Suivi de partition pour l'alignement de la voix chantée

Rong Gong

Rapport de stage Master 2 ATIAM

SUIVI DE PARTITION POUR L'ALIGNEMENT DE LA VOIX CHANTÉE

Auteur :
Rong GONG

Encadrants :
Philippe CUVILLIER
Nicolas OBIN

Équipe-projet MuTant, Équipe Analyse-Synthèse des sons
Institut de Recherche et de Coordination Acoustique / Musique

Mars - Juillet 2014

Résumé

La hauteur du son de la voix chantée est relativement moins juste et moins stable que celle de la musique instrumentale et il existe de nombreuses variabilités, comme le vibrato, l'intonation, etc. Le système d'alignement « musique/partition » AnteScofo se fonde sur l'information de la hauteur du son qui s'adapte bien à la musique instrumentale, mais qui pose des problèmes de fiabilité pour le chant. Ce stage a deux objectifs, le premier est d'intégrer l'information phonétique dans le système pour réaliser l'alignement « voix chantée/texte » ; le deuxième est de fusionner cette information avec celle de la hauteur du son pour rendre le système d'alignement plus robuste. Pour accomplir le premier objectif, on intègre les gabarits de voyelle dans le système, et conduit une expérience pour mettre en évidence que l'enveloppe spectrale générée par la méthode « True envelope » est suffisamment distinctive pour les voyelles. Pour accomplir le deuxième objectif, deux méthodes de fusion d'information sont proposées et parmi lesquelles la méthode de la moyenne arithmétique des probabilités d'observation est prouvée d'être la plus robuste par l'évaluation.

Mots-clefs : AnteScofo, Suivi de partition, Information de phonétique, Enveloppe spectrale, gabarits de voyelle, fusion d'information

Abstract

The pitch of singing voice is relatively less accurate and less stable than that of musical instruments and there are many variabilities, such as vibrato, intonation, etc. The music/score alignment system AnteScofo is based on the pitch information which adapts well to instrument music, but raises reliability problems for singing voice. This internship has two objectives, the first is to integrate the phonetic information into the system alignment to achieve the singing voice/text alignment and the second is to merge this information with that of pitch to make the system more robust. To accomplish the first objective, we integrate the vowel templates into the system, and conduct a preliminary experiment to prove that the spectral envelope generated by the method "True envelope" is sufficiently distinctive for the vowels. To accomplish the second objective, two information fusion methods are proposed and among those the arithmetic averaging of the observation probabilities is proven to be the most robust by the evaluation.

Keywords : AnteScofo, Score following, phonetic information, spectral envelope, vowel templates, information fusion

Remerciements

Je tiens tout d'abord à remercier mes encadrants, que sont Philippe Cuivillier et Nicolas Obin, pour leur aide précieuse et ainsi que leurs conseils avisés afin de suivre ce stage dans les meilleures conditions possibles.

Merci beaucoup au directeur de l'équipe Arshia Cont m'avoir donné les idées inspirantes et m'avoir encouragé pour surmonter ma difficulté d'élocution en public.

Je remercie ensuite mes camarades de bureau Henri, Diego, Marion et Alberto pour leur compagnie pendant ces cinq mois de stage.

Mes camarades de la promo du Master ATIAM ont partagé avec moi une excellente année d'études en France.

Table des matières

Chapitre 1

Introduction

Le suivi de partition possède une longue histoire de recherche [5] [38]. Il consiste à synchroniser une interprétation d'un morceau de musique avec la partition de ce morceau de musique, celle-ci étant connue par l'ordinateur.

Une description minimale de suivi de partition est la suivante : le système possède une représentation de la partition symbolique à l'avance qui est donnée par l'utilisateur et introduit dans le système hors ligne. L'objectif du système est de faire correspondre le flux audio entrant à cette représentation et de décoder la position de partition et le tempo.

Motivation

Notre travail consiste à étendre les fonctionnalités du système AnteScofo (le logiciel de suivi de partition et d'écoute artificielle de STMS [4]) fondé sur un modèle espace-état dans le chapitre 2. Antescofo utilise des gabarits de hauteur et une observation de la hauteur qui s'adapte bien à la musique instrumentale, mais qui pose des problèmes de fiabilité pour le chant. En effet, la hauteur du chant est relativement moins juste que celle de la musique instrumentale et il existe de nombreux ornements, comme le vibrato, l'intonation, etc. (voir la figure 1.1) De plus, si deux états consécutifs possèdent la même hauteur, l'utilisation d'une seul information de hauteur n'est pas suffisante pour distinguer ces deux états.

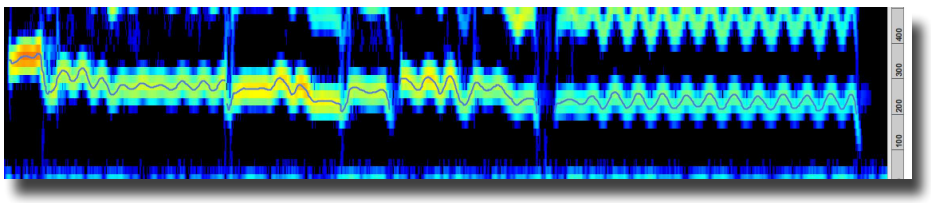

FIGURE 1.1 – Spectrogramme d'un extrait de chant : la courbe grise trace la fréquence fondamentale, révélant le vibrato et la variabilité de la hauteur.

Ces considérations motivent l'utilisation d'une source d'information plus spécifique : l'information phonétique de la voix. Les figures 1.2 et 1.3 montrent les spectrogrammes d'un extrait instrumental et d'un extrait de chant. Nous percevons une information supplémentaire dans la musique chantée, qui n'existe pas dans la musique instrumentale : les paroles. D'un point de vue phonétique, chaque événement possède un phonème. Les voyelles, la subdivision dominante des phonèmes, représentant environ 90% du temps de phonation deviennent naturellement le centre de notre recherche. Il est donc possible d'imaginer un alignement voix chantée/texte, et plus

spécifiquement un alignement voyelle/texte.

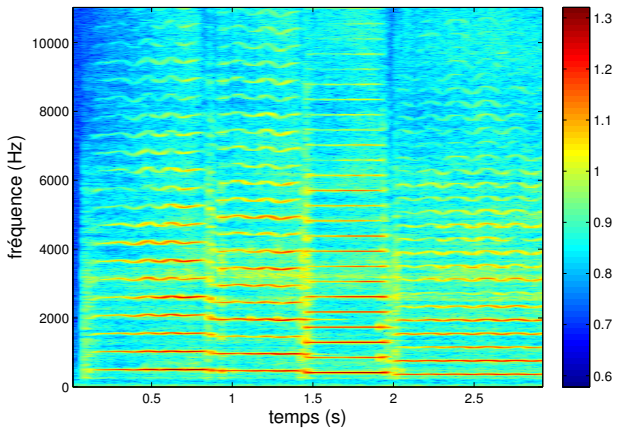

FIGURE 1.2 – Spectrogramme d'un extrait de musique instrumentale

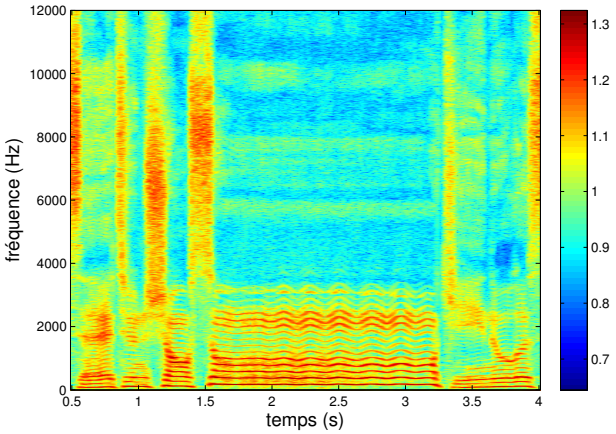

FIGURE 1.3 – Spectrogramme d'un extrait de musique de chant

Plan et contributions principales

Nous proposons d'intégrer l'information phonétique dans le système de suivi de partition Antescofo. Dans un premier temps, nous utilisons uniquement cette information pour la comparer à celle de hauteur. Dans un deuxième temps, nous proposons des méthodes de fusion de cette information avec celle de hauteur pour rendre le système plus robuste.

D'un point de vue de MIR (Musical Information Retrival), l'information de formant provenant de l'enveloppe spectrale permet de classer les voyelles différentes. Par conséquent, nous effectuons un apprentissage statistique d'un ensemble de gabarits de voyelle, à partir d'un descripteur de l'enveloppe spectrale, la « True Envelope » [23] et intégrons cet ensemble de gabarits dans le système d'alignement. Le modèle d'observation est modifié, le signal est représenté par l'enveloppe spectrale au lieu spectre à court-terme.

Une autre contribution de notre travail est d'avoir trouvé une stratégie de la fusion de ces deux informations afin d'améliorer la robustesse du système. Nous proposons deux stratégies : « Early fusion » et « Late fusion ». Pour la première, la nature du problème est de refaçonner les gabarits selon le modèle source/filtre pour qu'ils aient non seulement l'information de la hauteur mais aussi celle de l'enveloppe spectrale. Nous pouvons également étudier le problème dans la perspective de la fusion des probabilités d'observation - « Late fusion ».

Cette mémoire est organisée comme suit : Dans le chapitre 2, nous introduisons les modèles probabilistes, les chaînes de Markov cachées et l'architecture de AnteScofo. Dans le chapitre 3, nous introduisons le modèle source/filtre comme la base théorique de l'enveloppe spectrale, mettons en évidence la capacité distinctive de la « True Envelope », construisons les gabarits de voyelle et les intégrons dans le système d'alignement. Dans le chapitre 4, nous détaillons les stratégies « Early fusion » et « Late fusion ». Nous évaluons la performance du système dans le chapitre 5 suivi par la conclusion et la perspective.

Chapitre 2

État de l'Art : Alignement partition/audio

2.1 Suivi de partition par modèle probabiliste : Modèle de Markov caché

À cause du caractère temporel intrinsèque de la musique, la capacité de représenter et décoder les événements temporels constitue la base de tous les système de suivi de partition. Dans un contexte musical, un état de processus concerne un événement musical étant donné une durée. Un moyen courant pour modéliser les données de séries temporelles est d'implémenter le modèle état-espace. Dans cette section, nous rappelons brièvement la définition d'un modèle de Markov Caché (HMM). Plus de détails se trouvent dans [33].

Notre donnée d'observation est un signal acoustique généré par le musicien. Nous découpons ce signal en τ trames de longueur fixée. Ainsi on note la séquence d'observation $x_0^\tau \stackrel{\text{def}}{=} (x_0, x_1, \ldots, x_\tau)$.

BPM	60	
NOTE	0	0
NOTE	F#4	1/2
NOTE	G4	1/2
NOTE	F#4	1/2
NOTE	F#4	1/2
NOTE	E4	1

Table 2.1 – Un extrait de la partition d'AnteScofo

Nous adoptons une approche générative avec l'hypothèse implicite que le signal audio peut être généré par le modèle d'état-espace de la partition. Un modèle de HMM est défini comme une paire de processus stochastique (S_t, X_t). L'observation x_0^τ est une réalisation d'un processus aléatoire $\{X_t\}$ généré par une séquence d'états cachés s_0^τ qui est une réalisation de $\{S_t\}$. Nous supposons que le processus $\{S_t\}$ est une chaîne de Markov cachée, sur un espace d'états discret construit à partir de la partition (exemple d'une partition sur le tableau 2.1), comme dans l'exemple de la figure 2.1. Par conséquence, le problème de suivi de partition est un problème inverse : retrouver l'état le plus probable associée à la séquence d'observation.

Le processus d'état $\{S_t\}$ n'est pas directement observable mais il peut être estimé à partir de $\{X_t\}$. Selon l'hypothèse de Markov, l'observation X_t est indépendante des états passés et futurs, conditionnellement à l'état présent. Cela correspond à l'hypothèse de quasi-stationnarité

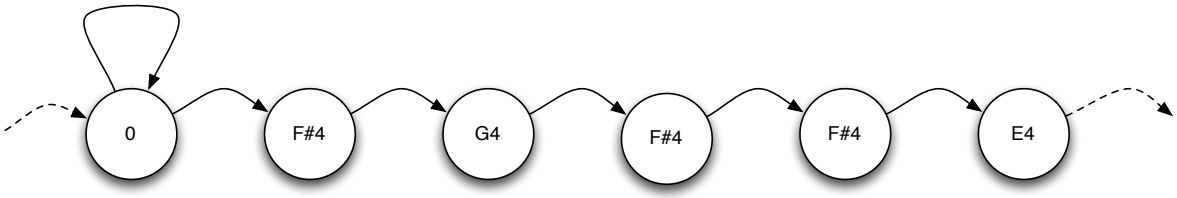

FIGURE 2.1 – Topologie d'une chaîne de Markov caché gauche-droite construit à partir de la partition 2.1

du signal audio pendant un événement musical. De même, l'état présent S_t est indépendant du passé lointain S_0^{t-2}, conditionnellement à son passé immédiat S_{t-1}. Le graphe de dépendance de ces variables aléatoires est récapitulé dans la figure 2.2.

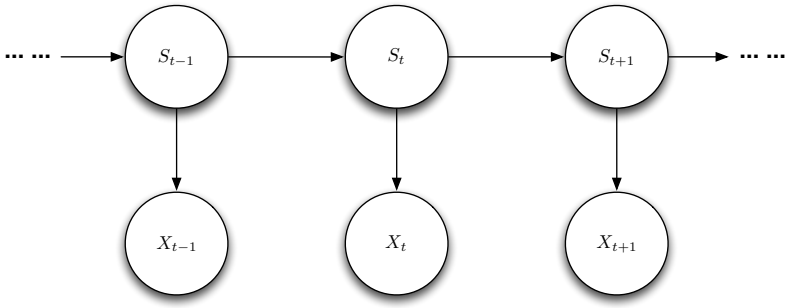

FIGURE 2.2 – Architecture général d'un HMM

Par conséquent, un modèle HMM est défini par deux types de distributions de probabilités :

- Les probabilités de transition du processus d'état $\{S_t\}$

$$\tilde{p}_{jk} = P(S_{t+1} = k | S_t = j) \tag{2.1}$$

  avec $\sum_k \tilde{p}_{jk} = 1$.
- Les probabilités d'observation, qui lient l'observation $\{X_t\}$ à l'état $\{S_t\}$

$$b_j(x_t) = P(X_t = x_t | S_t = j) \quad \text{avec} \quad \sum_y b_j(y) = 1. \tag{2.2}$$

Bien que l'on n'arrive pas à retrouver la séquence d'états exacte, on peut calculer la probabilité $P(X_0^\tau = x_0^\tau | S_0^\tau = s_0^\tau)$. De ce fait, un certain nombre d'estimateurs de la séquence d'états sont possibles. Le plus simple d'entre eux est l'algorithme **Forward**, qui estime à chaque instant t l'état courant le plus probable [33] :

$$\hat{s}_t = \arg\max_{s_t} P(S_t = s_t \mid X_0^\tau = x_0^\tau) \tag{2.3}$$

Dans la section suivante, nous présentons l'architecture de AnteScofo, qui implémente ce modèle.

