omega) = \frac{2}{P(\exp j\omega) + Q(\exp j\omega)} \tag{4.8}$$

De plus, les LSF sont les arguments des zéros de $P$ et de $Q$. Aussi, si une paire de LSF $(\omega_i, \theta_i)$ est proche d'une fréquence $\omega_0$, alors $|P(\exp(j\omega_0)) + Q(\exp(j\omega_0))|$ devient petit et $|H(\exp(j\omega_0))|$ devient grand. On observe alors un pic d'amplitude sur la réponse fréquentielle. La figure 4.6 illustre cette propriété, les cercles représentant les LSF.

Chaque fenêtre du signal de parole peut donc être décrite par un vecteur de LSF.

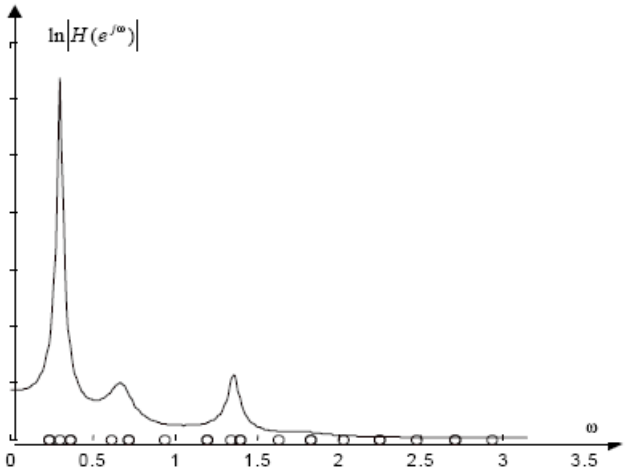


FIG. 4.6 – Relation entre les LSF et la fonction de transfert du filtre auto-régressif introduit dans la modélisation LPC

### Utilisation des LSF pour la classification des trames de parole

Nous proposons maintenant de montrer par l'expérience que cette représentation permet également de classifier les trames de parole en trois catégories :

- Les trames voisées
- Les trames non-voisées
- Les silences

Pour chaque trame audio analysée, nous traçons sur un même graphe l'amplitude des différents LSF (soit 12 coefficients pour une analyse LPC d'ordre 12). Trois classes de trajectoires se dégagent ; elles sont illustrées sur la figure 4.7. L'écoute des trames audio correspondantes, permet d'étiqueter ces classes. Il s'agit bien des trames

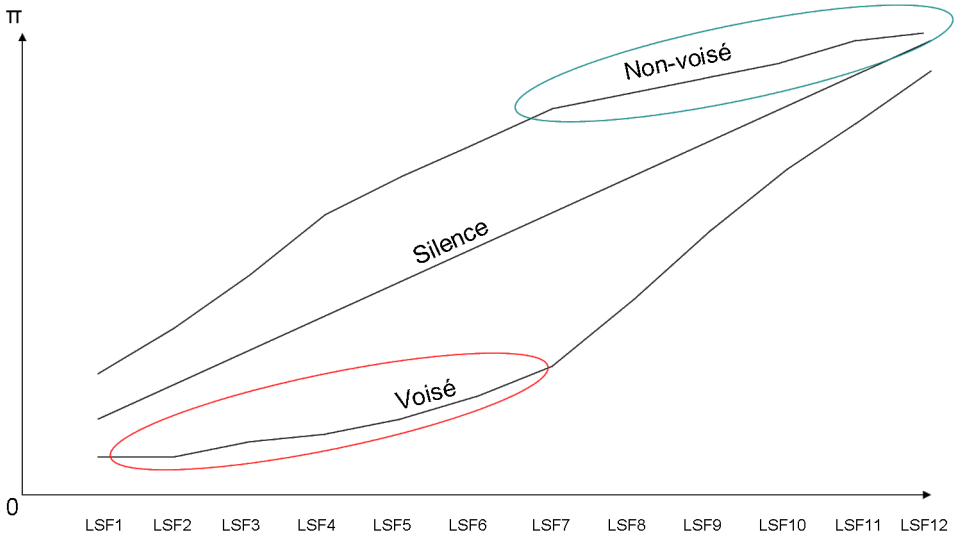


FIG. 4.7 – Amplitude des LSF - Classification du signal de parole

voisées, non-voisées et des zones de silence. La discussion suivante tente d'expliquer cette propriété.

Le spectre d'une trame de silence est quasiment plat, donc les LSF sont uniformément espacés : la trajectoire des LSF est bien une droite. Le spectre des trames voisées présente des pics dans les basses fréquences, le **formants**, dont la position caractérise notamment les voyelles. Dans cet intervalle, les premières LSF sont proches (zone rouge). En hautes fréquences, le spectre est quasiment plat, les LSF étant uniformément espacées, ce qui est cohérent avec la trajectoire observée. Enfin, le spectre d'une trame non voisée, par exemple celui d'une fricative, est riche en hautes fréquences. Les LSF se rapprochent donc dans cette bande (zone bleue), ce qui apparaît dans cette représentation. La position relative des trajectoires étiquetées « voisées »et « non-voisées », de part et d'autre de la diagonale de silence, peut également s'expliquer.

Le premier formant est généralement à basse fréquence ; il est modélisé par une agglomération de LSF de très faible amplitude. La modélisation des autres formants contraint un maximum de LSF à rester dans les basses fréquences. La pente observée est donc faible pour les premiers LSF, la valeur des autres LSF est contrainte à rester relativement faible. Dans le cas des trames non-voisées, le phénomène est inversé. Le spectre est riche en hautes fréquences, un maximum de LSF, de forte amplitude, décrivent cet intervalle. Les premiers LSF sont contraints à des valeurs plus élevées. La relation d'ordre entre les LSF impose la forme de ces trajectoires.

Ainsi, cette représentation permet de classer de manière simple les trames du signal de parole à analyser. D'autres représentations du signal de parole existent, dont une très utilisée, la représentation MFCC, qui est décrite dans la section suivante.

### 4.2.4 Analyse MFCC

En l'absence d'activité glottale, il peut être intéressant de décrire le signal de parole de manière totalement indépendante des variations prosodiques du locuteur. La représentation MFCC [Davis et Mermelstein, 1980] tente de séparer les contributions respectives de l'excitation et du conduit vocal.

Comme il a été décrit précédemment, la production de la parole peut être assimilée à un système « source - filtre ». Le signal $s(n)$ vocal peut donc s'écrire dans le domaine temporel :

$$s(n) = e(n) * h(n) \tag{4.9}$$

avec e(n) la réponse impulsionnelle du signal d'excitation, h(n), celle du filtre modélisant le conduit vocal et $(*)$, le produit de convolution.

Dans le domaine de Fourier, cette relation devient :

$$S(\omega) = E(\omega)H(\omega) \tag{4.10}$$

En partant de l'idée selon laquelle la contribution à la perception des sons de la parole des hautes fréquences est plus faible que celle des basses fréquences, on interprète $S(\omega)$ sur une échelle fréquentielle particulière dite de Mel [Laprie, 2002] définie par :

$$M = \frac{1000}{\log_2} * \log(1 + \frac{f}{1000}) \tag{4.11}$$

où $f$ est la fréquence en Hz.

Cette échelle est connue pour rendre compte de la perception humaine. Elle est linéaire en basses fréquences et logarithmique en hautes fréquences. Elle définit le banc de filtres triangulaires représenté à la figure 4.8. L'échelle en fréquence entre 0 et $\frac{f_e}{2}$ est ainsi partitionnée en $N$ bandes sur l'échelle de Mel, si $f_e$ est la fréquence d'échantillonnage du signal audio (typiquement $N = 24$). Ensuite, le filtrage est effectué en multipliant le spectre

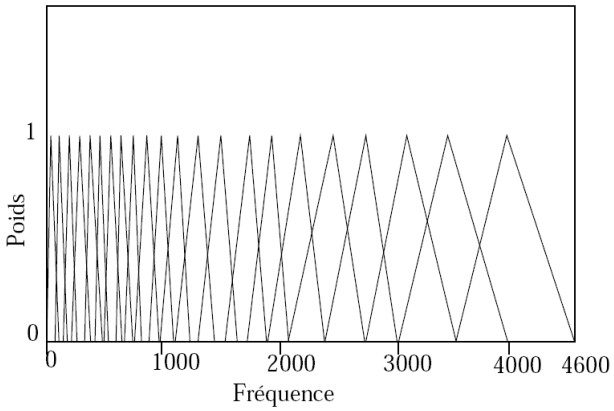


FIG. 4.8 – Banc de filtres sur l'échelle de Mel

du signal $S(n)$ (module de la FFT) par le gabarit des filtres. En notant $X_k$ le logarithme de l'énergie du signal en sortie de ces filtres triangulaires pour $k = 1, 2, ..., N$, on définit les coefficients cepstraux [7] Mel (MFCC) à l'aide de la transformée en cosinus discret :

$$MFCC_i = \sum_{k=1}^{N} X_k \cos(i(k - \frac{1}{2})\frac{\pi}{N}) \tag{4.12}$$

[7] Le cepstre est défini comme la Transformée de Fourier inverse du logarithme de la densité spectrale.

Les coefficients cepstraux permettent une déconvolution entre la source des sons produits et le conduit vocal, de plus ils ont la propriété d'être fortement décorrélés. Les premiers coefficients MFCC modélisent la contribution du conduit vocal [8].

