Modélisation des paramètres de contrôle pour la synthèse de voix chantée

L Ardaillon, Axel Roebel, Céline Chabot-Canet

CFA/VISHNO 2016

Modélisation des paramètres de contrôle pour la synthèse de voix chantée

L. Ardaillon, A. Roebel et C. Chabot-Canet
IRCAM, 1 Place Igor-Stravinsky, 75004 Paris, France
luc.ardaillon@ircam.fr

L'état de l'art de la synthèse vocale, et en particulier la synthèse concaténative, nous permet à ce jour d'obtenir une qualité d'élocution proche de la voix réelle, aussi bien pour la parole que pour le chant. Mais une synthèse à la fois naturelle et expressive ne peut être conçue sans un contrôle approprié, recouvrant de nombreux aspects à la fois timbraux et prosodiques, ainsi que leurs interdépendances. Pour le chant, la fréquence fondamentale (*F0*), portant la mélodie ainsi que certains aspects stylistiques, est à considérer en premier lieu. Une méthode de modélisation de la courbe de *F0* à partir de la partition, basée sur l'utilisation de B-splines, a été mise en place. Celle-ci permet une représentation paramétrique des variations expressives de la *F0* telles que le vibrato, les attaques, ou les transitions entre notes, avec un contrôle intuitif. Une première étude a permis d'établir qu'une telle représentation permet de reproduire de façon satisfaisante les variations propres à différents styles de chant. Mais le réglage manuel de l'ensemble des paramètres reste une tâche fastidieuse. Une gestion automatique de ces paramètres, basée sur un apprentissage et certaines règles, s'avère donc nécessaire, afin de réduire la quantité de réglages manuels à fournir. Les différents paramètres considérés varient d'un style de chant à l'autre. L'extraction de ces paramètres à partir d'enregistrements, ainsi que des contextes liés à la partition, doit donc permettre de capturer les caractéristiques propres au style interprétatif du chanteur, tout en conservant une certaine variabilité et la cohérence nécessaires à la production d'un chant naturel.

1 Introduction

Si les méthodes actuelles de synthèse de voix chantée, et en particulier la synthèse concaténative, permettent désormais une qualité satisfaisante, ouvrant la voie à de nouvelles applications artistiques [1], des progrès restent à faire pour obtenir une qualité comparable à la voix d'un chanteur professionnel. Le système de synthèse, au-delà de générer une voix présentant un timbre naturel comparable à celui d'une voix réelle, doit pouvoir également reproduire les différentes variations expressives des chanteurs, afin de rendre la synthèse plus vivante par l'ajout d'intentions musicales particulières. Pour cela, un modèle de contrôle adapté doit être conçu, qui, à partir d'une partition et d'un texte donnés, génère l'ensemble des paramètres de la synthèse, les principaux étant : la fréquence fondamentale (*F0*), l'intensité, les durées des phonèmes, et d'éventuelles variations et effets de timbre. L'objectif majeur est notamment de pouvoir générer automatiquement une interprétation expressive, pour un style donné, sans nécessiter un réglage fastidieux des paramètres de la synthèse par l'utilisateur.

Bien que chacun des paramètres évoqués ait son importance, la *F0*, qui porte la mélodie ainsi que de nombreuses caractéristiques stylistiques, comme souligné dans [2, 3, 4], est à considérer en premier lieu. Le principal intérêt de recourir à la synthèse de chant est de permettre à un utilisateur d'en contrôler précisément le rendu. Pour cette raison, il semble nécessaire de concevoir un modèle paramétrique permettant de caractériser les variations expressives de la *F0* de façon simple et intuitive. Dans [5], un tel modèle paramètrique a été proposé. Une première étude a montré que ce modèle, avec une paramétrisation adéquate, est capable de générer des courbes de *F0* comparables à celles extraites de réels enregistrements pour différents styles de chant. Mais la méthode présentée ne proposait pas de moyen pour sélectionner les paramètres du modèle de façon automatisée.