2.2 Architecture de AnteScofo

AnteScofo est un logiciel de suivi de partition pour la composition musicale développé par l'équipe-projet MuTant à l'IRCAM. Il permet de reconnaître automatiquement la position de partition et le tempo d'un flux audio produisant par un interprète en temps réel.

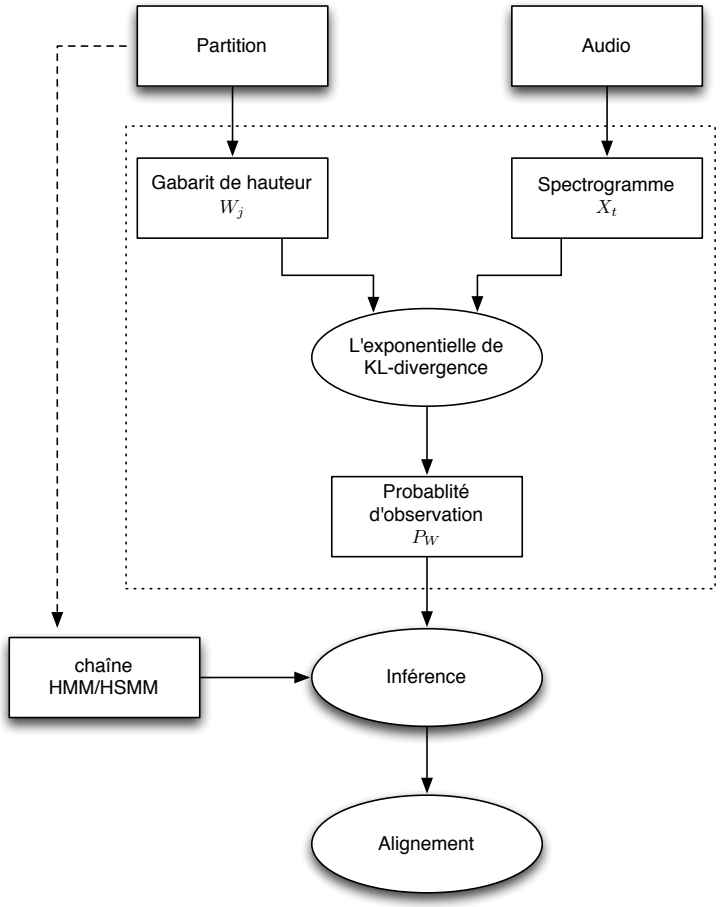

FIGURE 2.3 – Le processus général du système AnteScofo, la partie dans le rectangle pointillé concerne à notre travail.

Antescofo se fonde sur une extension des modèles HMM présentés la section précédente. Il s'agit des modèles cachés hybrides de Markov (HMM) et semi-Markov (HSMM) tels que définis dans l'article [27]. Comme leur étude n'est pas au cœur de notre travail, nous les décrivons succinctement.

2.2.1 Modèle hybride de HMM/HSMM

L'espace d'état généré à partir d'une partition comporte des états Markoviens et semi-Markoviens. L'avantage du modèle de Semi-Markov est qu'il définit explicitement la distribution de probabilité de l'occupation d'un état j : $d_j(.)$.

La figure 2.3 montre le processus général du système. Une fois reçu la partition, AnteScofo construit la chaîne d'états à partir des événements. L'événement temporel correspond à l'état semi-Markovien tandis que celui atemporel correspond à l'état Markovien [4].

Le tableau 2.1 montre un extrait de la partition. La première ligne de ce extrait est un

Meta-événement BPM (*beat per minute*) qui définit le tempo global de ce morceau de musique. À partir de la deuxième ligne, AnteScofo va lire les événements musicaux. Par exemple, « NOTE 0 0 » indique que c'est un silence atemporel correspondant un état Markovien et « NOTE G4 1/2 » une note de hauteur G4 qui dure 1/2 « beat » correspondant un état semi-Markovien. Un état Markovien est paramétré par $i, f0_i$ tandis qu'un état semi-Markovien l'est par $i, f0_i, l_i$, où i est le numéro de l'événement depuis le début de la partition, $f0_i$ est sa hauteur et l_i est sa durée.

2.2.2 Modèle d'observation

Le choix d'un bon modèle de probabilités d'observation b_j est le centre de notre travail. La probabilité d'observation est calculée à partir d'une représentation à court terme du signal sonore. Dans les expériences présentées dans notre travail, la fenêtre d'analyse a pour longueur 92 ms et pour facteur de recouvrement 4, ce qui fixe le compromis entre la résolution temporelle et fréquentielle. Dans Antescofo, le spectrogramme SP_t est choisi comme le descripteur audio pour la trames x_t. Elle est suffisante pour mettre en évidence l'information de hauteur, car elle montre les harmoniques dans le domaine fréquentiel. On compare cette observation avec le gabarit de hauteur construit directement par l'information de hauteur s_j.

Ce modèle d'observation est fondé sur une hypothèse que le son musical est stationnaire pendant un événement. Ceci est approximativement vrai le temps d'une note de hauteur constante, comme celle-ci a tendance à préserver sa structure harmonique au cours de sa durée de vie. L'hypothèse est donc bonne pour les sons harmoniques quasi-stationnaires.

Gabarit de hauteur

S_j est considéré comme la vérité de la distribution de fréquence et construit à partir de la hauteur de l'état s_j. Par conséquent, S_j, autrement dit, le gabarit de hauteur (notons W_j afin de faire la distinction avec le gabarit de voyelle) consiste en les harmoniques comme les pics dans le domaine fréquentiel.

$$W(f) = \sum_{k=1}^{K} e(kf_0)\mathcal{N}(f; kf_0, \sigma^2_{f_0,k}) \tag{2.4}$$

Chaque pic est modelé comme la distribution gaussienne

$$\mathcal{N}(f; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.5}$$

centrée sur les harmoniques kf_0 et la variance σ^2 est relative à la fréquence centrale sur l'échelle musicale logarithmique. L'enveloppe d'harmonique $e(kf_0)$ est une exponentielle décroissante qui correspond à la plupart des notes instrumentales. Dans le système d'AnteScofo, par défaut, un gabarit consiste en $K = 10$ harmoniques et l'écart-type est fixé au demi-ton : $\sigma_{f_0,k} = 2^{1/12}kf_0$. Un exemple du gabarit de hauteur est montré sur la figure 2.4.

La probablité d'observation

La probabilité d'observation définit dans AnteScofo est

$$p_W(x_t|s_j) = \exp[-\beta D(W_j||SP_t)] \tag{2.6}$$

où $D(W_j||SP_t)$ est la divergence Kullback-Leibler entre le gabarit de hauteur W_j de l'état s_j et le spectrogramme SP_t de la trame x_t (afin de faire la distinction avec la « True envelope » de x_t). La probabilité dans l'équation 2.6 a besoin de la normalisation de SP_t et de W_j de telle

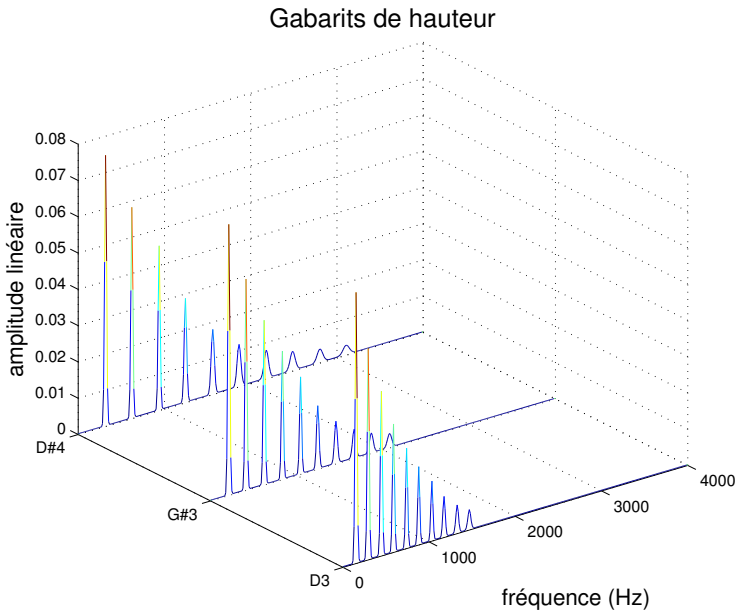

FIGURE 2.4 – Exemples de gabarits pour trois hauteurs différentes.

sorte que leurs sommes respectives valent 1. Par conséquent, on travaille uniquement avec le spectrogramme normalisé en amplitude, ce qui rend la probabilité invariante au volume sonore.

$$D(W_j||SP_t) = \sum_i W_j(f) \log \frac{W_j(f)}{SP_t(f)} \tag{2.7}$$

Notons que la divergence KL de l'équation 2.7 n'est pas une métrique utilisée comme une fonction de la probabilité d'observation. Parce que ses valeurs tombent sur l'intervalle $[0, +\infty]$. Pour convertir l'intervalle de l'équation 2.7 à $[0, 1]$, on la passe à travers l'exponentielle de KL divergence de l'équation 2.6. Le facteur β est fixé à 0,5 dans AnteScofo. Ce modèle correspond à l'hypothèse d'une loi multinomiale, comme expliqué dans [20].

2.2.3 Exemple d'alignement

La figure 2.5 montre un résultat d'alignement sur un extrait du chant. La vérité-terrain saute au prochain état à l'attaque d'événement. Le chemin d'alignement suit généralement bien la vérité-terrain, sauf qu'aux états 9 et 10 qui possèdent la même hauteur mais pas la même parole, donc, le même gabarit est utilisé pour calculer la probabilité d'observation de ces deux états. Ce cas illustre la remarque suivante : si deux états consécutifs possèdent la même hauteur, l'utilisation d'une seule information de hauteur n'est pas suffisante pour distinguer ces deux états.

Également, les bandes verticales aux états 5, 9 et 10 montrent la variabilité de la fréquence fondamentale du chant. Ces considérations motivent l'utilisation d'une source d'information de la voix : l'information phonétique.

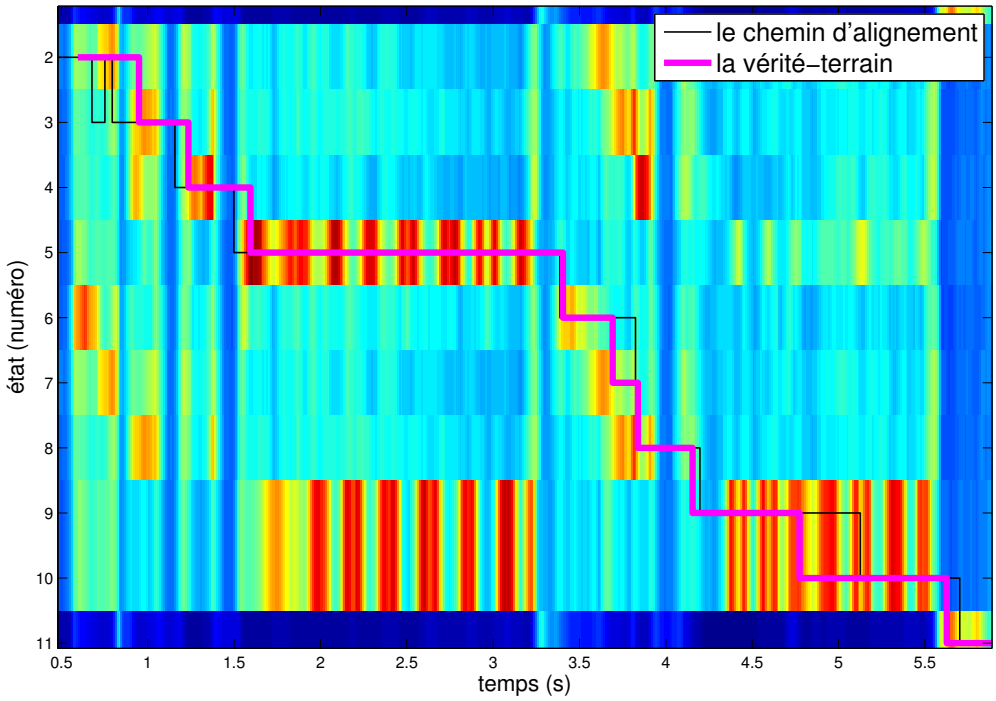

FIGURE 2.5 – La matrice de la probabilité d'observation et le chemin d'alignement. La couleur rouge foncé indique la grande probabilité.

Chapitre 3

Alignement Texte/voix chantée

L'information phonétique est introduite dans ce chapitre comme une source d'information complémentaire aux hauteurs pour l'alignement de la voix chantée.

Dans la section 3.1, le modèle source/filtre de la voix est introduit pour permettre la séparation des informations de la source d'excitation (F0, la hauteur) des informations du conduit vocal (enveloppe spectrale, les phonèmes). Dans la section 3.2, on présente le cepstre et la méthode True Enveloppe pour l'estimation de l'enveloppe spectrale. Dans la section 3.3, on présente une méthode d'extraction du gabarit de voyelle pour intégrer l'information de voyelle dans le système d'alignement. Dans la section 3.4, on conduit une expérience préliminaire pour évaluer la capacité de l'enveloppe spectrale à représenter les voyelles en voix chantée.

3.1 Le modèle source-filtre de la voix

3.1.1 Introduction

Modèle physique	Modèle signal	Perception
vibration des cordes vocales	source d'excitation (F0)	hauteur
résonances du conduit vocal	filtre résonateur (enveloppe spectrale)	« timbre »

TABLE 3.1 – La voix : représentation physique, signal, et perceptive

L'appareil vocal humain est constitué d'un ensemble d'organes susceptibles de produire une grande variété de sons. On présente deux organes ici qui sont liés à notre travail : les cordes vocales et le conduit vocal.

La source d'excitation comprend : 1) la vibration des cordes vocales à une fréquence fondamentale (F0) qui détermine la hauteur du signal vocal ; 2) et du bruit. Le conduit vocal assure une fonction de filtrage acoustique (filtre résonateur) des signaux de source qui joue un rôle essentiel pour le « timbre » de la voix. La cavité formée par le conduit vocal possède des résonances et des anti-résonances [6].

3.1.2 Phonèmes

Dans la communication humaine (la voix parlée et la voix chantée), les sons émis véhiculent un message linguistique. En linguistique, le phonème constitue la plus petite unité distinctive utilisée dans la communication orale et chantée. Du point de vue de la production, les phonèmes sont définis par une configuration spécifique de l'appareil phonatoire, et donc des résonances du

conduit vocal [32, 13]. La subdivision élémentaire qui apparaît est liée au mode d'excitation du conduit vocal, et à la stabilité de ce dernier : c'est la séparation voyelles/consonnes [6].