La description MFCC est notamment très utilisée pour la reconnaissance de la parole. En effet, ce type de codage est indépendant de l'excitation, donc de la prosodie du locuteur.

Contrairement à l'analyse LPC, la re-synthèse du signal de parole, à partir des coefficients cepstraux n'est pas immédiate. En effet, certaines étapes de la décomposition du spectre ne sont pas réversibles :

- L'information sur la phase du signal est perdue suite à l'utilisation de l'amplitude du spectre pour le filtrage Mel.
- La quantification du spectre sur l'échelle de Mel engendre une perte d'information

L'inversion des MFCC, et le retour à l'amplitude du spectre est une opération difficile. De plus, comme nous l'aborderons plus loin, les coefficients MFCC n'ont pas été introduits dans l'objectif d'être utilisés directement pour la synthèse de la parole. Ils permettent, nous le verrons, d'utiliser le système d'analyse-synthèse ALISP, décrit dans la section suivante.

Cependant, une méthode d'inversion des coefficients MFCC est décrite dans [Chazan *et al.*, 2000].

Nous abordons maintenant une approche très différente de l'analyse de la parole : la description segmentale.

## 4.3 Description segmentale *ALISP*

L'approche segmentale classique consiste à rechercher les limites temporelles des unités linguistiques telles que les phones, les diphones, les mots ... Une segmentation linguistique est nécessaire. Cependant, cette étape indispensable est extrêmement difficile car une connaissance phonétique *a priori* est requise. De plus, étiqueter le signal audio est une approche conceptuellement difficile. En effet, la réalisation de deux phonèmes identiques peut donner lieu à des signaux de parole assez différents. La parole étant un phénomène essentiellement dynamique, la réalisation d'un phonème est très dépendante du contexte dans lequel il se trouve. Les phones environnants influent fortement sur la dynamique du conduit vocal. Les mouvements articulatoires peuvent être modifiés de façon à minimiser l'effort à produire pour les réaliser à partir d'une configuration articulatoire donnée, ou pour anticiper une configuration ultérieure. Ces effets sont connus sous le nom de réduction, d'assimilation et de co-articulation. Ces effets sont décrits dans [Dutoit, 2003]. La segmentation automatique ALISP [9], développée lors de la thèse de Jan Cernocky [Cernocky, 1998] vise à s'affranchir de ces difficultés. La recherche de segments et leur classification n'est plus contrainte par des considérations phonétiques ; seules interviennent les caractéristiques du signal.

La segmentation ALISP s'effectue en deux étapes [Bimbot *et al.*, 1988] :

- une segmentation initiale du signal du corpus de parole en utilisant une décomposition temporelle et une quantification vectorielle.
- une segmentation statistique en utilisant les modèles de segments initiaux.

La figure 4.9 illustre le principe de fonctionnement de la segmentation ALISP.

### 4.3.1 Segmentation initiale

Le signal de parole est fenêtré. Un filtrage est effectué sur chaque trame afin de pré-accentuer les aigus. Une estimation du voisement et de la fréquence fondamentale (en cas de voisement) et une extraction des coefficients MFCC sont effectuées. Une décomposition temporelle est ensuite appliquée sur la séquence des vecteurs MFCC. Cette méthode recherche, dans la séquence de MFCC, des zones de stabilité spectrale, ainsi que les instants de transitions qui les séparent. Cette étape est décrite dans [Bimbot *et al.*, 1988]. L'étape suivante est une quantification vectorielle des segments trouvés par décomposition temporelle. Une classification non-supervisée répartit les segments parmi 64 classes. Les unités ainsi constituées sont appelées unités ALISP. Cette segmentation initiale aboutit à un étiquetage du corpus de parole, qui est utilisé ensuite comme base d'apprentissage pour la segmentation statistique.

### 4.3.2 Segmentation statistique

Les unités ainsi constituées sont modélisées par des modèles de Markov cachés (HMM). Les détails de cette technique de modélisation ne seront pas décrits dans le cadre de ce document. En revanche, l'article historique de Rabiner [Rabiner, 1990] fournit une étude approfondie sur ce sujet. Replaçons néanmoins l'utilisation des HMM dans le contexte du traitement de la parole. Un modèle de Markov caché est un puissant outil de modélisation statistique, qui peut être considéré comme un ensemble discret d'états et de transitions entre ces états. Un tel modèle caché est caractérisé par une distribution de probabilités pour chaque état et par des probabilités de

[8] Typiquement, on considère que le conduit vocal peut être modélisé par les 12 premiers coefficients MFCC pour $N = 24$.

[9] ALISP : *Automatic Language Independent Speech Processing*

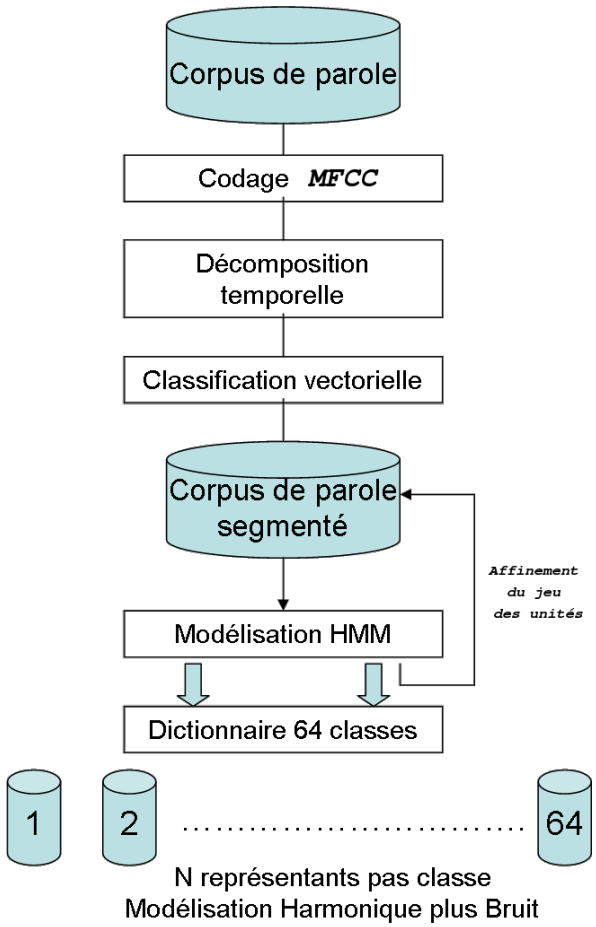


FIG. 4.9 – Principe de la segmentation ALISP

transition entre ces états. Afin de modéliser une unité acoustique, l'approche couramment utilisée, consiste à choisir un modèle de Markov caché à trois états par phonème. On suppose alors que le second état modélise la partie stationnaire du phonème, et les états extérieurs modélisent la co-articulation avec les phonèmes voisins. L'apprentissage d'un modèle de Markov caché consiste en l'estimation des probabilités de transition à partir de séquences d'états observées. Dans le cas de la modélisation d'unités ALISP, un raisonnement similaire peut être tenu.

La dernière étape est une succession de segmentation et de ré-estimation des paramètres des modèles HMM. Cette technique permet d'affiner la segmentation du corpus. Ainsi, une base de données a été constituée. Elle est organisée en 64 classes, dans lesquelles figurent un certain nombre d'unités présentant des caractéristiques statistiques proches. Ces unités sont appelées **représentants**. Le codage ALISP fournit donc :

- une base de représentants organisée en 64 classes.
- une modèle statistique HMM de ces 64 classes.

Le codage des représentants dans la base de données s'effectue à l'aide du modèle « Harmonique Plus Bruit »(la forme d'onde n'est pas stockée) décrit dans [I.Stylianou, 1996].

Le système ALISP peut être utilisé dans de multiples applications comme le codage à très bas débit, la reconnaissance de la parole, la vérification du locuteur et la transformation de la voix. Pour plus d'informations sur le système ALISP, le lecteur est invité à consulter [Padellini, 2006] et [Mosbah, 2005]. Une utilisation du système ALISP a été envisagée dans le cadre du projet *Ouisper*. Elle sera décrite dans le chapitre suivant.

## 4.4 Conclusion

Dans le cadre de la description acoustique, le signal de parole est décrit à intervalles réguliers par un vecteur de coefficients. L'analyse LPC, via sa représentation LSF, permet de modéliser la contribution du conduit vocal au processus de production de la parole. Dans le cadre d'une modélisation du lien visio-acoustique à partir d'images du conduit vocal, ce point est capital. L'utilisation d'une de ces représentations est donc cohérente avec l'objectif poursuivi dans cette étude. Cependant, sans information sur l'activité glottale, ce type de représentation ne permet pas une synthèse de qualité du signal de parole. En effet, une excitation du filtre AR

à l'aide d'un bruit blanc gaussien uniquement produit un signal proche de la voix chuchotée, mais difficilement intelligible. L'utilisation d'un système d'analyse segmentale peut éviter ce problème. La synthèse ALISP est basée sur l'utilisation de véritables segments de parole, à savoir les représentants. Le choix de la bonne séquence de représentants suffit à produire une parole de qualité.

Nous disposons donc maintenant d'une description complète des données de la base *Ouisper*. Chaque image, et chaque trame audio est décrite par une série de coefficients. Nous cherchons à présent à modéliser la relation qui existe entre ces différents descripteurs, cette étude est l'objet du dernier chapitre.