Différentes approches de contrôle de la *F0* pour la production de synthèses expressives et la modélisation des styles de chant ont déjà été proposées. Les systèmes de règles tels que [6] ont l'avantage de s'appuyer sur des connaissances musicales, et peuvent être progressivement améliorés, tout en apportant de nouvelles connaissances, dans une procédure d'analyse par la synthèse. Mais un travail approfondie est d'abord nécessaire pour déterminer les règles à utiliser pour chaque style représenté, ce qui en fait une approche peu flexible pour la modélisation de nouveaux styles. Les approches basées sur des HMM telles que [7] peuvent au contraire aisément modéliser de nouveaux styles de chant en choisissant une base d'apprentissage adaptée, sans nécessiter de connaissances particulières. Ceux-ci permettent notamment un contrôle globale de haut niveau et l'interpolation entre plusieurs styles vocaux, mais pas de contrôle local, à l'échelle de la note, de l'expressivité. Mais la principale limite à cette approche est qu'une quantité importante de données est nécessaire pour la phase d'apprentissage. L'approche par sélection d'unités proposée dans [8] permet, par l'utilisation de contours réels extraits d'enregistrements, de modéliser un style à l'aide de peu de données, tout en évitant d'éventuels problèmes de lissage (ou « *oversmoothing* ») liés à la modélisation statistique des paramètres dans les approches par HMMs. Mais une telle approche n'offre aucun contrôle à l'utilisateur pour modifier le résultat, à moins de redessiner manuellement la courbe. Dans [9], les auteurs proposent de modéliser les variations expressives de la *F0* par la sélection de templates paramétriques dans une base d'exemples extraits d'enregistrements commerciaux. Le principal intérêt de cette approche est qu'elle permet de caractériser quantitativement, à l'aide d'un jeu de paramètres restreint, les différentes variations expressives de la *F0*.

Dans [8] et [9], les unités et paramètres sont choisis dans une base d'exemples en fonction des contextes originaux dans lesquels ils ont été observés, et des contextes cibles de la partition à synthétiser, à l'aide de fonctions de coûts ou de règles déterminées. Pour cette raison, seul un ensemble restreint et figé de contextes sont pris en compte. Or, l'importance relative de ces différents contextes dans les choix interprétatifs peut être variable d'un chanteur à l'autre, ce qui ne peut être pris en compte avec ces approches.

Similairement à [9], nous proposons ici une approche basée sur la sélection de templates paramétriques pour la modélisation des styles de chant. Afin d'inclure dans la modélisation une représentation plus riche des contextes musicaux, et de mieux prendre en compte leur influence variable sur les choix interprétatifs des chanteurs, nous proposons d'utiliser une technique de regroupement des contextes (« *context-clustering* ») similaire à celle utilisée dans les méthodes de synthèse par HMM pour la parole

Le travail présenté ici est supporté par l'ANR dans le cadre du projet ChaNTeR (ANR-13-CORD-0011)

et le chant [10, 11], basée sur des arbres de décision. Dans une première phase d'apprentissage, les différents paramètres expressifs, ainsi que les contextes associés, sont extraits d'enregistrements, de façon à constituer une base d'exemples pour un style donné. Des arbres de décision sont ensuite construits pour ordonner automatiquement ces exemples en fonction de leurs contextes d'apparition et de l'influence de ceux-ci sur les valeurs des paramètres. Lors de la synthèse, ces arbres permettent ensuite de choisir dans la base les templates les plus appropriés, en fonction des contextes rencontrés dans la partition cible. Enfin, certaines règles peuvent également être appliquées afin de contraindre et corriger, si nécessaire, les paramètres des différents templates sélectionnés.

Cette approche peut également être adaptée à la modélisation d'autres paramètres. Les durées des consonnes, en particulier, varient de façon non linéraire avec le tempo et sont également souvent utilisées par le chanteur comme un moyen expressif pour accentuer certaines notes. Nous proposons donc ici d'inclure également ces effets dans notre modélisation des styles.

L'organisation de cet article se présente comme suit. Dans la section 2, le modèle de *F0* utilisé est présenté, puis la procédure d'extraction des paramètres de ce modèle sur des enregistrements est expliquée dans la section 3. La section 4 présente la méthode utilisée pour le choix des paramètres de la *F0*, ainsi que son application pour la modélisation des durées de phonèmes. Les premiers résultats obtenus par cette méthode sont ensuite évoqués dans la section 5, avant de présenter nos conclusions en section 6.