Les voyelles correspondent d'une part à une excitation quasi-périodique (donc assez longue) délivrée par les cordes vocales et d'autre part à une conformation stable du conduit vocal. Les voyelles sont des segments relativement stables et d'énergie assez élevée. On distingue les différentes voyelles grâce à l'emplacement des premiers formants - les modes de résonances du conduit vocal.

La position des deux premiers formants permettent de caractériser les voyelles d'un français. On établit par un schéma classique la répartition des voyelles dans le plan F1, F2 (fréquence du premier formant, fréquence du deuxième formant) qui fait apparaître le « triangle vocalique » (figure 3.1) [6].

phonème	XSAMPA	exemple	phonème	XSAMPA	exemple
ø	2	p**eu**	œ	9	j**eu**ne
œ̃	9∼	br**un**	ə	@	p**e**tit
ɛ	E	c**è**pe	ɔ	O	s**o**rt
ɑ	A	p**â**te	a	a	l**a**
ã	a∼	**an**chois	e	e	beaut**é**
ɛ̃	e∼	pa**in**	i	i	v**i**e
o	o	rése**au**	õ	o∼	pige**on**
u	u	f**ou**	y	y	ch**u**te

TABLE 3.2 – Codage XSAMPA (ASCII) des voyelles du français. [39]

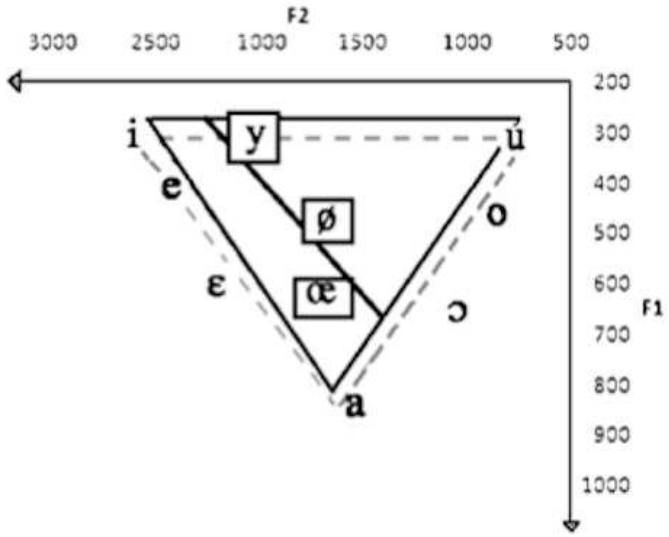

FIGURE 3.1 – Triangle vocalique, position relative des deux premiers formants en français (en Hz) [17]

3.1.3 Voix parlée et voix chantée : voyelles vs. consonnes

Les voyelles et les consonnes occupent des fonctions différentes en voix parlée et en voix chantée.

En voix parlée, la transmission du message linguistique est primordial : l'ensemble des phonèmes (voyelles et consonnes) sont également importants dans la transmission de ce message.

En voix chantée, il est généralement admis que la transmission du message musical (la partition musicale) prévaut sur la transmission du message linguistique (le texte linguistique) [26, 14].

Par rapport aux voyelles, les consonnes sont produites avec un resserrement du conduit vocal par les rapides occlusions du conduit vocal possèdent une durée très brève. Elles constituent en ensemble de sons très hétérogène aussi bien du point de vue articulatoire qu'acoustique [29].

En outre, les voyelles constituent un support idéal pour porter la ligne mélodique (partie stable) tandis que les consonnes constituent des obstacles à l'écoulement d'air continu nécessaire pour porter la ligne mélodique (partie transitoire). Par exemple et en conséquence, les voyelles représentent environ 90% du temps de phonation en chant lyrique [3].

En conséquence, nous concentrerons dans la suite la question de l'alignement du texte et de la voix chantée à l'alignement des voyelles et de la voix chantée. D'un point de vue pratique, l'alignement des voyelles devrait fournir une précision temporelle suffisamment satisfaisante dans le cadre d'un système de suivi de partition pour la voix chantée.

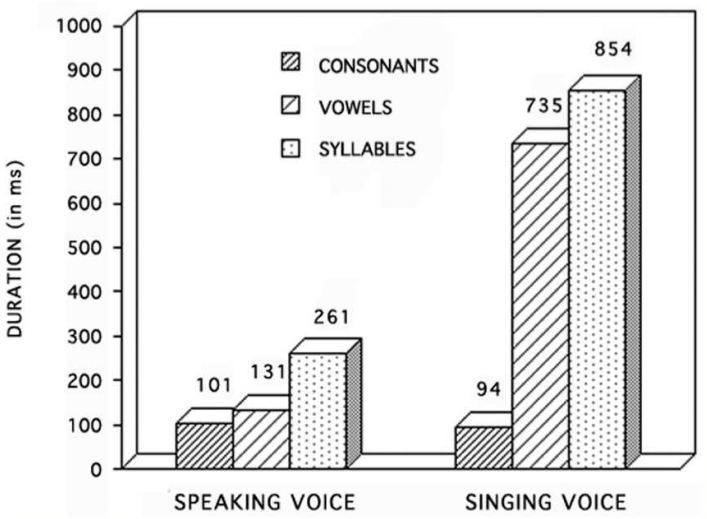

FIGURE 3.2 – La durée moyenne des consonnes, voyelles et syllabes dans la voix parlée et la voix chantée [3]

3.1.4 Modèle source/filtre

Un modèle standard de représentation du signal vocal est le modèle source/filtre : le signal de voix correspond à la convolution de la source d'excitation des cordes vocales (F0+bruit) et du filtre du conduit vocal.

Il est commode de simplifier le modèle acoustique linéaire source/filtre en choisissant une source de spectre plat, et en réunissant dans un filtre la contribution du conduit vocal. Si p est le signal source d'excitation : un train périodique d'impulsions, et h le filtre, la réponse impulsionnelle s'écrit par définition :

$$s(t) = (p \otimes h)(t) \tag{3.1}$$

Et la réponse fréquentielle :

$$S(f) = P(f)H(f) \tag{3.2}$$

où $P(f)$ est la réponse fréquentielle la source d'excitation, $H(f)$ est la réponse fréquentielle du filtre et aussi l'enveloppe spectrale de $S(f)$. Les voyelles, étant par essence portées par la voix,

sont naturellement voisées. Dans le modèle source/filtre, la source d'excitation correspond à la hauteur du son $P(f)$ et l'enveloppe spectrale $H(f)$ est une estimation du filtre du conduit vocal, et donc encode l'information portée par les phonèmes. Elle est par exemple couramment utilisée en traitement automatique de la parole : depuis la reconnaissance [40], la transformation [9] à la synthèse [22].

Nous allons nous concentrer sur l'estimation des caractéristiques du filtre : l'enveloppe spectrale.

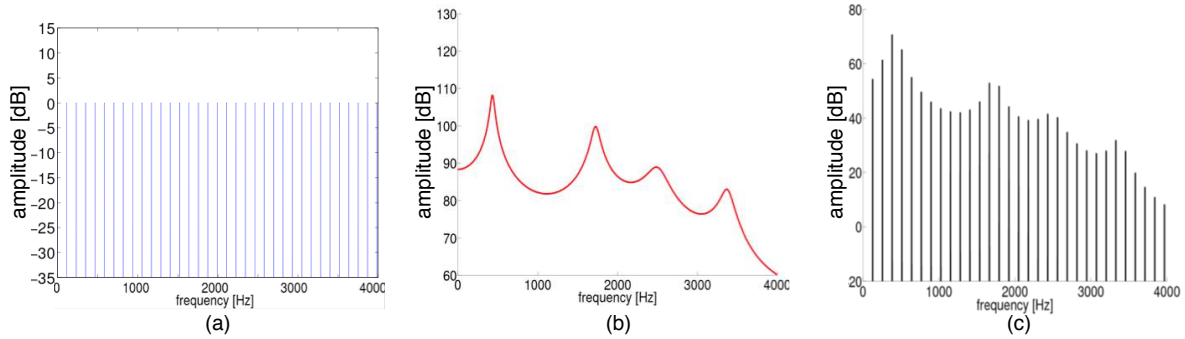

FIGURE 3.3 – Le modèle source/filtre : (a) spectre du signal source $P(f)$; (b) réponse fréquentielle du filtre $H(f)$; (c) spectre résultant $S(f)$

3.2 Enveloppe spectrale

Dans le cadre du modèle source/filtre, la question essentielle pour caractériser les voyelles se concentre sur l'estimation de l'enveloppe spectrale. Par définition, l'enveloppe spectrale représente le filtre du conduit vocal à l'exclusion de la source d'excitation, et donc de la fréquence fondamentale F0. Pratiquement, l'enveloppe spectrale est une fonction qui représente les résonances du filtre du conduit vocal, mais suffisamment lisse pour ne pas encoder la périodicité due à la source d'excitation (la F0).

Dans le modèle source/filtre, l'enveloppe spectrale correspond, à un facteur près, au module de la réponse en fréquence du filtre. Deux représentations prédominent pour l'estimation de l'enveloppe spectrale : la modélisation auto-régressive [28], et la modélisation cepstrale [31].

La représentation auto-régressive (autrement connu sous le nom de codage par prédiction linéaire - LPC), est fondée sur l'hypothèse que le filtre ne contient que des pôles [28]. La représentation cepstrale est fondée sur l'hypothèse d'un filtre à réponse impulsionnelle finie de type « cepstre » [31]. Quelle que soit la représentation, l'adaptation de la représentation au signal est cruciale pour l'estimation du filtre du conduit vocal.

L'avantage du cepstre pour l'estimation de l'enveloppe spectrale est double : la séparation théorique de la source d'excitation et du filtre, et l'adaptation théorique de la représentation au signal. Nous présentons ici la représentation cepstrale, et la méthode de la « True Envelope » [25, 23] pour l'estimation de l'enveloppe spectrale. Cette représentation sera dans la suite intégrés au système d'alignement de suivi de partition pour l'alignement du texte et de la voix chantée.

3.2.1 Le cepstre

On considère le modèle source/filtre tel que présente dans la partie précédente : soit dans le domaine fréquentiel :

$$S(f) = P(f)H(f) \tag{3.3}$$

En ne considérant que le module de la transformée de Fourier, et en passant le logarithme :

$$\log(|S(f)|) = \log(|P(f)|) + \log(|H(f)|) \tag{3.4}$$

En prenant la représentation en logarithme, nous sommes passés d'une multiplication des composantes source et filtre à une simple addition ; ainsi, en supposant que les deux composantes interviennent à des quéfrences (la transformée de Fourier inverse appliquée au logarithme de la transformée de Fourier du signal) différentes, nous pouvons donc simplement séparer les deux contributions. La partie « haute quéfrence » et le filtre et la partie « basse quéfrence » est la source d'excitation. Par la transformée en cosinus, on obtient :

$$c(n) = DCT(\log(|S(f)|)) \tag{3.5}$$

où DCT est la transformée en cosinus En utilisant les propriétés de parité et de périodicité de $\log(|S(f)|)$, on en déduit :

$$\log(|S(f)|) = c(0) + 2\sum_{k=1}^{K} c(k)\cos(2\pi f k) \tag{3.6}$$

où K est le nombre de points de la DCT. Le k-ième coefficients cepstral représente donc la contribution de la cosinusoïde de fréquence $2\pi fk$ au spectre d'amplitude logarithmique.

En revenant sur la séparation des contributions de la source et du filtre, nous pouvons trouver l'ordre p optimal de l'estimation de l'enveloppe spectrale. C'est-à-dire l'ordre du cepstre maximal permettant de ne pas modéliser les partiels.

Cet ordre est donné par la relation [9] :

$$p \leq \frac{f_e}{2f_0} \tag{3.7}$$

où f_e est la fréquence d'échantillonnage du signal, et f_0 est la fréquence fondamentale du signal considéré comme harmonique ou quasi-harmonique.

3.2.2 La « True Envelope »

Le principal inconvénient du cepstre est que l'estimation a tendance a sous-estimer les résonances du filtre (un exemple de cepstre est représenté en ligne bleue sur la figure 3.4).

Deux méthodes ont été présentées pour pallier aux limitations de la représentation cepstrale : le cepstre discret [24] et la « True Envelope » [25].

La méthode du cepstre discret consiste à limiter le cepstre aux partiels du signal sonore. La méthode nécessite en outre l'estimation des partiels du signal sonore.

La méthode de la « True Envelope » est une méthode itérative de l'estimation de l'enveloppe spectrale [25, 23] qui permet de prendre en compte implicitement l'harmonicité du signal, c'est-à-dire qu'elle ne nécessite pas l'estimation préalable des paramètres précis des partiels. Par ailleurs, la représentation peut s'adapter théoriquement au signal sonore à partir de l'estimation de la fréquence fondamentale du signal (F0).

Soit $S(z)$ le spectre discret de K points fréquentiels d'une trame temporelle, et soit $V_i(z)$ la représentation spectrale donnée par les coefficients cepstraux à l'itération i, c'est-à-dire la transformée de Fourier des p premiers coefficients cepstraux :

$$V_i(z) = c(0) + 2\sum_{k=1}^{K} c(k)\cos(2\pi f k) \tag{3.8}$$

La méthode d'estimation fonctionne itérativement de la manière suivante :

1 On pose $A_0(k) = \log(|X(k)|)$ et $V_0(k) = -\infty, \forall k \in [1, \ldots, K]$.

2 L'amplitude du spectre « cible », à l'itération i, est :

$$A_i(k) = \max(A_{i-1}(k), V_{i-1}(k)), \forall k \in [1, \ldots, K] \tag{3.9}$$

3 Les coefficients cepstraux du spectre $A_i(k)$, et par conséquence la nouvelle représentation spectrale de l'enveloppe $V_i(k)$ est calculée.

Les étapes 2 et 3 sont répétées jusqu'à ce que qu'un critère d'arrêt soit vérifié. Le critère d'arrêt est par exemple la distance θ entre l'enveloppe estimée V_i et les points fréquentiels du spectre dont l'amplitude est supérieure à cette enveloppe :

$$A_i(k) - V_i(k) \leq \theta, \forall k \in [1, \ldots, K] \tag{3.10}$$

Une exemple typique de valeur de θ est celle qui correspond à 2 dB.