# Chapitre 5

# Modélisation visio-acoustique

## 5.1 Introduction

Dans les chapitres précédents, une description synchrone de l'appareil vocal et du signal de parole vocalisé, a été mise en œuvre. La modélisation visio-acoustique que nous présentons maintenant s'appuie sur l'hypothèse suivante.

**Il existe une fonction $f$ qui établit une relation entre les descripteurs visuels et les descripteurs audio.**

Autrement dit, la configuration du conduit vocal détermine les caractéristiques du signal de parole. Cette hypothèse paraît raisonnable.

Dans le cas de l'analyse LPC ou MFCC du signal de parole, les descripteurs audio décrivent de manière continue l'espace acoustique. De même, la suite des positions de la langue et des lèvres décrit également une séquence continue. Aussi, la modélisation visio-acoustique proposée ici est un problème de **régression**. Chaque vecteur de descripteurs visuels constitue un vecteur d'entrée du modèle. De même, chaque vecteur acoustique est un vecteur de sortie.

Tout d'abord, nous décrirons succinctement la méthode du descripteur sonde, qui consiste à évaluer quantitativement la contribution de chacun des descripteurs dans la relation visio-acoustiques. Puis nous aborderons la modélisation à l'aide d'une approche linéaire. Enfin, nous présenterons l'utilisation de réseaux de neurones pour le cas de la modélisation non-linéaire. Les concepts théoriques de ce chapitre s'appuient sur [Dreyfus *et al.*, 2004].

## 5.2 Pertinence des descripteurs visio-acoustique

Cette étude cherche à répondre à la question suivante : Toutes les informations présentes dans les variables du modèle sont-elles pertinentes pour la grandeur que l'on cherche à modéliser ? La méthode du descripteur sonde permet de répondre à cette question.

### 5.2.1 Sélection des variables par la méthode du descripteur sonde

De manière générale, une méthode de sélection des variables cherche à classer celles-ci par ordre de pertinence décroissante, afin de savoir quelle entrée explique le mieux la grandeur de sortie à modéliser. Considérons en premier lieu un modèle linéaire par rapport aux paramètres. Une telle sélection des entrées peut s'effectuer à l'aide de la procédure d'orthogonalisation suivante, dans l'espace des observations, c'est-à-dire dans un espace dont la dimension est égale au nombre d'exemples de la base d'apprentissage ; dans cet espace, chaque variable, et la grandeur à modéliser, sont représentées par un vecteur dont les composantes sont les valeurs de cette variable, ou de la grandeur à modéliser, présentes dans la base d'apprentissage.

1. Choisir la variable la plus corrélée avec la grandeur de sortie
2. Projeter le vecteur de sortie et les vecteurs des autres entrées sur le sous espace orthogonal à l'entrée sélectionnée lors de l'étape (1)
3. Itérer 1. et 2.

Cette procédure peut être réalisée à l'aide de l'algorithme de Gram-Schmidt. Ensuite, il est nécessaire de choisir les variables qui doivent être conservées, en fonction de leur capacité à expliquer la grandeur de sortie. Pour aider à ce choix difficile, on ajoute aux vecteurs d'entrée une variable aléatoire, nommée **sonde**, sans relation avec la sortie. Schématiquement, on procède au classement et on élimine toutes les variables qui sont moins bien classées que la sonde.

### 5.2.2 Résultats obtenus et Interprétation

Dans le cadre de ce document, plusieurs vecteurs d'entrée ont été proposés.
- **E1** : Intervalle de définition et coefficients de la spline d'ordre 4 modélisant la surface de la langue (6 coefficients)
- **E2** : Coordonnées de l'image ultrasonore dépliée et filtrée dans le *TongueSpace* (Nous ne gardons que 20 *EigenTongue* soit 20 coefficients)
- **E3** : Coordonnées 2D des points de plus forte courbure du profil du visage modélisant la position des lèvres et de la commissure du locuteur, courbure en ces points et angle d'ouverture de la bouche [1]. Nous résumerons cet ensemble par « profil des lèvres ».

De même plusieurs vecteurs de sortie sont disponibles :
- **S1** : Coefficients LSF (12 valeurs)
- **S2** : Coefficients MFCC (12 valeurs)

La modélisation visio-acoustique utilisant la segmentation ALISP sera abordée à la section 5.6. **Dans les sections suivantes, nous nous plaçons dans le cas de la description synchrone des images avec l'audio**. Dans ce cas, une série de descripteurs visuels décrivant la configuration du conduit vocal à une cadence de 33 ms, doit être associée à une série de descripteurs acoustiques décrivant le signal de parole sur le même intervalle temporel. Bien que cette description du signal de parole soit grossière, cette formalisation permet une modélisation plus simple, au moins dans une première approche.

Les résultats obtenus sont surprenants. La méthode du descripteur sonde indique que tous les descripteurs de l'image ultrasonore, à savoir les entrées **E1** et **E2** « explique »de manière pertinente les sorties **S1** comme **S2**. En revanche, les entrées **E3**, regroupant les informations sur la position des lèvres sont assez mal classées, notamment les valeurs de courbure qui obtiennent un score proche de celui du descripteur sonde. Il est assez difficile d'interpréter l'ordre de classement des entrées **E2**, les résultats obtenus par chacune des *EigenTongues* étant similaires.

Ainsi, la plupart des descripteurs acoustiques et visuels semblent pertinents, et nous pouvons alors envisager la modélisation visio-acoustique.

## 5.3 Modélisation linéaire de la relation visio-acoustique

Posons l'hypothèse suivante : « La fonction $f$ qui établit une relation entre les descripteurs visuels et les descripteurs acoustique est linéaire par rapport aux entrées ».

On note $N$, le nombre d'images considérées pour la modélisation, $N_v$ le nombre de descripteurs visuels et $N_a$ le nombre de descripteurs acoustiques. Soit X la matrice de taille $N_v$ par $N$, composée des valeurs des $N_v$ descripteurs visuels, pour les $N$ images. Soit $Y_k$ le vecteur composé des valeurs du $k^{eme}$ descripteur acoustique pour chacune des $N$ images.

Le problème direct peut s'écrire sous forme matricielle : Résoudre le problème de régression linéaire, c'est trouver les coefficients de la matrice $A$ telle que :

$$AX = Y_k \tag{5.1}$$

En notant $X^T$ la transposée de la matrice des entrées X, nous obtenons :

$$A = (X^T X)^{-1} X^T Y_k \tag{5.2}$$

La résolution de cette équation pour chacun des $N_a$ descripteurs acoustiques nous fournit $N_a$ modèles visio-acoustiques linéaires. Les résultats obtenus seront présentés à la section 5.5.

## 5.4 Modélisation non-linéaire de la relation visio-acoustique

Posons maintenant l'hypothèse suivante : « La fonction $f$, qui établit une relation entre les descripteurs visuels et les descripteurs acoustiques est non linéaire par rapport aux variables ». Le problème de régression devient donc non-linéaire. Il peut être traité à l'aide d'un réseau de neurones artificiels que nous décrivons brièvement maintenant.

### 5.4.1 Principe des Réseaux de neurones artificiels

Les Réseaux de Neurones Artificiels (RNA) sont des combinaisons de fonctions non linéaires élémentaires appelées neurones « formels »ou simplement « neurones ».

[1] Cette angle est défini comme l'angle entre les vecteurs $\overrightarrow{CL_{sup}}$ et $\overrightarrow{CL_{inf}}$ ou $C$ est le point représentatif de la commissure et $L_{sup}$ (resp. $L_{inf}$), celui de la lèvre supérieure (resp. inférieure)

### Neurone formel

Un neurone formel est est une fonction non linéaire paramétrée, qu'il est commode de représenter graphiquement comme indiqué sur la figure 5.1.

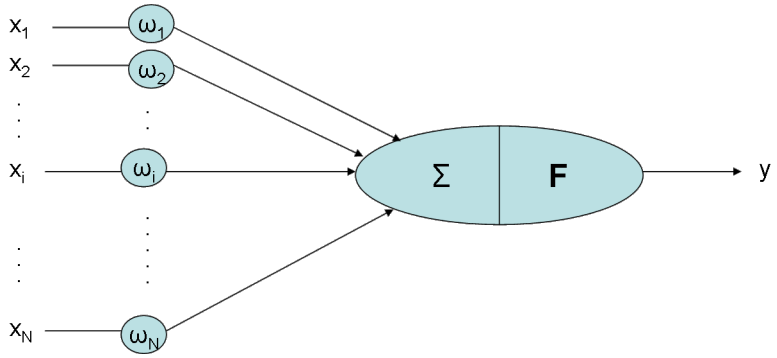


FIG. 5.1 – Représentation graphique d'un neurone formel

Soit $x = (x_1, x_2, ..., x_N)$ le vecteur des variables et $\omega = (\omega_1, \omega_2, ..., \omega_N)$ des paramètres, parfois appelés « poids synaptiques ». La sortie $y$ est définie par :

$$y = F(\sum_{i=1}^{N} \omega_i x_i) \tag{5.3}$$

où $F$ une fonction **non-linéaire**, appelée fonction d'activation. Une des fonctions les plus couramment utilisée est la fonction sigmoïde (tangente hyperbolique). La sortie de chaque neurone est donc une fonction non-linéaire des entrées et des paramètres.