2 Modélisation de la *F0*

Dans [12], les auteurs suggèrent que les variations temporelles de la *F0* telles que le vibrato ou les transitions, utilisées pour ajouter de l'expressivité, sont des indices importants de l'identité d'un chanteur, et sont donc à modéliser en priorité. Dans [5], un modèle de génération de courbes de *F0* pour la synthèse de voix chantée a été proposé. Celui-ci décompose la courbe en 4 couches additives : une couche mélodique et une couche de vibrato (qui modélisent les composantes expressives de la *F0*), ainsi qu'une couche de jitter et une couche « phonétique », ou « micro-prosodique » (qui modélisent des variations non contrôlées induites par le mécanisme de la voix).

Dans ce modèle, la couche mélodique est découpée en 5 types de segments élémentaires : les silences ; les attaques, caractérisées par une montée de la *F0* en début de phrase musicale, après un silence ; les sustains, qui constituent les parties stables des notes portant le vibrato ; les transitions entre 2 notes successives ; et les releases, constituées d'une possible baisse de la *F0* en fin de phrase, avant un silence. Chacun de ces segments est caractérisé par un jeu restreint de paramètres. La figure 1 illustre la paramétrisation de ces différents segments. Les paramètres utilisés pour les attaques et les releases sont la durée L (en s) et la profondeur D (en *cents*). Les transitions sont contrôlées par 4 paramètres : d_L et d_R (s) qui déterminent les durées des parties gauche et droite de la transition, et A_L et A_R (*cents*) qui représentent les amplitudes de possibles inflexions à gauche (« préparation ») et à droite (« overshoot »). Le vibrato est caractérisé par sa fréquence f_{vib} (Hz) et une courbe d'amplitude de type

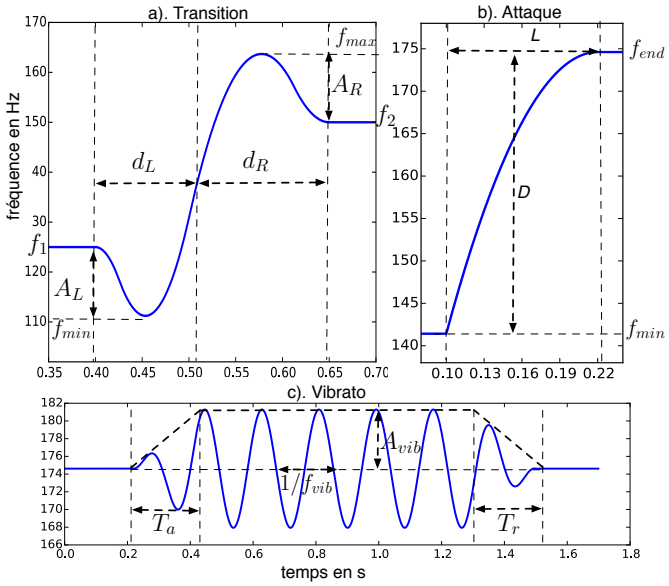

FIGURE 1 – Paramétrisation des différents segments du modèle de *F0*

Attack-Sustain-Release (ASR) déterminée par une amplitude globale A_{vib} (*cents*), un temps d'attaque T_a, et un temps de release T_r (s).

Pour générer la courbe de *F0*, la séquence de ces segments est tout d'abord déterminée à partir des notes de la partition, et leurs paramètres sont fixés. La courbe est ensuite générée par l'utilisation de B-splines dont les positions des noeuds et les poids sont déterminés par ces paramètres, comme expliqué dans [5], avant d'être additionnée aux autres couches du modèle. Certaines améliorations ont été appliquées à cette première version du modèle de *F0*. La principale modification est que le vibrato est désormais modélisé au sein de la couche mélodique, plutôt que dans une couche séparée, afin de permettre une meilleure fluidité dans l'enchaînement des transitions et du vibrato. Mais le détail de ces modifications ne sera pas abordé ici.