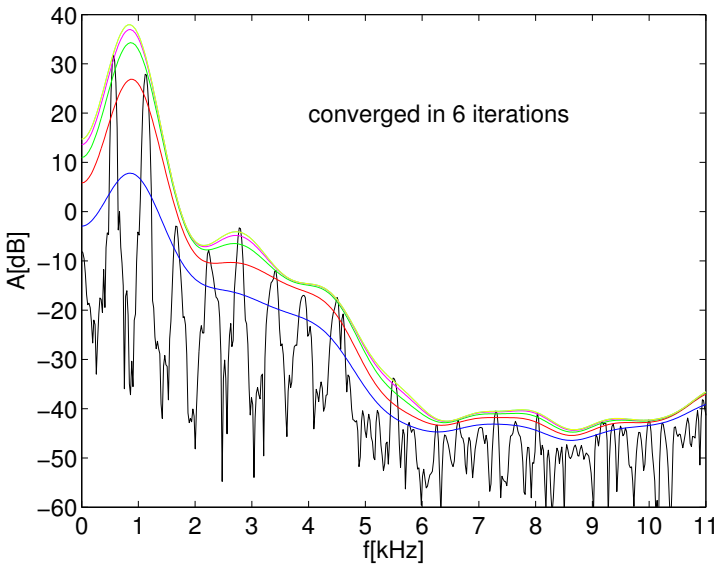

FIGURE 3.4 – Estimation de l'enveloppe par la méthode True Enveloppe pour 6 itérations [30]

Dans la suite de notre travail, nous avons choisi d'utiliser l'estimation de l'enveloppe spectrale par la méthode True Envelope. La figure 3.4 montre l'avantage de la méthode True Envelope : l'estimation du filtre du conduit vocal converge rapidement vers une solution adéquate. En bleu, le filtre estimé par le cepstre "classqiue", et en vert le filtre estimé par la méthode True Envelope après quelques itérations.

3.2.3 Sources de variabilité de l'enveloppe spectrale

L'enveloppe spectrale est censée représenter uniquement l'information des résonances du conduit vocal, et donc des phonèmes. Néanmoins, il existe un grand nombre de sources de variabilité pouvant affecter les caractéristiques de l'enveloppe spectrale. Dans le cadre de la voix chantée, nous pouvons lister les facteurs principaux suivants :

1 la physiologie du chanteur (identité, taille, sexe) [12]. Par exemple, les femmes ont un conduit vocal plus petit que ceux des hommes, et par conséquent des modes de résonances plus élevés en fréquences.

2 le style de chant. Par exemple : chant « pop » vs. chant lyrique, et l'apparition du "formant du chanteur" (3000 - 4000Hz) caractéristique du chant lyrique. [35, 37]

3 la fréquence fondamentale [10] [30]. Dans le modèle source/filtre, les résonances du conduit vocal sont supposées être indépendantes de la fréquence d'oscillation des cordes vocales. Il s'agit seulement d'une approximation : en réalité, la fréquence d'oscillation des cordes vocales entraîne une modification de la configuration de l'appareil phonatoire, et donc des résonances du conduit vocal. Un exemple extrême sont les modes de phonations en voix chantée : voix de poitrine, voix de tête qui correspondent à des configurations très spécifiques de l'appareil phonatoire [16].

Les sources de variabilité 1 et 2 viennent du caractère du locuteur et ce caractère reste invariant pendant une chanson. Donc, la solution de ces deux problèmes est d'adapter à ce caractère de locuteur en temps réel. Les méthodes courantes pour adapter le caractère de locuteur sont : MAP (Maximum a posteriori) [21] et MLLR (Maximum likelihood linear regression) [18] cf. 3.3.4.

3.3 Intégration des voyelles dans le système d'alignement

3.3.1 Partition

On change le format de partition d'Antescofo pour y ajouter l'information de voyelle. La première ligne est le tempo du texte « beat per minute ». Les événements musicaux sont à partir de la deuxième ligne. « NOTE » est l'étiquette d'événement ; la deuxième colonne est la hauteur du son en Midicent ; la troisième colonne est la durée ; la dernière colonne est l'étiquette de voyelle en XSAMPA.

Par exemple, « NOTE F#4 1/2 i » désigne un événement « NOTE » de la hauteur de F#4 qui dure 1/2 « beat » et cet événement est une voyelle « i ». Le silence dans le partition est marqué par la hauteur 0. Au début dans la partition, on peut ajouter un silence artificiel.

BPM	60		
NOTE	0	2	
NOTE	F#4	1/2	@
NOTE	G4	1/2	i
NOTE	F#4	1/2	@
NOTE	F#4	1/2	a
NOTE	E4	1	i

TABLE 3.3 – Un extrait de la partition de « Petite Marie »

3.3.2 L'apprentissage du gabarit de voyelle

Un ensemble des gabarits est construit en moyennant les True Envelopes de voyelle dans la base de données d'apprentissage. Par exemple, si l'on veut construire le gabarit de la voyelle j, on cherche l'ensemble de réalisations de cette voyelle dans une base de données, et puis on calcule les enveloppes spectrales te^j sur les spectrogrammes des réalisations des voyelles. Supposons que le nombre de trame de la voyelle j est n_j. Le gabarit d'une voyelle j est alors calculé comme la

moyenne de l'enveloppe spectrale estimée sur l'ensemble des trames correspondant à la voyelle :

$$V_j = \frac{te_1^j + te_2^j + \ldots + te_{n_j}^j}{n_j} \tag{3.11}$$

La figure 3.5 montre un ensemble des gabarits en moyennant dix réalisations de chacune de voyelle.

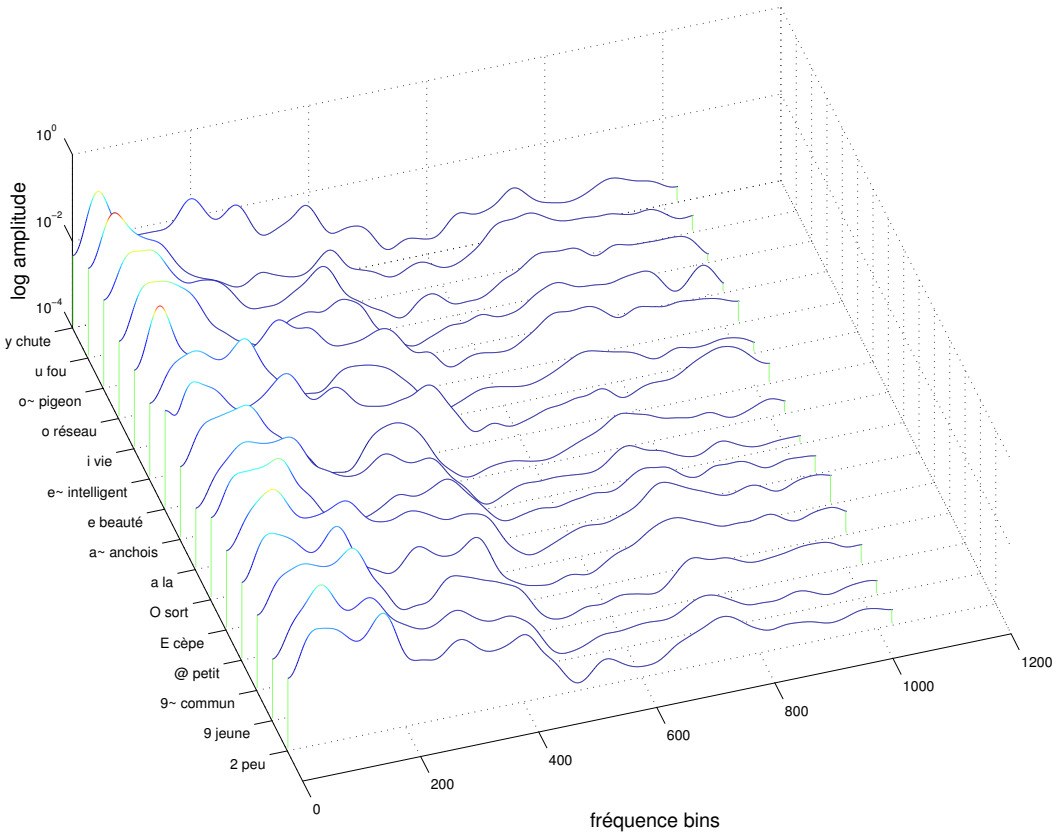

FIGURE 3.5 – Ensemble des gabarits de voyelle calculés

3.3.3 Calcul de la probabilité d'observation

Pour le gabarit de hauteur, la probabilité d'observation est calculée comme la KL–divergence entre le gabarit de hauteur et le spectrogramme. Pour le gabarit de voyelle, la probabilité d'observation est calculée de manière similaire : la KL-divergence entre le gabarit de voyelle et la « True Envelope » du spectrogramme.

La figure 3.6 montre un exemple de calcul de la probabilité d'observation à partir du gabarit de voyelle V_j et la « True Envelope » du spectrogramme X_t à l'instant t d'un enregistrement, notons T_t. Une fois que l'on a les informations de voyelle de la partition, on va chercher leurs gabarits de voyelle correspondants V_j dans l'ensemble des gabarits. Et puis, on utilise la formule de l'exponentielle de KL-divergence 3.12 pour calculer la probabilité d'observation de chaque voyelle.

$$p_V(x_t|s_j) = \exp[-\beta D(V_j||T_t)] \tag{3.12}$$

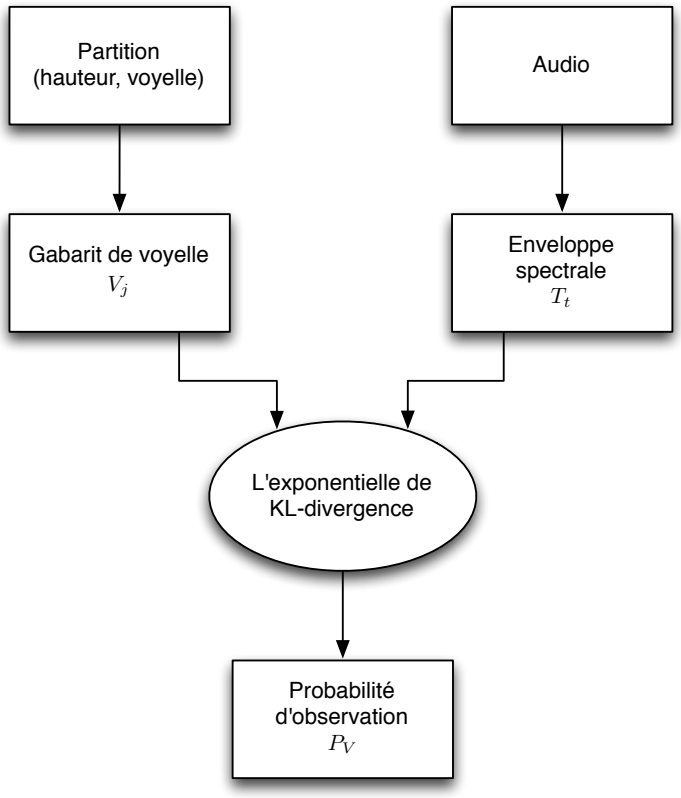

FIGURE 3.6 – Processus du calcul de la probabilité d'observation du gabarit voyelle

3.3.4 Adaptation du gabarit par l'estimation du MAP

Dans la section 3.2.3, nous avons listé trois facteurs principaux de variabilité de l'enveloppe spectrale. Nous présentons dans cette section une adaptation des gabarits de voyelle à de nouveaux enregistrements et aux sources de variabilité sus-mentionnées de l'enveloppe spectrale avec la méthode du "Maximum A Posteriori" (MAP) [21].

Le processus de l'adaptation du gabarit de voyelle est montré sur la figure 3.7. Les événements ayant la même information de voyelle mais différentes hauteurs possèdent ses propres gabarits de voyelle. Lorsqu'un événement est inféré, le gabarit de voyelle correspondant peut être repéré selon non seulement l'information de voyelle mais aussi celle de hauteur de cet événement.

L'adaptation du gabarit de voyelle est effectuée est effectuée "en ligne" lors de l'alignement.. Donc, l'utilisation du seuil de confiance pour préserver les données d'adaptation est importante. Le score de confiance utilisé dans notre travail est la probabilité de l'observation (enveloppe spectrale) x_t généré par l'état s_j, donc la probabilité d'observation $P_V(x_t|s_j)$. Les observations avec les scores de confiance au-dessus/au-dessous du seuil ρ sont marquées « retenue »/« jetée ». L'adaptation ne réalise qu'avec les observations « retenues ».

Selon l'équation 3.11, le gabarit de voyelle V_j est la moyenne d'une gaussienne de n-dimensions, où n est le nombre de points fréquentiels du gabarit. Supposons que les observations « retenues » pour l'adaptation sont $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, le gabarit de voyelle adapté $\tilde{V}_j$ est

$$\tilde{V}_j = \alpha E(\mathbf{x}) + (1 - \alpha)V_j, \tag{3.13}$$

où $\alpha = T/(T + r)$ est le coefficient d'adaptation avec le facteur de pertinence r, $E(\mathbf{x})$ est la

statistique exhaustive (la moyenne) des observations X

$$E(x) = \frac{\sum_{t=1}^{T} \mathbf{x}_t}{T}. \tag{3.14}$$

L'équation 3.13 est la forme simplifiée de l'estimation du MAP (maximum a posteriori) [21] dans le cas de gaussienne uni-modale. On constate que $\tilde{V}_j$ est simplement la somme pondérée entre le gabarit de voyelle précédent V_j et les observations « retenues » X.