### Perceptron multi-couches

Dans sa version la plus simple, qui est celle que nous avons mise en œuvre dans ce travail, un Perceptron multi-couches (PMC) est une combinaison linéaire de neurones décrits au paragraphe précédent, appelés « neurones cachés ». La figure 5.2 montre un exemple de PMC. Soit $N$ le nombre d'entrées, et $M$ le nombre de neurones cachés, la sortie $y_i$ est définie par :

$$y_i = \sum_{j=1}^{M} \omega_{ij} F(\sum_{k=1}^{N} \omega_{jk} x_k) \tag{5.4}$$

Un PMC est donc défini par le nombre de variables, le nombre de « neurones cachés »et le nombre de sorties, ce qui détermine le nombre de paramètres du modèle. Un PMC est un outil de modélisation très puissant. En effet, il présente la propriété d'**approximateur universel**. Cette caractéristique est décrite dans [Hornik *et al.*, 1989] et dans [Dreyfus *et al.*, 2004] ; nous la rappelons ici :

Toute fonction bornée, suffisamment régulière, peut être approchée uniformément, avec une précision arbitraire, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire.

Cette propriété prouve l'existence d'un tel réseau mais ne fournit pas de méthode pour trouver les paramètres du réseau [2] L'estimation des paramètres ou « poids synaptiques »est l'objectif de l'étape d'apprentissage, que nous décrivons maintenant.

### Apprentissage supervisé

Soit $X$ une observation, $f$ la fonction à modéliser et $Y$ tel que $Y = f(X)$ On appelle ensemble d'apprentissage, l'ensemble des observations (données) utilisées pour cette estimation des paramètres.

L'ensemble d'apprentissage est de taille finie. On connaît donc en certains points de l'espace des variables, les valeurs que doit avoir la sortie du réseau. L'apprentissage « supervisé »consiste, à partir de ces exemples, à

[2] Les réseaux de neurones présentent une autre caractéristiques très intéressante, la parcimonie. Cette propriété est décrite dans [Dreyfus *et al.*, 2004], mais par soucis de concision, elle ne le sera pas dans ce document.

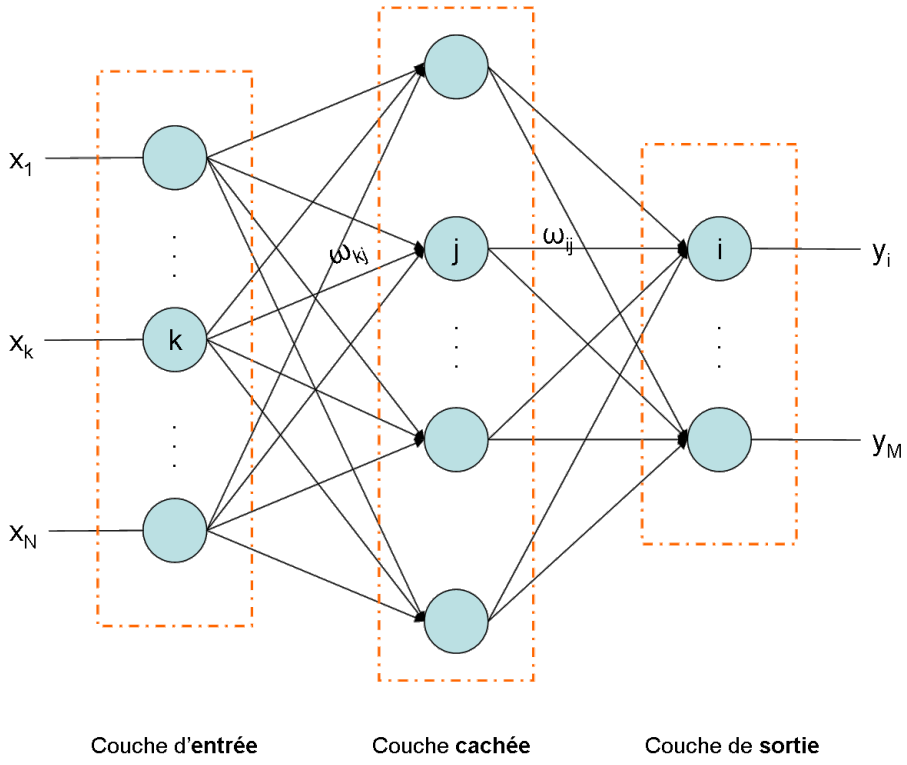


FIG. 5.2 – Architecture d'un Perceptron Multi-Couches

trouver l'ensemble des paramètres pour lesquels une fonction de coût, qui reflète l'écart entre les observations qui figurent dans la base d'apprentissage et les prédictions du réseau, est minimale. Il s'agit donc d'un problème d'optimisation et différents algorithmes peuvent être utilisés pour le résoudre. Ces derniers ne seront pas décrits dans le cadre de ce document ; citons néanmoins la méthode de *Levenberg-Marquardt* et celle de *Polak-Ribiere*, méthode qui sera utilisée dans le cadre de cette étude. Le plus souvent et notamment dans la présente étude, la fonction de coût est celle des moindres carrés, c'est-à-dire la somme des carrés des écarts entre les observations et les prédictions.

### Sélection de la complexité du modèle

La modélisation par apprentissage à partir d'exemples nécessite que la complexité du modèle, (c'est-à-dire, schématiquement, le nombre de ses paramètres), soit adaptée à la complexité des données à modéliser. En effet, il faut non seulement que le modèle fournisse des prédictions satisfaisantes pour les observations de l'ensemble d'apprentissage, mais également qu'il soit capable de fournir des résultats corrects pour d'autres points. Cette propriété est appelée **capacité de généralisation du modèle**. Si le modèle (réseau de neurones ou tout autre modèle paramétré) possède un trop grand nombre de paramètres, sa sortie peut passer très précisément par les exemples d'apprentissage mais fournir des prédictions inexactes entre ces points. Ce phénomène est appelé **surajustement**. A l'inverse, si le réseau possède un nombre de paramètres trop faible, il n'est pas suffisamment « riche »pour modéliser la fonction de régression inconnue. Le choix de la complexité du modèle fait donc l'objet d'un compromis.

La technique de la validation simple est couramment utilisée pour la sélection de modèles. L'ensemble des observations disponibles est segmenté en deux sous-ensembles disjoints. Le premier constitue l'ensemble d'apprentissage. Le second, généralement de taille plus petite, est nommé ensemble de validation. Il n'est pas utilisé lors de l'apprentissage mais il sert à évaluer la capacité de généralisation du modèle. L'étape de validation se déroule selon le schéma suivant :

- 1. Un modèle est fixé et l'estimation de ses paramètres est effectuée sur l'ensemble d'apprentissage.
- 2. La performance du modèle sur les entrées de l'ensemble de validation est évaluée.
- Les étapes 1 et 2 sont répétées, le modèle choisi est celui qui a conduit à une performance maximale sur l'ensemble de validation.

Nous pouvons maintenant mettre en œuvre un PMC pour la modélisation visio-acoustique.

### 5.4.2 Mise en œuvre pratique de la modélisation visio-acoustique

Les propriétés du corpus IEEE/Harvard (cf 3.2.1) permettent de définir facilement des sous-ensembles pour l'apprentissage et la validation. Ainsi sur les 72 listes disponibles, 71 sont dédiées à l'apprentissage et une seule à la validation.

**Modélisation 1** Les descripteurs visuels sont :
- Intervalle de définition et coefficients de la spline d'ordre 4, modélisant la surface de la langue (6 coefficients)
- Profil des lèvres (10 coefficients)

Les descripteurs de sortie sont les coefficients LSF, calculés selon la méthode du **fenêtrage grossier** (12 coefficients par image, *cf.* figure 4.1).

**Modélisation 2** Les descripteurs visuels sont :
- Coordonnées de l'image ultrasonore dépliée et filtrée dans le *TongueSpace* (20 coefficients,*cf.* section 3.3.3)
- Profil des lèvres

Les descripteurs de sortie sont identiques à ceux utilisés pour l'expérience 1.

**Modélisation 3** Dans cette modélisation, nous utilisons les mêmes descripteurs que ceux utilisés dans la modélisation 2. Nous introduisons une notion **dynamique** en complétant les descripteurs visuels de la trame $n$ avec ceux des trames $(n-1)$ et $(n-2)$.

**Modélisation 4** Cette modélisation est identique à la modélisation 3. Les descripteurs audio de type LSF sont remplacés par les coefficients MFCC.

Chaque modélisation à l'aide d'un PMC est précédée de sa version linéaire. Le résultat apporté par la régression linaire permet de quantifier la non-linéarité de la fonction recherchée.

Rappelons enfin que **la méthode de sélection des variables a retenu toutes les variables candidates**, pour chacune des modélisations proposées.

## 5.5 Résultats de la modélisation et interprétations

**Techniques d'estimation de la performance du modèle**

Comme nous l'avons indiqué précédemment, l'étape d'apprentissage adapte les paramètres du modèle, en minimisant une fonction de coût. On note EQMA (*resp.* EQMV) l'erreur quadratique moyenne entre les prédictions du modèle et les exemples de la base d'apprentissage (*resp.* validation). Au terme de l'apprentissage, la valeur de l'EQMA permet de conclure sur la performance générale du modèle et l'EQMV sur sa capacité de généralisation.