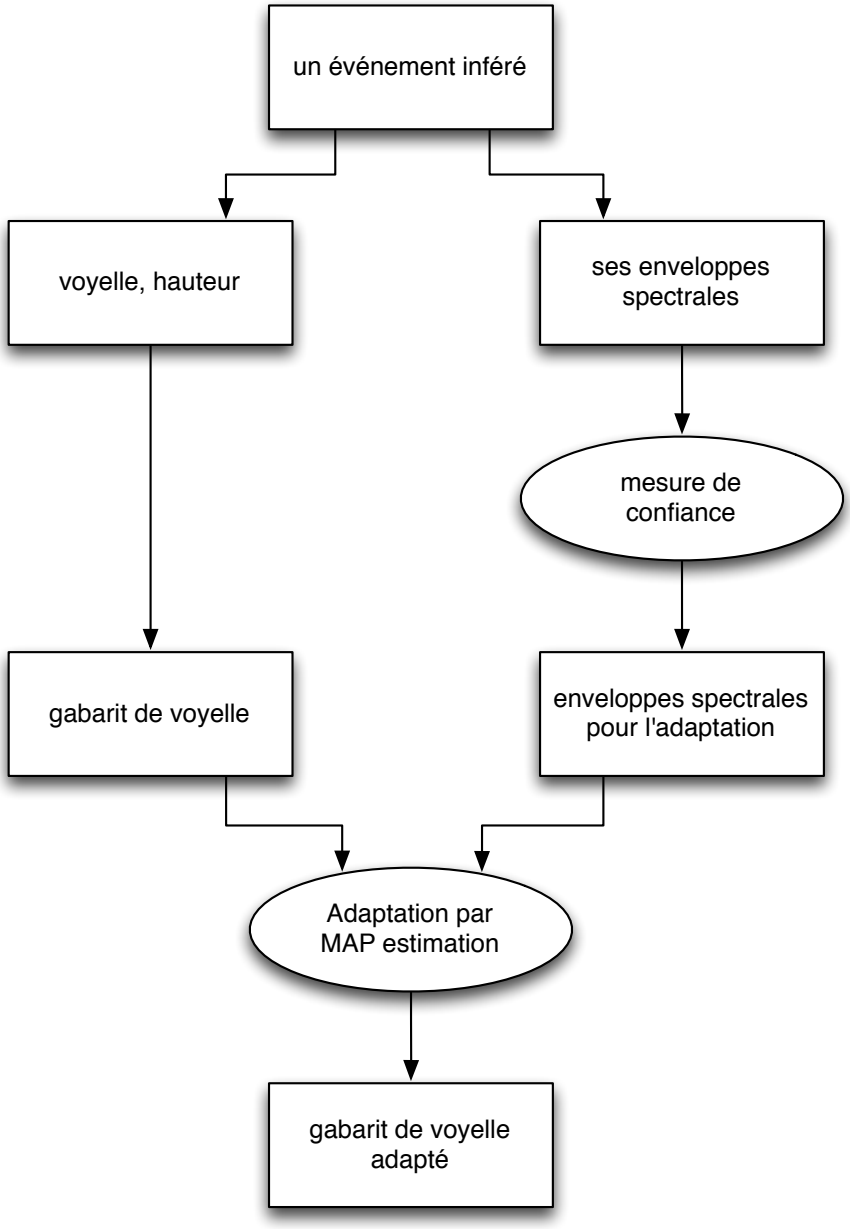

FIGURE 3.7 – Le processus de l'adaptation d'un gabarit de voyelle

3.4 Expérience préliminaire : preuve de concept

Pour mettre en évidence la capacité de « True Envelope » à distinguer les différentes voyelles françaises, on effectue une expérience préliminaire. La base de données de cette expérience est un ensemble de diphones de hauteur fixe. Pour chaque voyelle, le signal de voix est découpé en trames de spectre du spectrogramme, puis l'enveloppe spectrale de chaque trame est estimé par la méthode « True Enveloppe ». La KL divergence symétrique est utilisée comme une métrique pour calculer la distance entre chaque pair de trames, et le résultat global de distance est représenté sur une matrice de dissimilarité. Finalement, on projette toutes les trames sur une espace de MDS (en anglais Multidimensional scaling, en français Positionnement multidimensionnel) selon leur distance. Nous obtenons ainsi une visualisation dans un espace à faible dimension de la distribution des voyelles.

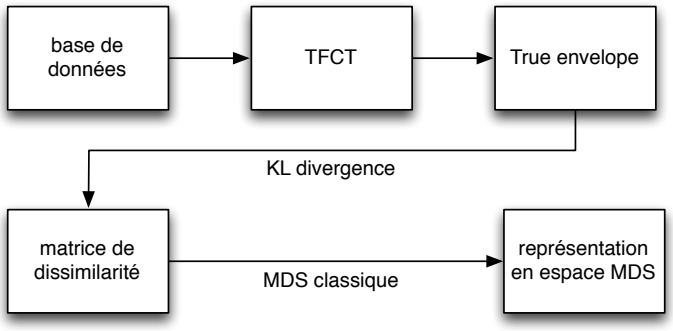

FIGURE 3.8 – Processus de l'expérience

3.4.1 Description de la base de données

La base de données contient les enregistrements d'un chanteur masculin (1149 fichiers audio, environ 40 minutes). Chaque enregistrement correspond à la réalisation d'un diphone à une hauteur fixée - 147 Hz (D3) pour l'ensemble des enregistrements. Les fichiers sont enregistrés dans des conditions optimales (studio professionnel), et format non compressé (wav), et codés en 48kHz/16bits.

La liste de voyelle utilisée pour l'expérience est montrée sur le tableau 3.4. D'abord, On collectionne les trames de trois réalisations pour chacune de voyelle. On concatène les trames de la même voyelle dans les unités, et après on concatène ces unités par ordre de gauche à droite du tableau 3.4.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
phonème	ø	œ̃	ə	ɑ	ɛ	ɔ	a	ã	e	ɛ̃	i	o	õ	u	y
XSAMPA	2	9∼	@	A	E	O	a	a∼	e	e∼	i	o	o∼	u	y
nombre de trame	512	473	415	422	417	852	402	507	344	529	484	874	524	482	186

TABLE 3.4 – Voyelles utilisée pour l'expérience.

3.4.2 KL-divergence symétrique

La KL divergence utilisée dans Antescofo est défini comme

$$D(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{3.15}$$

Cette formule n'est pas une métrique - par exemple, elle n'est pas symétrique : $D(P||Q) \neq D(Q||P)$. Par conséquent, la matrice de dissimilarité générée par cette formule n'est pas symétrique non plus. Cela pourra poser un problème lorsque l'on réduit la dimension par la méthode MDS (la section 3.4.5). Au lieu d'implémenter la formule (3.15), on propose la définition de la KL divergence symétrique :

$$D_{KL}(P||Q) = \frac{D(P||Q) + D(Q||P)}{2} \tag{3.16}$$

qui est symétrique et non négative. Cette quantité est parfois utilisée pour la sélection des descripteurs dans les problèmes de classification, où P et Q sont les pdfs (densité de probabilité) conditionnelles d'un descripteur dans deux classes différentes.

3.4.3 Extraction de l'envelope spectrale

Le calcul de spectrogramme est effectué sur les réalisations de voyelle. La taille de fenêtre est choisie à 30 ms, parce qu'elle devrait être supérieure à quatre fois de la période de la f_0, qui vaut $4/147 \approx 27$ms ; Le pas d'incrément est choisi à 5 ms. La fréquence de coupure de spectre est $f_e/4$, où $f_e = 48$kHz est la fréquence d'échantillonnage.

Après le calcul de spectrogramme, on obtient un « nuage » de trame de ces réalisations. Le calcul de True Envelope peut ensuite effectué sur ces trames, avec un ordre optimal..

3.4.4 Mesure de similarité entre les enveloppes spectrales

La matrice de dissimilarité est calculée entre chaque paire de trame en utilisant la métrique « KL-divergence symétrique » décrite en équation 3.16.

La figure 3.9 montre cette matrice. La couleur bleue signifie que la distance est petite et que les True Envelopes sont similaires, la couleur rouge signifie que la distance est grande et que les True Enveloppes sont dissimilaires. Tout d'abord, on remarque que la distance intra-voyelles est petite par rapport à la distance inter-voyelles (carrés bleus sur la diagonale de la matrice de similarité). Nous voyons ainsi que les true enveloppe des trames d'une même voyelle se regroupent en classes similaires. Par ailleurs, on peut également noter que les trames de voyelles similaires sont similaires acoustiquement. Par exemple : /2/, /9∼/, et /@/ ; /A/, /a/, et /a∼/ ; /O/, /o/ et /o∼/.

3.4.5 Positionnement multidimensionnel

La méthode MDS [34] projette les données issues d'un espace à haute dimension dans un espace euclidiens à faible dimension et est couramment utilisé pour visualiser des données dans un espace à haute dimension (typiquement, dimension = 2 ou 3). Dans notre cas, nous allons utiliser cette méthode pour visualiser la distribution des trames de voyelles à partir de la matrice de dissimilarité calculée dans la section précédente.

Pratiquement, la MDS prend une matrice de distance entre des points dans un espace de dimension N (matrice de similarité/dissimilarité) , et retourne une matrice de coordonnées dans un espace de dimension P ($P < N$, typiquement : $P = 2, 3$). [34].

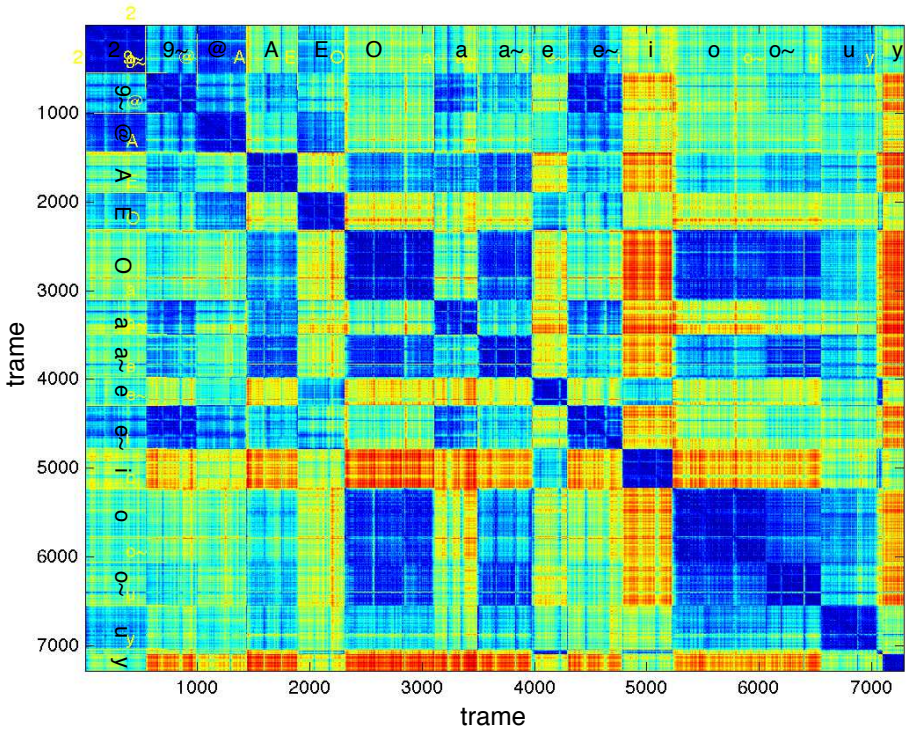

FIGURE 3.9 – La matrice de dissimilarité

On représente les moyennes des nuages de voyelle. En observant la figure 3.10, on remarque qualitativement une correspondance entre la répartition des voyelles en espace MDS et le triangle vocalique si on renverse l'abscisse « dimension 1 » et si on incline un peu le triangle vocalique vers la gauche.

3.5 Conclusion

Dans cette section, nous nous sommes intéressés à l'alignement du texte et de la voix chantée. Pour cela, le texte est converti en une séquence de voyelles qui représentent les phonèmes dominants pour l'alignement de la voix chantée. Pour chaque voyelle, nous construisons un gabarit de voyelle à partir d'une analyse de l'enveloppe spectrale du son à court-terme. Ainsi, l'information de voyelle peut être intégrée directement dans le système de suivi de partition AnteScofo.

Nous avons réalisé une expérience préliminaire pour évaluer l'efficacité de l'enveloppe spectrale à représenter les voyelles dans l'espace acoustique. Nous avons obtenu des résultats extrêmement encourageants : la distribution de l'enveloppe spectrale des voyelles reproduit le triangle vocalique théorique de la voix [17].

On peut donc raisonnablement espérer que les gabarits de voyelles pourront être efficacement utilisés pour l'alignement du texte et de la voix chantée. Nous présentons dans la suite l'intégration du texte dans le cadre du suivi de partition de la voix chantée.

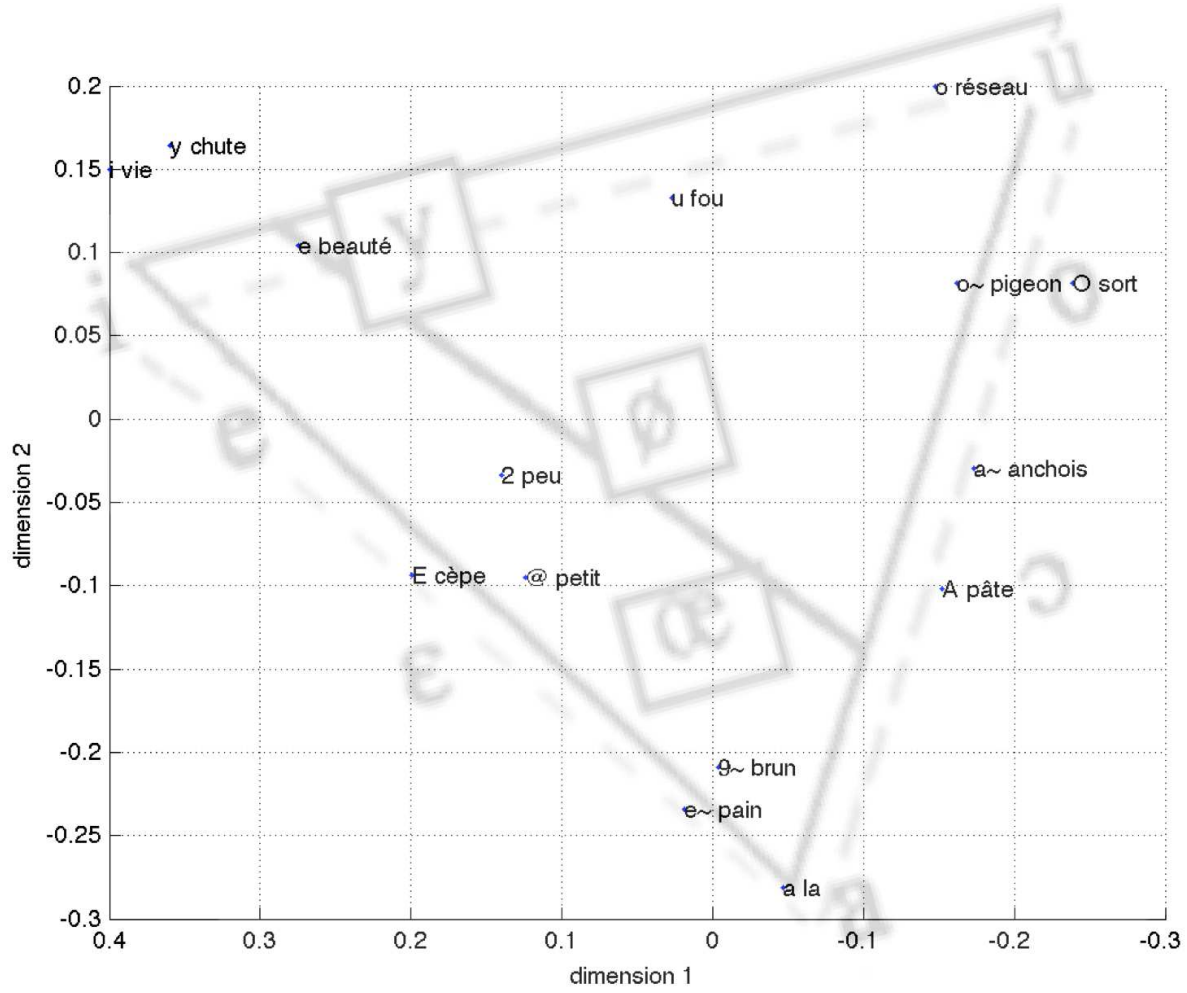

FIGURE 3.10 – Représentation des centroïdes de voyelle en espace MDS et Triangle vocalique. Le triangle en gris : le triangle vocalique, les points en noir : les centroïdes de chaque voyelle

Chapitre 4

Fusion d'information hauteur/voyelle

Dans les chapitres précédents, on a présenté les gabarits de hauteur et d voyelle. On peut prévoir qu'ils ont leurs propres avantages dans l'application d'alignement : les gabarits de hauteur marcheraient bien lorsque les hauteurs des événements se varient consécutivement. Par contre, comme les états de la même hauteur possèdent la même probabilité d'observation, le système d'inférence n'aura pas de caractéristique à décider le moment où saute-t-il vers l'état suivant. En conséquence, les gabarits de hauteur marcheraient moins bien quand le chanteur chante les mêmes hauteurs consécutives.