Cependant, la valeur de l'EQMA, comme celle de l'EQMV, dépend de l'intervalle de variation du descripteur de sortie considéré. Or, un coefficient LSF n'évolue pas sur le même intervalle qu'un coefficient MFCC. Aussi, afin de comparer les modélisations mettant en jeu des descripteurs acoustiques différents, nous utiliserons les indices suivants :

$$\alpha_a = \frac{\sqrt{EQMA}}{y_{imax} - y_{imin}} \tag{5.5}$$

$$\alpha_v = \frac{\sqrt{EQMV}}{y_{imax} - y_{imin}} \tag{5.6}$$

$y_i$ étant le descripteur acoustique considéré.

Un représentation graphique des résultats est également utilisée. Elle consiste à tracer les prédictions du modèle en fonction des valeurs originales : pour un modèle parfait, les points de ce diagramme devraient être alignés. En pratique, la dispersion du nuage de points est un bon indice sur la capacité d'apprentissage et de généralisation du modèle.

Enfin, dans le cadre de la modélisation visio-acoustique, le meilleur estimateur des performances du modèle est l'écoute. Un signal de parole est synthétisé à partir des prédictions acoustiques, il est ensuite comparé au signal de parole original. Dans le cadre des modélisations 1 à 4, le signal de parole est synthétisé en excitant le filtre LPC prédit (via les coefficients LSF), par une fonction d'activation construite à partir des vraies valeurs de la fréquence fondamentale pour les sons voisés et d'un bruit blanc gaussien pour les sons-non voisés. Cette technique permet de mieux estimer la qualité de la modélisation du conduit vocal. Le problème de la prédiction de la fréquence fondamentale n'est pas résolu.

Comme nous l'avons vu dans le cadre de la présentation des coefficients MFCC, l'oreille humaine et l'audition en général, sont des systèmes complexes, dont la capacité à évaluer la qualité d'un signal sonore ne peut être assimilée à une simple erreur quadratique moyenne. Aussi, des méthodes d'évaluation plus sophistiquées sont disponibles ; une des plus utilisée est la distorsion spectrale [Marques et Almeida, 1986] et [Masri, 1996]. Dans cette étude, nous privilégions la mise en œuvre de la modélisation visio-acoustique ; les techniques de synthèse utilisées ne fournissent pas une qualité de signal suffisante pour être évaluée par mesure de la distorsion spectrale. Aussi, n'utiliserons pas ce type de méthode.

**Présentation des résultats et commentaires**

**Sélection du modèle :** Le tableau ci-dessous précise les modèles utilisés pour chaque modélisation (nombres de neurones cachés du PMC), et pour chaque descripteur acoustique $y_i$. La base d'apprentissage est constituée de 71595 exemples, la base de validation en comporte 878.

| ... | $y_1$ | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | $y_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M1 | 8 | 12 | 14 | 12 | 9 | 11 | 10 | 14 | 13 | 11 | 14 | 14 |
| M2 | 10 | 10 | 11 | 9 | 10 | 12 | 13 | 14 | 14 | 15 | 15 | 24 |
| M3 | 8 | 10 | 9 | 10 | 10 | 10 | 11 | 11 | 13 | 12 | 12 | 12 |
| M4 | 15 | 14 | 16 | 15 | 17 | 14 | 17 | 15 | 20 | 14 | 17 | 20 |

Pour les modélisation M1, M2, M3, il semble nécessaire d'utiliser un réseau plus complexe pour modéliser les derniers coefficients LSF. Le spectre dans les hautes fréquences, et les coefficients LSF qui le décrivent, sont généralement assez bruités ; ce comportement est difficile à modéliser.

La comparaison des résultats obtenus pour la modélisation M3, avec ceux obtenus pour la modélisation M4, montre que la prédiction des coefficients MFCC nécessite l'emploi de modèles plus complexes que pour celle des coefficients LSF.

Enfin, l'utilisation des descripteurs visuels *EigenTongues* ne semble pourvoir se faire qu'au prix de modèles plus complexes que ceux nécessaires pour les descripteurs de type « contour ». Les différents modèles étant sélectionnés, nous décrivons maintenant leur performance.

**Performance de la modélisation :** Le tableau suivant présente les valeurs de $\alpha_a$ et $\alpha_v$ (*cf* Eq 5.5 et 5.6) obtenues pour chaque descripteur acoustique $y_i$, pour chaque modélisation [3].

| . | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
|---|---|---|---|---|---|---|
| M1 - Linéaire | (0.12, 0.17) | (0.14, 0.17) | (0.14, 0.17 ) | (0.14, 0.15) | ( 0.14, 0.17) | (0.13, 0.16) |
| M1 - Non-Linéaire | (0.11, 0.17) | (0.12, 0.16) | (0.13, 0.16) | (0.14, 0.15) | (0.13, 0.16) | (0.12, 0.14) |
| M2 - Linéaire | (0.11, 0.18) | (0.12, 0.16) | (0.12, 0.14) | (0.12, 0.13) | (0.12, 0.15) | (0.11, 0.13) |
| M2 - Non-Linéaire | (0.08, 0.16) | (0.10, 0.11) | (0.11, 0.13) | (0.11, 0.11) | (0.10, 0.13) | (0.09, 0.12) |
| M3 - Linéaire | (0.11, 0.18) | (0.12, 0.15) | (0.12, 0.14) | (0.12, 0.13) | (0.11, 0.14) | (0.19, 0.16) |
| M3 - Non-Linéaire | (0.07, 0.13) | (0.11, 0.13) | (0.10, 0.11) | (0.10, 0.11) | (0.09, 0.12) | (0.09, 0.11) |
| M4 - Linéaire | (0.18, 0.20) | (0.15, 0.16) | (0.11, 0.15) | (0.10, 0.13) | (0.09, 0.13) | (0.10, 0.13) |
| M4 - Non-Linéaire | (0.15, 0.17) | (0.13, 0.13) | (0.10, 0.14) | (0.08, 0.11) | (0.08, 0.12) | (0.08, 0.11) |

| . | $y_7$ | $y_8$ | $y_9$ | $y_{10}$ | $y_{11}$ | $y_{12}$ |
|---|---|---|---|---|---|---|
| M1 - Linéaire | (0.15, 0.18) | (0.13, 0.16) | (0.14, 0.17) | (0.14, 0.15) | (0.13, 0.17) | (0.12, 0.12) |
| M1 - Non-Linéaire | (0.14, 0.17) | (0.12, 0.15) | (0.13, 0.17) | (0.13, 0.14) | (0.13, 0.17) | (0.11, 0.16) |
| M2 - Linéaire | (0.13, 0.16) | (0.11, 0.13) | (0.11, 0.14) | (0.11, 0.12) | (0.11, 0.14) | (0.11, 0.15) |
| M2 - Non-Linéaire | (0.12, 0.14) | (0.10, 0.12) | (0.11, 0.13) | (0.10, 0.11) | (0.10, 0.14) | (0.11, 0.15) |
| M3 - Linéaire | (0.12, 0.15) | (0.11, 0.13) | (0.11, 0.14) | (0.11, 0.12) | (0.11, 0.12) | (0.11, 0.15) |
| M3 - Non-Linéaire | (0.11, 0.13) | (0.09, 0.12) | (0.10, 0.12) | (0.10, 0.11) | (0.10, 0.14) | (0.11, 0.15) |
| M4 - Linéaire | (0.11, 0.13) | (0.11, 0.14) | (0.09, 0.14) | (0.09, 0.12) | (0.09, 0.11) | (0.09, 0.11) |
| M4 - Non-Linéaire | (0.09, 0.12) | (0.10, 0.13) | (0.08, 0.13) | (0.08, 0.11) | (0.07, 0.10) | (0.08, 0.10) |

On observe en premier lieu que les erreurs sur les ensembles d'apprentissage et de validation sont du même ordre de grandeur, ce qui signifie que les modèles ne souffrent pas de surapprentissage : les coefficients LSF mal prédits ont également été mal appris. De manière générale, les modélisations non-linéaires sont plus performantes que les modélisations linéaires. Toutefois, les valeurs obtenues sont relativement proches, ce qui pourrait laisser

[3] Les performances présentées dans ce paragraphe sont celles obtenues sur la base de validation ; en toute rigueur, la qualité d'un modèle ne peut être estimée que sur une base de test, qui n'a été utilisée ni pour l'apprentissage, ni pour la sélection de modèle. Néanmoins, compte tenu du caractère exploratoire de cette étude, nous n'avons pas fait de distinction entre base de validation et base de test.

penser que la modélisation non-linéaire est inutile ; ce qui n'est pas le cas. Traçons par exemple, les prédictions du modèle M3 en fonction des valeurs originales, sur la base de validation, dans le cas de la modélisation linéaire (figure 5.3) et dans le cas non-linéaire (figure 5.4). Sur les différents diagrammes de dispersion, les variables sont normalisées.