Du côté des gabarits de voyelle, ils marcheraient mieux pour les événements qui varient consécutivement au niveau de la voyelle, mais pour la même raison, ils ne marcheraient pas quand le chanteur chante les mêmes voyelles consécutives. Donc, dans ce chapitre, on propose deux stratégies de fusion des informations hauteur et voyelle pour rendre le système d'alignement plus robuste.

On propose dans les sections 4.1 et 4.3 deux stratégies de fusion. 1) fusion dans l'espace acoustique, ce que l'on appelle « **early fusion** ». 2) fusion dans l'espace statistique, ce que l'on appelle « **late fusion** ».

4.1 « Early fusion »

4.1.1 Fusion des gabarits hauteur/voyelle

L'idée de la fusion des gabarits est inspirée par le modèle de source/filtre. Du point de vue de Traitement du Signal, la source d'excitation d'une voyelle est une vibration périodique représentée par le son harmonique - le gabarit de hauteur au spectre plat. Le filtre est décrit par l'enveloppe spectrale - le gabarit de voyelle.

$$S_j^{\text{fusion}} = W_j . * V_j \tag{4.1}$$

Pour obtenir le gabarit d'Early fusion, on multiplie le gabarit de hauteur et de voyelle point par point (figure 4.1, eq 4.1, où $.*$ signifie la multiplication point par point). On se concentre sur la basse fréquence, puisque les hautes fréquences sont davantage bruitées et donc plus variables, comme elles concentrent moins de puissance. D'un autre côté, le nombre d'harmoniques du gabarit de hauteur doit recouvrir le 2ème formant. Les fréquences de 2ème formant de la voyelle du français sont inférieures à 3000 Hz [29], donc, on choisit le nombre d'harmoniques pour qu'il atteigne 3000 Hz.

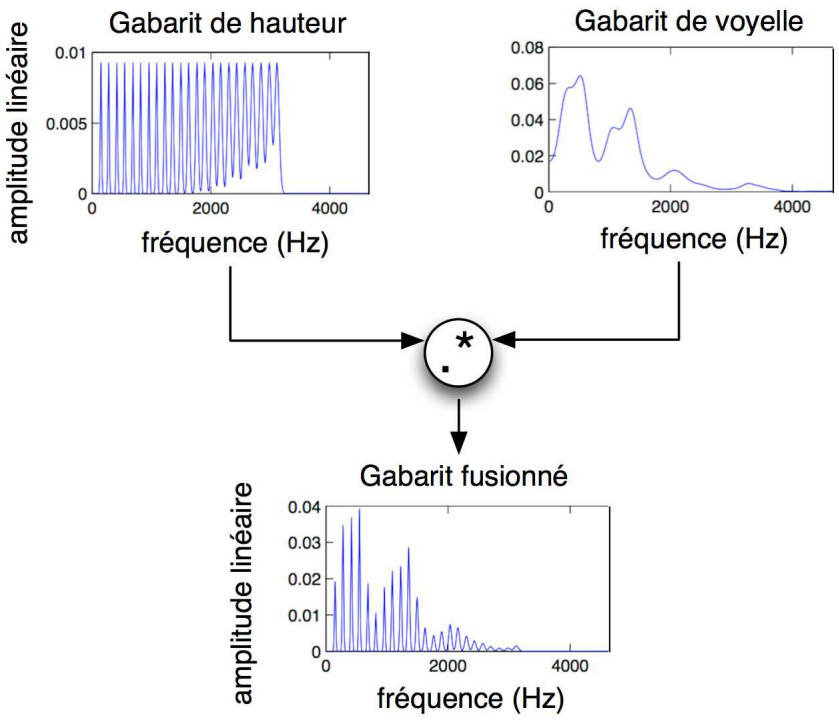

FIGURE 4.1 – Fusion des gabarits de hauteur et de voyelle

4.1.2 Calcul de la probabilité d'observation

Lorsque le gabarit « Early fusion » est généré, la probabilité d'observation s'obtient en calculant l'exponentielle de KL-divergence entre les gabarits et le spectrogramme de l'enregistrement à aligner, de la même manière que dans le système de base (la section 2.2.2).

La figure 4.2 montre un exemple de calcul de la probabilité d'observation à partir d'un gabarit fusionné de gabarit de voyelle /@/ et de celui de hauteur F#4. On conduit cette fusion dans l'espace acoustique. On obtient la probabilité d'observation dans l'espace statistique après le calcul de l'exponentielle de KL-divergence.

Un processus de calcul de la probabilité d'observation à partir de la partition et le ficher audio est montré sur la figure 4.3. On extrait les informations de hauteur et de voyelle en parallèle dans la partition, et puis on cherche leurs gabarits correspondants dans la base de données de gabarit. La fusion des gabarits de hauteur et de voyelle est ensuite effectuée. La probabilité d'observation de chaque état de la partition est finalement calculée.

4.2 Conversion de l'échelle de fréquence

À cause du fonctionnement de la cochlée, le système auditif humaine offre une résolution plus haute en basse fréquence qu'en haute fréquence. La cochlée peut ainsi être modélisée par un banc de filtres auditifs. Pour la raison de trouver l'échelle de fréquence le meilleur efficace en matière de l'alignement, on va tester la performance d'alignement en convertissant l'échelle de fréquence du spectrogramme observé ainsi que des gabarits de hauteur.

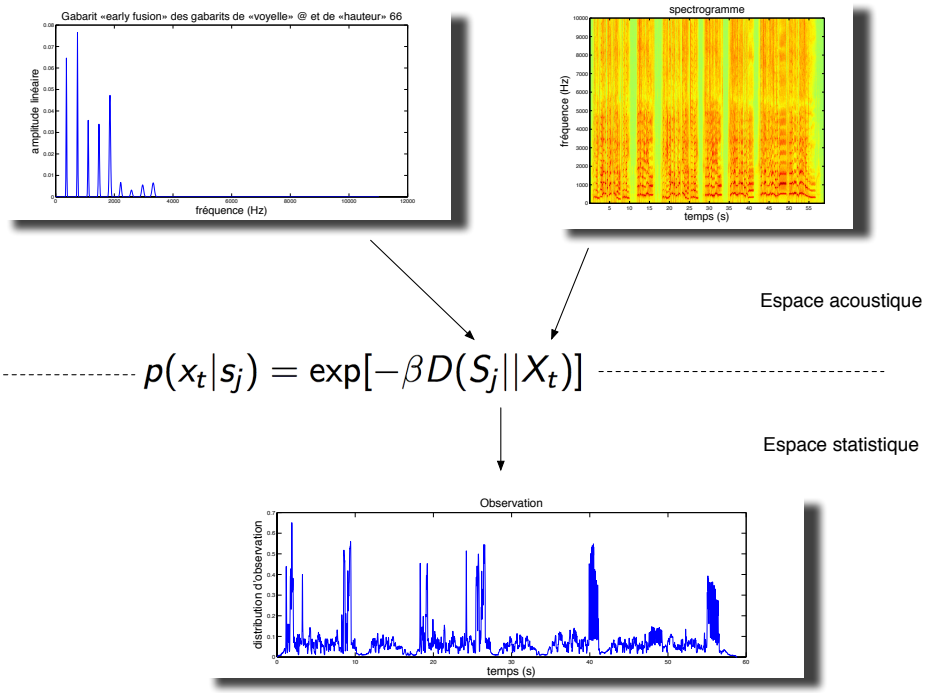

FIGURE 4.2 – Probabilité d'observation du gabarit « early fusion » à l'échelle de fréquence linéaire. La hauteur 66 en Midicents est F#4.

4.2.1 Échelles de Mel et de Bark

Les échelles de fréquence non linéaires basées sur la perception humaine les plus utilisées sont les échelles de Mel et de Bark.

L'échelle de Mel décrit l'échelle de la perception humaine sur les octaves. Par exemple, un son ayant une fréquence de 3428 Hz (2000 mels) est donc perçu comme étant deux fois plus aigu qu'un son à 1000 Hz (1000 mels) [15]. Bark est une échelle déduite par Zwicker [19] proportionnelle à la largeur des bandes critiques qui décrient les capacités de résolution fréquentielle de l'oreille, liées à la perception des hauteurs.

Les équations de conversion de la fréquence Hz aux échelles de Mel [1] et de Bark [36] sont :

$$Mel = 1127 \cdot \ln(1 + \frac{f}{700}) \tag{4.2}$$

$$Bark = 13 \arctan(0.00076 f) + 3.5 \arctan((f/7500)^2) \tag{4.3}$$

L'interpolation linéaire est faite après le conversion pour augmenter la résolution de basse fréquence et diminuer celle de haute fréquence.

4.2.2 Transformée à Q constant

La transformée à Q constant (CQT) se réfère ici à une représentation temps-fréquence obtenue par un banc de filtres dont les fréquences centrales des fréquences bandes sont géométriquement espacées et leurs facteurs de qualité sont tous égal [8]. Cela signifie que la résolution de fréquence est meilleure pour les basses fréquences et la résolution temporelle est meilleure pour les hautes fréquences. Les fréquences centrales f_k des fréquences bins de la transformée à Q constant (CQT) se forment à la formule [2]

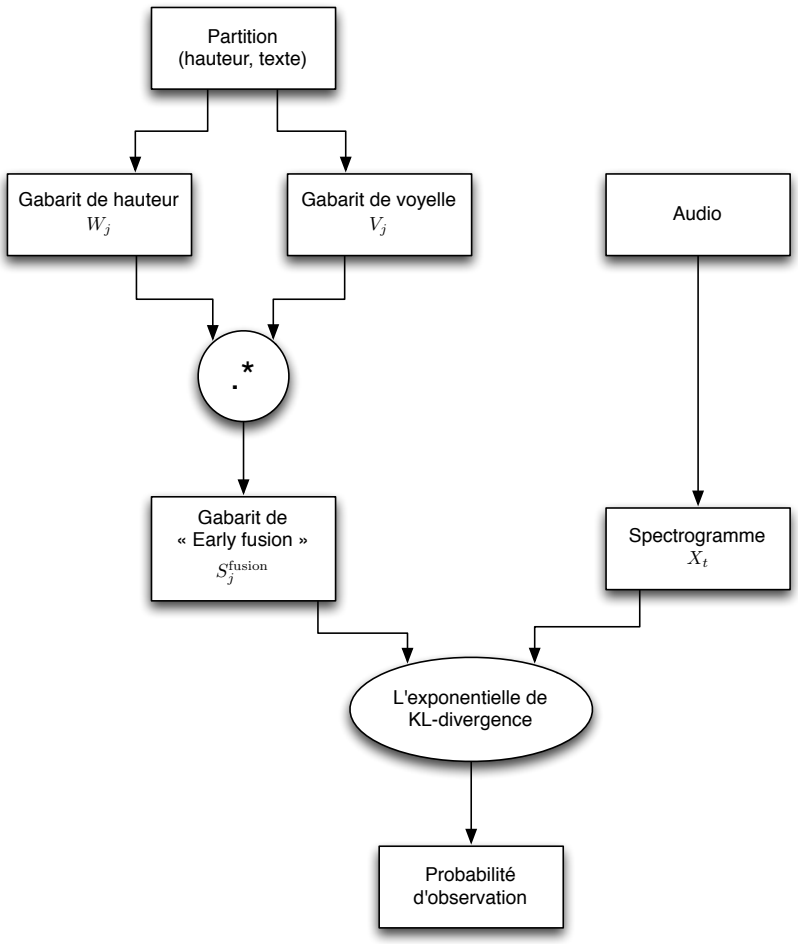

FIGURE 4.3 – Processus du calcul de la probabilité d'observation pour le gabarit « Early fusion »

$$f_k = f_0 2^{k/b} \quad (k = 0, \ldots) \tag{4.4}$$

Le CQT est bien motivé du point de vue musical. Selon la formule 4.4, si f_0 vaut la fréquence d'une note musicale du tempérament égal à 12 demi-tons, b vaut 12, f_k tombe ainsi sur les fréquences des notes au tempérament égal à 12 demi-tons.

Dans la phase d'implémentation, on choisit les paramètres $f_0 = 65.4$Hz qui correspond au note C2 et la fréquence maximum 8372 Hz (C9) qui recouvre largement le 2ème formant de toutes voyelles. $b = 12, 24, 48$ qui signifie que l'on divise une octave au 12, 24, 48 bancs de filtres au tempérament égal. En utilisant plus de bancs filtres signifie que l'on a plus de résolution fréquentielle.