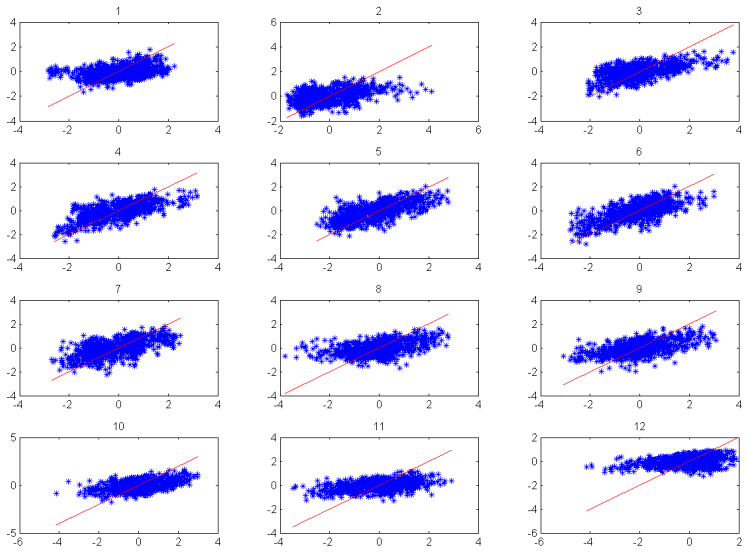


FIG. 5.3 – Représentation graphique des résultats de la modélisation visio-acoustique - Modélisation 3 - Cas linéaire - Base de validation

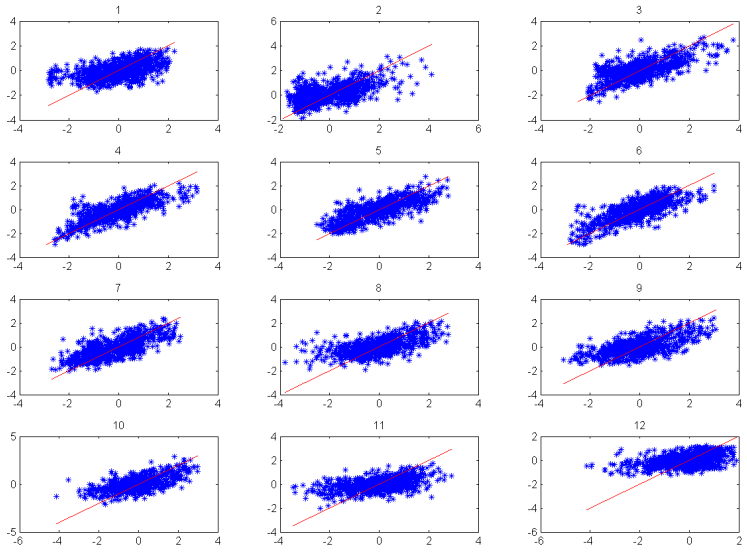


FIG. 5.4 – Représentation graphique des résultats de la modélisation visio-acoustique - Modélisation 3 - Cas non-linéaire - Base de validation

Dans le cas linéaire, les nuages de points ne sont pas centrés sur la diagonale. Dans le cas non-linéaire, les coefficients LSF 3, 4, 5, 8, 9, 10, semblent mieux modélisés. En revanche, la prédiction des coefficients LSF 1, 2, 11, 12, reste assez mauvaise : : la présence de nuages de points dont l'axe est essentiellement horizontal montre que le modèle a simplement appris la valeur moyenne des coefficients, mais n'est pas capable de modéliser leurs variations.

De plus, les descripteurs visuels *EigenTongues* (M2) aboutissent à une meilleure prédiction acoustique que ceux du type « contour »(M1). Ce phénomène est observable sur les figures 5.5 et 5.6, qui illustrent les résultats des modélisations non-linéaires M1 et M2 sur la base de validation. On notera l'amélioration de la prédiction

des coefficients LSF 4, 5, 6, et 7.

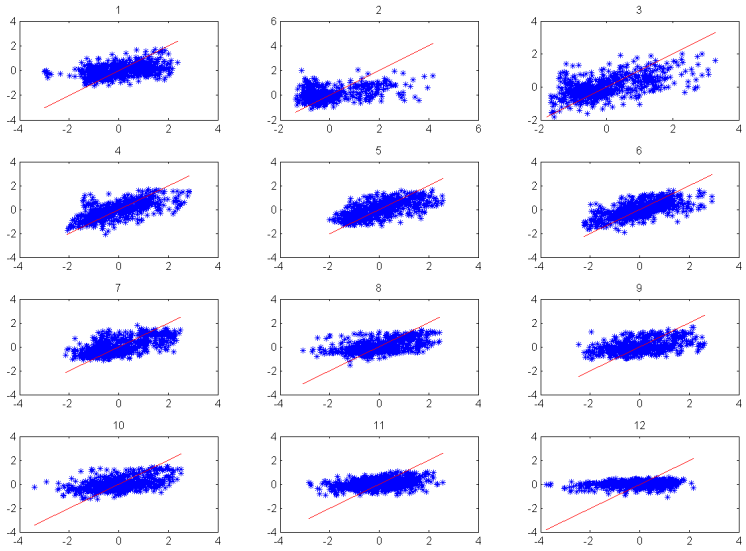

FIG. 5.5 – Représentation graphique des résultats de la modélisation visio-acoustique - Modélisation 1 non-linéaire - Base de validation

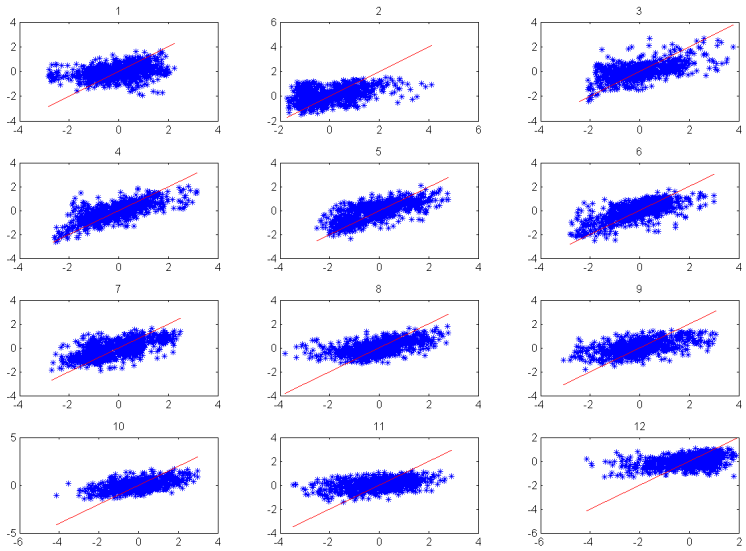

FIG. 5.6 – Représentation graphique des résultats de la modélisation visio-acoustique - Modélisation 2 non-linéaire - Base de validation

L'introduction de la dynamique améliore également la qualité de la modélisation, comme le montre la représentation graphique des résultats de la modélisation non-linéaire M3, sur la base de validation, (*cf.* figure 5.4) par rapport à ceux obtenus par la modélisation M2 (*cf.* figure 5.6). Pour M3, les nuages de points s'alignent de manière plus évidente sur la diagonale que pour M2 : la prédiction est globalement meilleure.

Les valeurs des couples $(\alpha_a, \alpha_v)$ présentées précédemment restent relativement élevées. L'erreur moyenne de prédiction oscille entre 8 et 20% selon le descripteur acoustique considéré, ce qui ne permet pas encore la synthèse d'un signal de qualité suffisante. Néanmoins, les résultats que j'ai obtenus montrent, pour la première fois, qu'il est possible d'effectuer une modélisation visio-acoustique sur un corpus de grande taille. Des progrès substantiels restent à accomplir, mais les pistes pouvant conduire à des améliorations ne manquent pas. Le paragraphe suivant présente l'une d'entre elles, qui sera testée en priorité dans le cadre de ma thèse.

## 5.6 Perspective de la synthèse ALISP

Pour l'instant, nous n'avons pas introduit le résultat de la segmentation ALISP dans le cas de la modélisation. En effet, le problème est tout autre. Il s'agit d'associer à une série de descripteurs visuels, décrivant de manière régulière, la configuration du conduit vocal, des unités segmentales de taille variable. Il ne s'agit plus d'un problème de régression non-linéaire mais de classification dynamique. Une telle approche ne peut être envisagée par l'utilisation de réseaux de neurones simples comme le PMC. On leur préférera des méthodes plus complexes, comme les TDNN[4]. Cependant, une utilisation du sytème ALISP est actuellement possible. L'approche proposée consiste à utiliser ALISP comme un système de reconnaissance de la parole. La modélisation 5 proposée fournit un modèle capable de prédire une description acoustique de la parole utilisant les coefficients MFCC. Or, la segmentation ALISP utilise cette représentation. Il devient alors possible d'utiliser les modèles HMM appris lors de cette segmentation, pour reconnaître, dans la suite des vecteurs MFCC prédite par le réseau, une suite de classes ALISP. La synthèse définitive du signal de parole s'effectue ensuite en choisissant un représentant pour chacune des classes ALISP trouvées, puis en concaténant les représentants choisis. Contrairement à la synthèse LPC, cette approche ne nécessite en théorie aucune prédiction de la fréquence fondamentale.

La représentation MFCC a été introduite dans cette optique. Cette approche est encore à ce jour en cours d'étude. Cette technique nécessite un modèle performant, capable dans un premier temps de prédire une description acoustique de bonne qualité, ce qui n'est aujourd'hui pas encore tout à fait le cas. La figure 5.7 illustre les résultats obtenus par notre modèle actuel de prédiction des coefficients MFCC à partir des descripteurs visuels (*EigenTongues* et profil des lèvres).

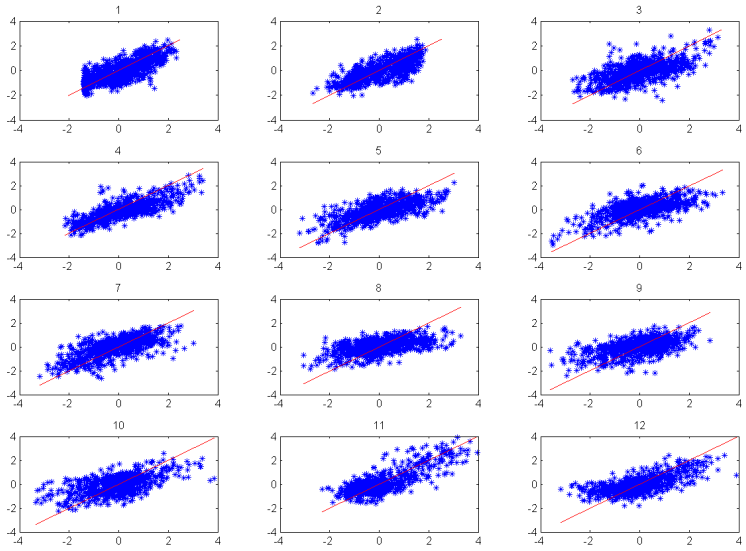


FIG. 5.7 – Représentation graphique des résultats de la modélisation visio-acoustique - Modélisation 4 - Cas non-linéaire - Base de validation

La qualité de la prédiction des coefficients MFCC est proche de celle obtenue pour la modélisation M3 (coefficients LSF) ; elle n'est cependant pas encore suffisante pour effectuer une étape de reconnaissance de parole à l'aide du système ALISP. L'utilisation de ce dernier dans le cadre du projet *Ouisper* reste néanmoins une perspective de première importance.

[4]TDNN : Time Delay Neural Network

# Conclusion générale et Perspectives

Le projet *Ouisper* propose un cadre d'étude pour le traitement multimodal de la parole. Les données acquises par le VTVL permettent d'étudier le mécanisme de production de la parole, d'une part, sous l'angle de la description des différentes configurations du conduit vocal et, d'autre part, sous l'angle de l'analyse acoustique du signal.

La visualisation de l'appareil vocal n'est pas chose facile. Nous avons pu voir les difficultés que présente l'imagerie ultrasonore de cette partie du corps humain. L'interprétation des données passe, dans un premier temps, par la compréhension du système d'acquisition échographique. Profitant de l'expérience du VTVL, nous sommes maintenant capables de comprendre les différentes structures qui constituent une image échographique de l'appareil vocal.

Afin d'en extraire un maximum d'informations, des techniques de pré-traitement ont été proposées. La première consiste à corriger la géométrie conique de l'image échographique tout en optimisant la région d'intérêt. Après avoir vérifié que cette étape ne modifiait pas la nature du bruit de *speckle* qui entache ce type d'image, nous avons sélectionné puis implémenté une technique de filtrage adaptée à ce bruit. Les résultats obtenus sont satisfaisants, les images ainsi nettoyées, sont prêtes à être décrites.

Tout d'abord, nous avons considéré la langue comme la seule structure importante dans les images ultrasonores. Une première approche a donc consisté à extraire son contour. Une méthode a été conçue à cet effet. Fondée sur l'algorithme des moindres carrés médians, elle propose une heuristique permettant l'approximation de la surface de la langue par une courbe spline. La performance de cette méthode se paye au prix d'un temps de calcul important. En outre, cette approche pré-suppose la présence de la langue dans chacune des trames vidéo, ce qui n'est pas systématique. Chez certains locuteurs, la réalisation de certains phonémes conduit à une position de la langue difficile à imager par échographie. Chercher un contour courbe, dans une image qui n'en présente aucun, est voué à l'échec. Néanmoins, ce problème (qui n'est pas propre à l'imagerie de l'appareil vocal) peut être résolu en introduisant un suivi temporel du contour. La mise en œuvre de ce type de technique, notamment celles à base de filtres de Kalman, est une des perspectives envisagées.

Par la suite, nous avons proposé une approche globale de la description des images ultrasonores, nommée *EigenTongue*. Cette méthode utilise la grande similarité des images, et les décrit dans un espace de taille plus petite que celle de l'espace image. Les *EigenTongues*, forment une base de cet espace et ordonnent l'information présente dans l'image. Nous avons montré qu'une image pouvait être assez bien décrite par un nombre restreint de ses projections sur ces *EigenTongues*.

Le traitement des images optiques consiste à paramétrer le profil des lèvres. Un algorithme basé sur la notion de courbure a été proposé. Il permet de localiser les lèvres sans a priori sur leur forme.

Dans le cadre ce cette approche mutimodale du traitement de la parole, nous avons présenté deux catégories de description du signal vocal. La première décrit à intervalles réguliers le contenu spectral du signal de parole. Nous avons introduit l'analyse LPC, sa représentation robuste par les LSF et l'estimation de la fréquence fondamentale. Cette analyse acoustique doit satisfaire la contrainte de synchronisme avec la vidéo tout en étant suffisamment précise pour décrire convenablement le signal de parole. Nous avons également étudié la représentation acoustique à l'aide des coefficients cepstraux. Ce codage vise l'introduction du système de synthèse ALISP. Cette technologie se base sur la décomposition du signal en unités de tailles variables puis en leur classification. L'utilisation de l'analyse ALISP constitue une perspective importante, mais sa mise en œuvre, difficile pour l'instant, n'est pas encore achevée.

L'analyse des données fournit ainsi une série de descripteurs visuels et acoustiques. Pour concevoir un système de synthèse de la parole à partir de la saisie du mouvement de l'appareil vocal, nous avons besoin d'un modèle reliant ces deux séries de descripteurs. Cette modélisation visio-acoustique est rendue possible grâce à l'utilisation de techniques d'apprentissage artificiel. Différentes expériences ont été réalisées. Les résultats obtenus sont comparables à ceux présentés dans [Denby *et al.*, 2006], mais ils ont été obtenus sur un corpus de beaucoup plus grande taille. De plus, les modèles obtenus restituent également les sons non-voisés, ce qui n'était pas le cas pour l'étude préliminaire. [5]. Cependant, le signal de parole synthétique, obtenu avec ce type de modélisation ne peut être considéré comme suffisamment intelligible.

[5] Dans [Denby *et al.*, 2006], les sons non-voisés sont enlevés de la base d'apprentissage pour faciliter la modélisation

Afin d'améliorer les résultats, nous envisageons les perspectives suivantes :

- Une amélioration du protocole d'acquisition des données (choix d'un corpus de parole plus approprié, enregistrement de phonèmes hors contexte)
- Une mise en place de techniques d'estimation de mouvement pour le *tracking* de la langue.
- Une modélisation dynamique à l'aide de TDNN ou HMM.
- Une extension de la technique d'analyse-synthèse ALISP au traitement multimodal de la parole.

Ces travaux seront réalisés dans le cadre de ma thèse au Laboratoire d'Électronique de l'ESPCI, dans le cadre d'un projet de l'Agence Nationale de la Recherche, toujours en collaboration avec Télécom Paris.