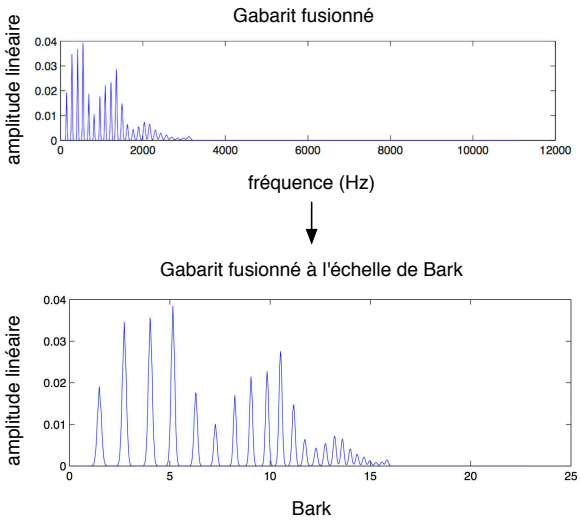

FIGURE 4.4 – Un exemple de la conversion d'échelle de fréquence pour un gabarit de « Early fusion »

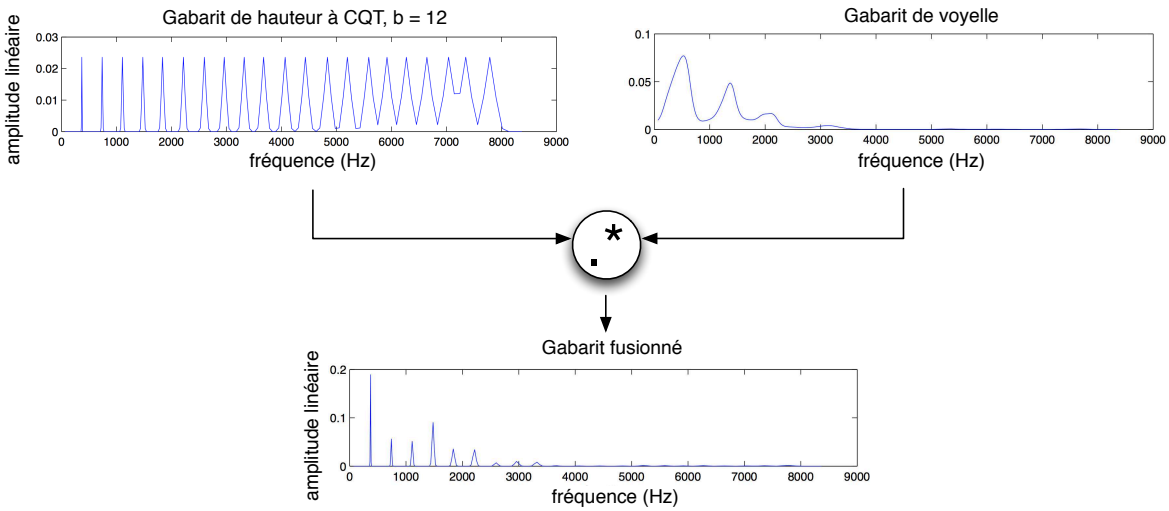

FIGURE 4.5 – Fusion des gabarits à CQT

4.3 « Late fusion »

« Late fusion » signifie la fusion des probabilités dans l'espace statistique au lieu de fusionner acoustiquement les gabarits.

$$p_{fusion} = \frac{p_W + p_V}{2} \tag{4.5}$$

Les probabilités d'observation du gabarit de hauteur p_W et du gabarit de voyelle p_V sont calculées en parallèle. La fusion des probabilités d'observation est la moyenne arithmétique de ces deux probabilités. La figure 4.6 montre ce processus :

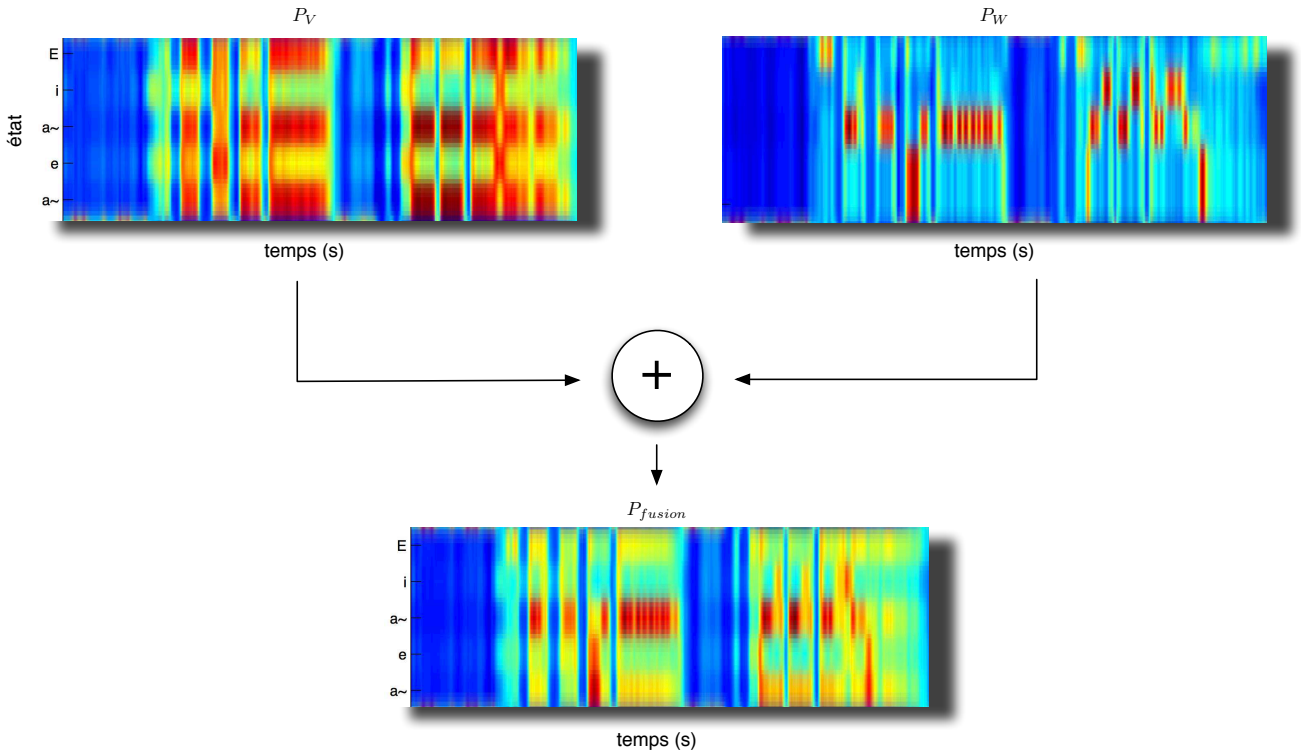

FIGURE 4.6 – Stratégie de « late fusion », la couleur rouge signfie que la probabilté est grande

Chapitre 5

Évaluation

L'évaluation donne une indication de la qualité d'un algorithme d'alignement et permet de comparer les différentes méthodes, implémentations, paramètres, etc [11]. Une fois les stratégies de fusion intégrées dans le système, l'évaluation de ce système d'alignement peut être effectuée. On décrit la procédure de l'évaluation et la comparaison des stratégies dans les quatre sections suivantes.

Dans la section 5.1, on présente les bases de données utilisées pour le système. Les bases de données sont celles d'apprentissage - l'ensemble des gabarits de voyelle, celles d'évaluation - les enregistrements de longs extraits, les partitions de ces extraits et les annotations. Dans la section 5.2, on présente le protocole d'évaluation qui donne une mesure objective de la performance des stratégies. Dans la section 5.3, on présente l'apprentissage du gabarit, les expériences effectuées avec les deux stratégies. Dans la section 5.4, on compare les résultats de l'évaluation des stratégies.

5.1 Description de la base de données

5.1.1 Gabarits de voyelle

Deux ensembles de gabarits de voyelle sont construits par la méthode d'apprentissage de la section 3.3.2 avec deux bases de données d'apprentissage correspondantes contenant les mêmes phrases conçues pour la synthèse du chant.

#	**Locuteur**	**Hauteur (Hz)**	nombre de voyelle
1	femme	392	15
2	homme	192	15

TABLE 5.1 – Deux ensembles des gabarits de voyelle utilisés dans l'évaluation

5.1.2 Les longs extraits d'évaluation

Les extraits audio de longes durées contiennent 22 minutes de paires audio/partition. Le format de ficher audio est AIFF-C, la fréquence d'échantillonnage et le nombre de bits par échantillon sont 44.1 kHz et 16 bits. Tous les enregistrements sont à tempo constant et la partition d'alignement indique ce tempo.

L'ensemble des longs extraits sont enregistrés par deux locuteurs homme et femme 5.1. Huit extraits de chanson monophonique sont choisis. L'un des extraits, "Petite Marie", a été enregistré deux fois dans deux tonalités différentes par le locuteur femme. Les durées et les nombres de

voyelle des longs extraits sont montrés dans les tableau 5.2.

		femme		homme	
#	**Nom d'enregistrement**	**Durée (s)**	**Nombre de voyelle**	**Durée (s)**	**Nombre de voyelle**
1	Serge Gainsbourg - La Javanaise	2 :21	180	1 :13	106
2	Zazie - Je suis un homme	1 :20	138	1 :17	182
3	Francis Cabrel - Petite Marie	0 :59	117	1 :00	127
4	Francis Cabrel - Petite Marie (en baisse de 4 demi-tons)	1 :16	156	–	–
5	Charles Aznavour - Emmenez-moi	2 :09	319	1 :00	164
6	Céline Dion - J'irai où tu iras	1 :51	329	–	–
7	Jean-Jacques Goldman - Envole-moi	1 :46	230	1 :08	160
8	Axelle Red - Sensualité	2 :05	221	0 :51	114
9	Johnny Hallyday - Toute la musique que j'aime	0 :53	102	0 :53	114
	somme	14 :40	1792	7 :22	967

TABLE 5.2 – Les durées et les nombres de voyelle des enregistrements

5.1.3 Les partitions

Les partitions symboliques sont converties des fichiers MIDI téléchargés sur internet et ont été manuellement vérifiées par rapport à la version originale des chansons.

Les événements successifs de la partition sont séparés par les hauteurs différentes ou les voyelles différentes. Certaines partitions comportes des mélismes, c'est-à-dire des voyelles qui sont tenues sur plusieurs hauteurs successives.

Toutefois, les performances ne respectent pas toujours la partition (parfois les hauteurs, plus souvent le rythme). On peut appeler cela des "déviations d'interprétation" par rapport à la partition.

5.1.4 Annotation manuelle

L'annotation (en anglais ground-truth) qui s'agit la référence d'alignement, est l'indication de l'instant d'attaque de chaque note dans la partition. Elle est décidée par écouter ou regarder le spectrogramme. On utilise le logiciel « Audiosculpt » pour faire ce tâche en mettant le « marker » à l'attaque de chaque voyelle.

5.2 Méthode d'évaluation

On cite la méthode d'évaluation de l'article [11] pour évaluer nos stratégies. Dans cet article, on a défini 3 *mesures d'événement basiques* (en anglais *measures basics*) pour l'alignement en temps-différé.

On défini **Error** $e_i = |t_i^e - t_i^r|$ comme le module du laps du temps entre la position du temps d'alignement décodé t_i^e et celle correspondant dans la vérité-terrain (ground truth) t_i^r. **Missed notes** sont les événements qui ne sont pas décodés, mais qui existent dans l'annotation. **Misaligned** sont les événements qui possèdent les erreurs $e > 300$ms.

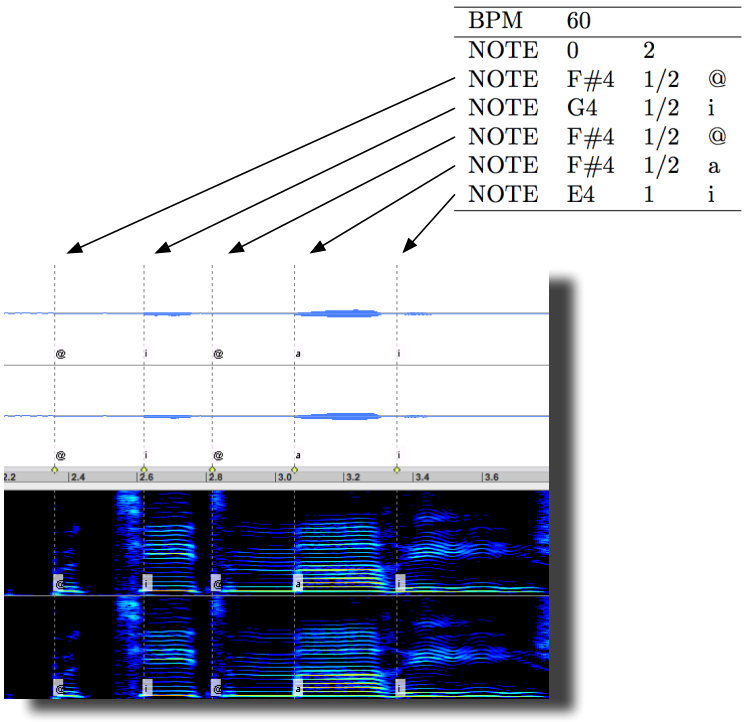

FIGURE 5.1 – La partition et l'annotation

Étant donné les mesures d'événement fondamentales, les *indicateurs d'évaluation* qui caractérisent la qualité de l'alignement sont : **Averge error** est défini comme la moyenne de tous les « errors » pour tous les non-« misaligned » notes. **Max. error** est défini comme la valeur maximum des « errors » [7](On tient compte de tous les notes qui incluent les « misaligned » notes). **Miss rate** et le **Misalign rate** sont également définis comme le rapport entre « Missed note » et le nombre des événements dans le morceau ainsi que celui entre « Misaligned » et le nombre des événements.

L'estimateur va lire le résultat de l'alignement t_i^e et la vérité-terrain t_i^r détaillé dans la section 5.1. Donc, toutes les mesures d'événement et les indicateurs d'évaluation peuvent être calculés facilement par ses définitions. Finalement, on assemble les résultats d'alignement des 8 morceaux dans un seul tableau pour montre la performance des stratégies.

5.3 Description des expériences

Une fois la méthode d'évaluation est bien définie. On peut ensuite préparer les expériences pour comparer les différentes stratégies.

Nos expériences vont se fonder sur le système AnteScofo que l'on a présenté dans le premier chapitre. Les parties de construire le réseau HSMM, de l'inférence et du décodage restent la même. Le changement du processus est sur les étapes de construire la matrice d'observation (figure 5.2).

On continue à utiliser les mêmes paramètres du système AnteScofo pour calculer le spectrogramme du fichier son et du gabarit de voyelle : la longueur de la fenêtre 92 ms et le facteur de recouvrement 4.

1. Les expériences principales sont conçues pour tester la performance du système avec un ensemble des gabarits de voyelle et les deux stratégies de fusion. Les deux ensembles des

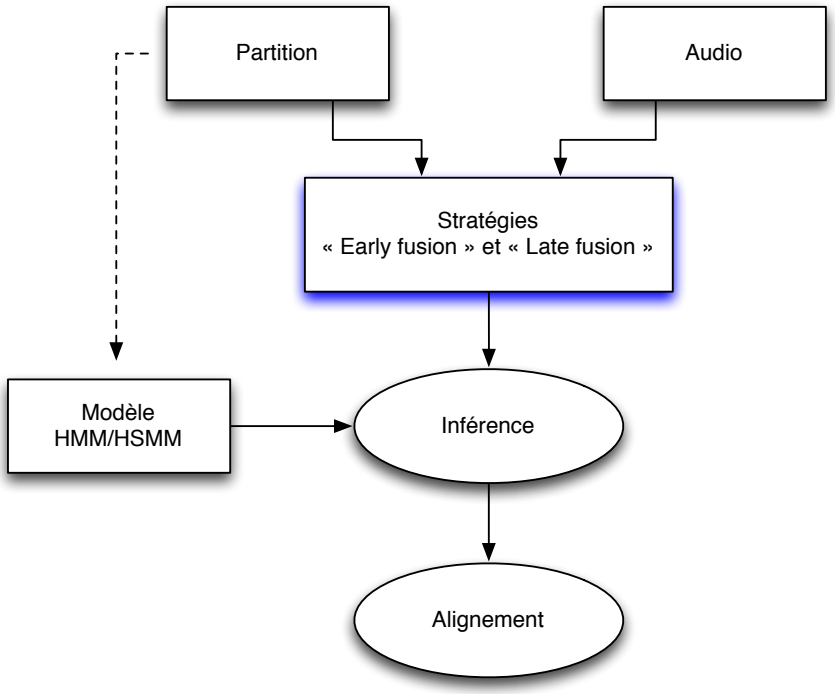

FIGURE 5.2 – Le processus du système d'expériences

gabarits de voyelle (tableau 5.1) sont utilisés pour leurs propres extraits d'évaluation. C'est-à-dire nous utilisons l'ensemble des gabarits femme pour évaluer les extraits femme, et ceux homme pour évaluer les extraits homme. Les gabarits sont à l'échelle de fréquence linéaire. Aucune méthode d'adaptation n'est utilisée.