# Annexe A

# Annexe - Code source du filtre de diffusion anisotrope

Cette annexe comprend le code source du **filtre anisotrope de diffusion** utilisé dans le cadre de cette étude. L'implémentation s'effectue dans l'environnement MATLAB.

```
% Speckle Filter - Anisotropic diffusion filter
% Wrote by Thomas Hueber
% Ouisper - ESPCI-ENST
% INPUT : im : Gray level input image
%         n  : Scale parameter
% OUTPUT: im_ret = Filtered image

function im_ret = speckle_filter(im,n)

% Config (default)
delta_t = 0.05; % Time discretization step
h = 1; % Spatial discretization step
global_cv = 1; rho = 0.166;

% Discretize space coordinates
x = [1:size(im,1)]*h; y = [1:size(im,2)]*h;

% Calculate logarithm of input data
lim = log(im);

% Run
for t = (1:n)*delta_t
%    fprintf('Iteration %i\n', round(t/delta_t));
    q0 = global_cv * exp(-rho*t);
    for i=1:length(x)
    for j=1:length(y)
        % Define current pixel
        I_i_j = lim(i,j);

        % Add symetric boundary conditions
        if i==1, I_im1_j = I_i_j; else I_im1_j = lim(i-1,j); end
        if j==1, I_i_jm1 = I_i_j; else I_i_jm1 = lim(i,j-1); end
        if i>=length(x)-1, I_ip1_j = I_i_j; else I_ip1_j = lim(i+1,j); end
        if j>=length(y)-1, I_i_jp1 = I_i_j; else I_i_jp1 = lim(i,j+1); end
        if i>=length(x)-2, I_ip2_j = I_i_j; else I_ip2_j = lim(i+2,j); end
        if j>=length(y)-2, I_i_jp2 = I_i_j; else I_i_jp2 = lim(i,j+2); end
        if (i<=length(x)-1) && (j<=length(y)-1), I_ip1_jp1 = lim(i+1,j+1); else I_ip1_jp1 = I_i_j; end
        if (i<=length(x)-1) && (j>1), I_ip1_jm1 = lim(i+1,j-1); else I_ip1_jm1 = I_i_j; end
        if (i>1) && (j<=length(y)-1), I_im1_jp1 = lim(i-1,j+1); else I_im1_jp1 = I_i_j; end

        % Calculate the derivative approximations and the
        % laplacian approximation
        grad_R_i_j = [(I_ip1_j - I_i_j)/h , (I_i_jp1 - I_i_j)/h ];
        grad_L_i_j = [(I_i_j - I_im1_j)/h , (I_i_j - I_i_jm1)/h ];
        laplacian_i_j = (I_ip1_j + I_im1_j + I_i_jp1 + I_i_jm1 - 4*I_i_j)/(h*h);
```

```
        grad_R_ip1_j = [(I_ip2_j - I_ip1_j)/h , (I_ip1_jp1 - I_ip1_j)/h ];
        grad_L_ip1_j = [(I_ip1_j - I_i_j)/h , (I_ip1_j - I_ip1_jm1)/h ];
        laplacian_ip1_j = (I_ip2_j + I_i_j + I_ip1_jp1 + I_ip1_jm1 - 4*I_ip1_j)/(h*h);

        grad_R_i_jp1 = [(I_ip1_jp1 - I_i_jp1)/h , (I_i_jp2 - I_i_jp1)/h ];
        grad_L_i_jp1 = [(I_i_jp1 - I_im1_jp1)/h , (I_i_jp1 - I_i_j)/h ];
        laplacian_i_jp1 = (I_ip1_jp1 + I_im1_jp1 + I_i_jp2 + I_i_j - 4*I_i_jp1)/(h*h);

        % Calculate the diffusion coefficient
        q2_i_j = ((sqrt(norm(grad_R_i_j)^2 + norm(grad_L_i_j)^2)*0.5/I_i_j)^2 - ...
                        0.25*(laplacian_i_j/I_i_j)^2)/((1+0.25*laplacian_i_j/I_i_j)^2);
        q2_ip1_j = ((sqrt(norm(grad_R_ip1_j)^2 + norm(grad_L_ip1_j)^2)*0.5/I_ip1_j)^2 - ...
                        0.25*(laplacian_ip1_j/I_ip1_j)^2)/((1+0.25*laplacian_ip1_j/I_ip1_j)^2);
        q2_i_jp1 = ((sqrt(norm(grad_R_i_jp1)^2 + norm(grad_L_i_jp1)^2)*0.5/I_i_jp1)^2 - ...
                        0.25*(laplacian_i_jp1/I_i_jp1)^2)/((1+0.25*laplacian_i_jp1/I_i_jp1)^2);

        c_i_j = 1/(1+(q2_i_j-q0*q0)/(q0*q0*(1+q0*q0))); % Use form 33
        c_ip1_j = 1/(1+(q2_ip1_j-q0*q0)/(q0*q0*(1+q0*q0)));
        c_i_jp1 = 1/(1+(q2_i_jp1-q0*q0)/(q0*q0*(1+q0*q0)));

        % Calculate divergence
        d_i_j = (1/(h*h)) * (c_ip1_j*(I_ip1_j-I_i_j) + ...
                 c_i_j*(I_im1_j-I_i_j) + ...
                 c_i_jp1*(I_i_jp1-I_i_j) + ...
                 c_i_j*(I_i_jm1-I_i_j));

        % Update pixel - SRAD
        lim(i,j) = lim(i,j) + 0.25*delta_t*d_i_j;
    end
    end
end

% Return Exponential image
im_ret = exp(lim);
```

# Bibliographie


[Atal et Hanauer, 1971] Atal, B. et Hanauer, S. (1971). *Journal of the Acoustical Society of America*, 50(2).

[Attneave, 1954] Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61:183–193.

[Bimbot *et al.*, 1988] Bimbot, F., Chollet, G., P.Deleglise et Montacie, C. (1988). Temporal decomposition and acoustic-phonetic decoding of speech. *International Conference on Acoustics, Speech, and Signal Processing*, 1:445–448.

[Cadic, 2003] Cadic, D. (2003). Implémentation d'une toolbox d'analyse et synthèse vocale selon un modèle harmoniques plus bruit. Mémoire de D.E.A., Télécom Paris.

[Cernocky, 1998] Cernocky, J. (1998). *Speech processing using automatically derived segmental units : applications to very low bitrate coding and speaker verification.* Thèse de doctorat, Université Paris VI.

[Chazan *et al.*, 2000] Chazan, D., Hoory, R., Cohen, G. et Zibulski, M. (2000). Speech reconstruction from mel frequency cepstral coefficients and pitch. *IEEE International Conference on Acoustics, Speech and Signal Processing.*

[Davis et Mermelstein, 1980] Davis, S. et Mermelstein, P. (1980). Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 28(4):357–366.

[de Cheveigné et Kawahara, 2002] de Cheveigné, A. et Kawahara, H. (2002). Yin, a fundamental estimator for speech and music. *Journal of the Acoustical Society of America*, 111:1917–1930.

[Denby *et al.*, 2006] Denby, B., Oussar, Y., Dreyfus, G. et Stone, M. (2006). Prospect for a silent speech interface using ultrasound imaging. *International Conference on Communication Audio and Speech Processing.*

[Denby et Stone, 2004] Denby, B. et Stone, M. (2004). Speech synthesis from real time ultrasound images of the tongue. *International Conference on Communication Audio and Speech Processing.*

[Doval, 1994] Doval, B. (1994). *Estimation de la fréquence fondamentale des signaux sonores.* Thèse de doctorat, Université Paris VI.

[Dreyfus *et al.*, 2004] Dreyfus, G., Samuelides, M., Martinez, J., Gordon, M., Badran, F., Thiria, S. et Hérault, L. (2004). Réseaux de neurones, ed eyrolles, collection algorithmes.

[Dutoit, 2003] Dutoit, T. (2003). Introduction au traitement automatique de la parole. *http ://tcts.fpms.ac.be/cours/1005-07-08/speech/parole.pdf.*

[Feldman et Singh, 2005] Feldman, J. et Singh, M. (2005). Information along contours and object boundaries. *Psychol Rev*, 112(1):243–252.

[G.Fant, 1970] G.Fant (1970). Acoustic theory of speech production with calculations based on x-ray studies of russian articulations.

[Goodman, 1975] Goodman, J. (1975). *Laser speckle and related phenomena*, volume 9, chapitre Statistical properties of laser speckle patterns, pages 9–75. Heidelberg édition.

[Hornik *et al.*, 1989] Hornik, K., Stinchcombe, M. et White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366.

[IEEE, 1969] IEEE (1969). Ieee recommended practice for speech quality measurements.

[I.Stylianou, 1996] I.Stylianou (1996). *Modèles harmoniques plus bruit combinés avec des méthodes statistiques pour la transformation de la parole et du locuteur.* Thèse de doctorat, Télécom Paris.

[Jorgensen *et al.*, 2003] Jorgensen, C., Lee, D. et Agabon, S. (2003). Sub auditory speech recognition based on emg/epg signals. *Proceedings of the International Joint Conference on Neural Networks*, 4:3128–3133.

[Kuan *et al.*, 1985] Kuan, D., Sawchuk, A., Strand, T. et Chavel, P. (1985). Adaptative noise smoothing filter for images with signal dependant noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-7, n°2, p.165-177.*

[Laprie, 2002] Laprie, Y. (2002). Analyse spectrale de la parole - cours. *http ://parole.loria.fr/Documents/analyseParole.pdf.*

[Lee, 1980] Lee, J. (1980). Digital image enhancement and noise filtering by use of local statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-2, p. 165-168.*

[Marques et Almeida, 1986] Marques, J. S. et Almeida, L. B. (1986). A background for sinusoid based representation of voiced speech. *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1233–1236.

[Masri, 1996] Masri, P. (1996). *Computer Modelling of Sound for Transformation and Synthesis of Musical Signals*. Thèse de doctorat, University of Bristol.

[M.Li *et al.*, 2003] M.Li, C.Kambhamettu et M.Stone (2003). Edgetrak, a program for band-edge extraction and its applications. *Sixth IASTED International Conference on Computers Graphics and Imaging*.

[Mosbah, 2005] Mosbah, B. B. (2005). *Utilisation de la mémoire de parole pour la reconnaissance (Application pour des personnes handicapées)*. Thèse de doctorat, Télécom Paris.

[Padellini, 2006] Padellini, M. (2006). *Optimisation d'un schéma de codage de la parole à très bas débit, par indexation d'unités de taille variable*. Thèse de doctorat, Université de Marne-La-Vallée.

[Perona et Malik, 1990] Perona, P. et Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(7):629–639.

[Rabiner et Juang, 1993] Rabiner, L. et Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall PTR.

[Rabiner, 1990] Rabiner, L. R. (1990). A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296.

[Schweitzer *et al.*, 2003] Schweitzer, P., Tisserand, E. et Denby, B. (2003). Design of an ultrasonic lingual profilograph. *World Congress On Ultrasound*.

[Stone, 2003] Stone, M. (2003). A guide to analysing tongue motion from ultrasound images. *Clinical linguistics and phonetics*, pages 359–366.

[Tauber, 2005] Tauber, C. (2005). *Filtrage anisotrope robuste et segmentation par B-spline snake : application aux images échographique*. Thèse de doctorat, Institut National Polytechnique de Toulouse.

[Turk et Pentland, 1991] Turk, M. A. et Pentland, A. P. (1991). Face recognition using eigenfaces. *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586–591.

[V.Dutt et G.Greenleaf, 1995] V.Dutt et G.Greenleaf (1995). *Statistical analysis of ultrasound echo envelope*. Thèse de doctorat, Mayo Clinic College Of Medicine.

[Yu et Acton, 2002] Yu, Y. et Acton, S. T. (2002). Speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 11(11):1260–1270.

[Zheng *et al.*, 1998] Zheng, F., Song, Z., Li, L., Yu, W., Zheng, F. et Wu, W. (1998). The distance measure for line spectrum pairs applied to speech recognition. *In Int. Conf. on Spoken Language Processing (ICSLP-98)*, pages 3 :1123–1126.