- Hauteur : utiliser l'ensemble des gabarits de hauteur toute seule
- Voyelle : utiliser l'ensemble des gabarits de voyelle
- Early fusion : utiliser le gabarit de « Early fusion »
- Late fusion : effectuer « Late fusion » qui combine les matrices d'observation des expériences « Hauteur » et « Voyelle ».

2. Les expériences supplémentaires ne sont pas nos travaux principaux, elles sont conçues pour chercher les méthodes possibles qui pourraient améliorer la performance d'alignement.

- Hauteur : utiliser l'ensemble des gabarits de hauteur aux échelles de fréquence non linéaire - **Bark**, **Mel** et de la transformée à Q constant (**CQT**).
- Voyelle : utiliser l'ensemble des gabarits de voyelle 1) avec **l'adaptation de $\mathbf{f_0}$** 2) avec l'adaptation du gabarit par l'estimation du **MAP** 2) en échangeant les gabarits de voyelle, c'est-à-dire nous utilisons l'ensemble des gabarits femme pour évaluer les extraits homme, et ceux homme pour évaluer les extraits femme (**Gabarit échange**).
- Early fusion : utiliser le gabarit « Early fusion » aux échelles de fréquence non linéaire et de la transformée à Q constant.
- Late fusion : effectuer « Late fusion » en échangeant les gabarits de voyelle (**Gabarit échange**).

5.4 Comparasion des stratégies

5.4.1 Expériences principales : rôle du gabarit voyelle et fusion des gabarits

Expérience	Averge error (s)	Max error (s)	Missalign rate %	Miss rate %
Hauteur	0.0758	2.2719	7.91	2.69
Voyelle	0.0840	2.7137	7.41	3.63
Early fusion	0.0688	2.9853	7.94	4.12
Late fusion	0.0678	1.2957	4.02	0.91

TABLE 5.3 – Les résultats d'alignement globals.

La stratégie « Late fusion » possède la meilleure performance d'alignement. Elle améliore 3.89% de « Missalign rate » et 1.78% de « Miss rate » par rapport à l'ancien système de AnteScofo. La stratégie « Early fusion » n'améliore pas la performance du système.

Les expériences « Hauteur » et « Voyelle » montre que le système avec l'information de hauteur ou de voyelle toute seule ont les performances similaires. Cela est un peu contraire à notre attente, parce que les chansons d'évaluation contiennent de nombreuses phrases « mono hauteur » qui mettraient en évidence l'avantage de l'alignement voyelle/texte. Une explication est que nos gabarits de voyelle sont tous à la même hauteur, donc ils ne s'adaptent pas aux enveloppes spectrales de chant, et cela dégrade la performance d'alignement.

5.4.2 Expériences supplimentaires

Expérience	Averge error (s)	Max error (s)	Missalign rate %	Miss rate %
Hauteur	0.0758	2.2719	7.91	2.69
Bark	0.0781	16.8135	16.31	12.96
Mel	0.0778	7.5871	11.12	6.65
CQT b = 12	0.0716	7.6103	7.58	4.40
CQT b = 24	0.0741	7.6916	9.09	4.43
CQT b = 48	0.0752	7.6103	10.22	5.53

TABLE 5.4 – Les résultats d'alignement des expériences a supplémentaires.

L'échelle de fréquence linéaire possède la meilleure performance d'alignement en utilisant le gabarit de hauteur tout seul, avec « Missalign rate » 7.91% et « Miss rate » 2.69%.

Expérience	Averge error (s)	Max error (s)	Missalign rate %	Miss rate %
Voyelle	0.0840	2.7137	7.41	3.63
MAP	0.0830	2.7137	7.49	4.12
Gabarit échange	0.0922	10.8020	11.28	7.81

TABLE 5.5 – Les résultats d'alignement des expériences b supplémentaires.

L'adaptation du gabarit par l'estimation MAP ne donne pas le meilleur alignement. La raison a besoin de continuer à explorer, nous proposons deux explications possibles : 1) les données de l'adaptation ne sont pas suffisantes. Dans les extraits qui durent au plus 2 minutes, l'occurrence d'un événement possédant la même hauteur et la même voyelle est faible. Par conséquent, un gabarit de voyelle n'obtient pas beaucoup de chance pour l'adaptation. 2) La mesure de confiance n'est pas bien choisie et ce qui conduit à la perte de données d'adaptation ou à l'inclusion de mauvaises données.

La statistique « Gabarit échange » montre qu'en utilisant un ensemble des gabarits d'un locuteur différent dégrade la performance.

Expérience	**Averge error (s)**	**Max error (s)**	**Missalign rate %**	**Miss rate %**
Early fusion	0.0688	2.9853	7.94	4.12
Bark	0.0793	9.9672	15.41	18.91
Mel	0.0758	9.0575	14.03	12.14
CQT b = 12	0.0678	2.2846	8.25	6.39
CQT b = 24	0.0666	2.4365	8.51	4.81
CQT b = 48	0.0667	3.3019	7.96	4.35

TABLE 5.6 – Les résultats d'alignement des expériences c supplémentaires.

L'échelle de fréquence linéaire possède la meilleure performance d'alignement en utilisant le gabarit « Early fusion », avec « Missalign rate » 7.94% et « Miss rate » 4.12%. En comparant le tableau 5.4 et le tableau 5.6, on constate que la stratégie « Early fusion » n'est pas robuste sous conditions de tous les échelles non linéaires.

Expérience	**Averge error (s)**	**Max error (s)**	**Missalign rate %**	**Miss rate %**
Late fusion	0.0678	1.2957	4.02	0.91
Gabarit échange	0.0734	2.5488	5.04	1.97

TABLE 5.7 – Les résultats d'alignement des expériences d supplémentaires.

La statistique de « Gabarit échange » montre que, même si un ensemble des gabarits de voyelle d'un locuteur différent sont implémentés dans le système, la stratégie « Late fusion » possède une performance excellente avec « Miss rate » 1.97%.

Chapitre 6

Conclusion

Récaptitulatif

Antescofo utilise des gabarits de hauteur et une observation de la hauteur qui s'adapte bien à la musique instrumentale, mais qui pose des problèmes de fiabilité pour le chant à cause de sa variabilité de la hauteur et de deux états consécutifs possédant la même hauteur. Pour résoudre ce problème par intégrer l'information phonétique dans le système et réaliser l'alignement voyelle/texte, dans le chapitre 3, nous construisons un ensemble des gabarits d'enveloppe spectrale de voyelle à partir de la partition et la base de données d'apprentissage. En parallèle, le signal du chant est représenté par l'enveloppe spectrale. Pour justifier l'estimation d'enveloppe spectrale - la « True envelope » est un bon descripteur de distinction des voyelles, nous conduisons une expérience vérification à la fin de ce chapitre.

Après la réalisation de l'alignement voyelle/texte, nous ne voulons pas abandonner l'information de hauteur parce que les deux informations ont leurs propres avantages. Dans le chapitre 4, nous concevons deux stratégies de fusion d'information : « Early fusion » et « Late fusion ». La première inspirée par le modèle de source/filtre est la multiplication point par point du gabarit de hauteur et du gabarit de voyelle. La deuxième est la moyenne arithmétique des deux probabilités d'observation.

L'évaluation du chapitre 5 montre que la stratégie « Late fusion » améliore la robustesse du système de toute évidence et aussi qu'il y a de nombreux problème à résoudre.

Perspectives

Tout d'abord, les résultats d'alignement doivent plus précisément être examiner sur chaque partie de la chansons (le couplet, le refrain, etc.) pour analyser les causes de « Missed note » et de « Missaligned note ».

On a évalué le gabarit de voyelle et les stratégies de fusion sur les bases de données de deux locuteurs qui contiennent un style de chant : pop. Afin de tester entièrement la robustesse du système, on doit introduire dans l'évaluation d'autres bases de données avec des locuteurs et des styles de chant différents.

Ensuite, dans l'ensemble des gabarits de voyelle, la variabilité d'enveloppe spectrale engendrée par la fréquence fondamentale f_0 n'est pas considérée. C'est-à-dire les états ayant la même voyelle mais différentes hauteur possèdent tout le même gabarit. Une amélioration possible est que nous construisons un ensemble des gabarits qui couvre la tessiture du chant. Cet ensemble des gabarits peut obtenu par enregistrer toutes les voyelles en toute la tessiture ou par l'appren-

tissage de la corrélation entre la f_0 et l'enveloppe spectrale sur une base de donnée assez grande [30].

Finalement, on néglige totalement la consonne dans l'alignement voix/texte et cela dégrade éventuellement la précision d'alignement, surtout pour les consonnes voisées qui sont parfois dominantes dans les événements musicaux. Par conséquent, il est nécessaire d'introduire des descripteurs pour décrire les consonnes et d'intégrer un ensemble des gabarits de consonne dans le système.

Bibliographie

[1] L. L. Beranek. *Acoustic Measurements*. New York : Wiley, 1949.

[2] J. C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1) :425–434, 1991.

[3] N. S. Di Carlo. Effect of multifactorial constraints on intelligibility of opera singing (ii). *Journal of Singing 63*, (4), 2007.

[4] A. Cont. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 32 :974–987, 2010.

[5] R. B. Dannenberg. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 193–198, IRCAM, France, 1984.

[6] B. Doval. Analyse et synthèse de la parole. Notes de cours. LAM-LJLRA. Master mention SDI, 2013-2014.

[7] N. Montecchio et A. Cont. A unified approach to real time audio-to-score and audio-to-audio alignment using sequential montecarlo inference techniques. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech, 2011.

[8] C. Schörkhuber et A. Klapuri. Constant-Q transform toolbox for music processing. In *Proceedings of the 7th Sound and Music Computing Conference*, Barcelona. Spain, 2010.

[9] F. Villavicencio et A. Röbel et X. Rodet. Applying improved spectral modeling for high quality voice conversion. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[10] A. Syrdal et A. Steele. Vowel F1 as a function of speaker fundamental frequency. *The Journal of the Acoustical Society of America*, 78, 1985.

[11] A. Cont et D. Schwarz et al. Evaluation of real-time audio-to-score alignment. In *International Symposium on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[12] R. Boite et H. Bourlard et al. Traitement de la parole. *Presses polytechniques et universitaires romandes*, 2000.

[13] C. Gussenhoven et H. Jacobs. *Understanding Phonology*. London : Hodder Arnold, New York : Oxford University press, 2nd edition, 2005.

[14] A. Novák et J. Vokrál. The speech intelligibility at the opera singing. *Sbornik Lekarsky*, 101(2) :153–164, 2000.

[15] S. S. Stevens et J. Volkman et E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3) :185–190, 1937.

[16] B. Roubeau et M. Castellengo. Revision of the notion of voice register. In *XIXth International CoMeT Congress*, Utrecht, Netherlands, 1993.

[17] S. Cornaz et N. Henrich et N. Vallée. L'apport d'exercices en voix chantée pour la correction phonétique en langue étrangère : le cas du français langue étrangère appliqué à des apprenants italiens d'âge adulte. *Phonétique, phonologie et enseignement des langues de spécialité - Volume 1*, 29(2) :103–119, 2010.

[18] C. J. Leggetter et P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9(2), 1995.

[19] E. Zwicker et R. Feldtkeller. *Psychoacoustique*. Ed. Masson, 1981.

[20] A. Cont et S. Dubnov et G. Assayag. On the information geometry of audio streams with applications to similarity computing. *IEEE Transactions on Audio, Speech and Language Processing*, pages 837–846, 2011.

[21] D. A. Reynolds et T. F. Quatieri et R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3), 2000.

[22] H. Zen et T. Nose et J. Yamagishi et S. Sako et T. Masuko et A.W. Black et K. Tokuda. The HMM-based speech synthesis system version 2.0. In *Speech Synthesis Workshop*, pages 294–299, Bonn, Germany, 2007.

[23] A. Röbel et X. Rodet. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *International conference on Digital Audio Effects (DAFx)*, Madrid, Spain, 2005.

[24] T. Galas et X. Rodet. Generalized discrete cepstral analysis for deconvolution of source-filter systems with discrete spectra. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 1991.

[25] S. Imai et Y. Abe. Spectral envelope extraction by improved cepstral method. *Electron. and Commun. (in Japan)*, 62-A(4) :10–17, 1979.

[26] J. Ginsborg. The influence of interactions between music and lyrics : What factors underlie the intelligibility of sung text ? *Empirical Musicology Review*, 9(1) :21–24, 2014.

[27] Y. Guédon. Hidden hybrid markov/semi-markov chains. *Computational Statistics and Data Analysis*, 49 :663–68, 2005.

[28] J. Makhoul. Linear prediction : A tutorial review. In *Proceedings of the IEEE*, volume 63, pages 561–580, Cambridge, Massachusetts, 1975.

[29] C. Meunier. Phonétique acoustique. *Chapitre 13. Les dysarthries. Auzou P. (Ed.)*, pages 164–173, 2007.

[30] N. Obin. Apprentissage de la corrélation de la fréquence fondamentale et de l'enveloppe spectrale : Application à la transposition de la voix parlée. Master's thesis, Université UPMC, IRCAM, Paris, Juin 2006.

[31] A. Oppenheim. Speech analysis-synthesis system based on homorphic filtering. *The Journal of the Acoustical Society of America*, 45(1) :458–465, 1969.

[32] L. Pierre. *Phonétisme et prononciation du français standard*. Paris, Nathan, 1992.

[33] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[34] G. A. F. Seber. *Multivariate Observations*. Hoboken, NJ : John Wiley & Sons, Inc, 1984.

[35] J. Sundberg. Articulatory interpretation of the 'singing formant. *The Journal of the Acoustical Society of America*, 55 :838–844, 1974.

[36] H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88 :97–91, 1990.

[37] W. Vennard. *Singing : The Mechanism and the Technique*. Carl Fisher, New York, NY, 1967.

[38] B. Vercoe. The synthetic performer in the context of live performance. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 199–200, IRCAM, France, 1984.

[39] J. Wells. Computer-coding the ipa : a proposed extension of sampa. `http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm`. Accessed : 2014-07-18.

[40] S. J. Young. The HTK Hidden Markov Model Toolkit : Design and Philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2 :2–44, 1994.