3 Extraction des paramètres

3.1 Corpus utilisé et annotations

Afin d'étudier et de modéliser différents styles de chants, un corpus a été constitué avec l'aide d'une musicologue. Lors de précédentes scéances d'enregistrements, il a été remarqué qu'il est difficile pour un chanteur d'imiter le style d'un autre en s'affranchissant totalement de ses propres caractéristiques stylistiques. Afin d'éviter ce biais, et de pouvoir également bénéficier de références culturelles bien connues ayant fait l'objet d'études musicologiques approfondies [2], nous avons choisi de baser notre étude sur des enregistrements commerciaux de chanteurs représentatifs de différents styles. Notre système étant prévu en premier lieu pour la synthèse du français, nous avons sélectionné 4 chanteurs français célèbres du 20^{e}siècle : Edith Piaf, Sacha Distel, Juliette Gréco, et François Leroux. Le choix de ces chanteurs a également été guidé par le fait qu'ils ont tous interprété, dans leur style propre, la chanson « *les feuilles mortes* », pouvant donc faire office de référence commune pour nos travaux

dans le cadre d'une évaluation.

Pour chaque chanteur, un ou deux morceaux (autre que « *les feuilles mortes* ») a été sélectionné, et chacun de ces morceaux a été annoté. Les annotations effectuées sont : la segmentation en phonèmes, la courbe de *F0* de la voix, et une annotation midi alignée sur l'interprétation du chanteur. Tout comme dans [9], les enregistrements utilisés sont des versions commerciales avec accompagnement instrumental. Les annotations ont donc due en grande partie être corrigées manuellement afin d'obtenir des données exploitables.

3.2 Extraction des paramètres de *F0*

Avant de pouvoir modéliser le style de chaque chanteur, il est tout d'abord nécessaire d'extraire les paramètres expressifs propres à chacun à partir des annotations du corpus. Pour cela, la courbe de *F0* de chaque chanson analysée est tout d'abord segmentée selon les unités définies par le modèle proposé, en *attaque*, *sustain*, *transition*, *release*, et *silence*. Cette segmentation est effectuée de façon semi-automatique, avec correction manuelle. Contrairement aux enregistrements utilisés dans [8], qui ne comportaient que des voyelles, les chansons de notre corpus contiennent les paroles originales. Afin de limiter l'impact des inflexions induites par certaines consonnes sur l'estimation des paramètres extraits, une interpolation linéaire de la *F0* est d'abord effectuée sur chaque consonne voisée (excepté les semi-voyelles), à l'aide de la segmentation phonétique du corpus. Les paragraphes suivants détaillent la procédure d'extraction des paramètres pour chaque type de segment.

3.2.1 Attaques et releases

Pour les attaques et release, D est calculé comme étant l'écart, en cents, entre la valeur minimum f_{min} et la valeur finale f_{end} (respectivement initiale) de la *F0* sur le segment selon la formule $D = 1200 \cdot log_2(\frac{f_{min}}{f_{end}})$. La durée L est directement extraite de la segmentation effectuée.

3.2.2 Sustains

Pour les segments de type sustain, un filtrage passe-bas est tout d'abord effectué sur le segment, afin d'isoler le vibrato et de le centrer en 0 par soustraction de la courbe filtrée à la courbe originale. Pour cela, un filtre RIF constitué d'une fenêtre de hanning de taille $2 \cdot sr \cdot T_{max}$ est utilisé, où sr est la fréquence d'échantillonage de la *F0*, et T_{max} est la période maximale supposée du vibrato pour le chanteur analysé. Les limites de chaque cycle de vibrato sont ensuite démarquées par l'extraction des points de croisement en 0 (« *zero-crossing* »). La fréquence du vibrato f_{vib} est alors déterminée comme l'inverse de la période moyenne des cycles appartenants au tiers central du segment. Cela permet d'éviter l'influence de possibles variations de la fréquence du vibrato en début ou fin de segment. Enfin, l'extrema de chaque demi-cycle de vibrato est repéré, et une courbe de type Attaque-Sustain-Release (ASR) est adaptée au mieux sur ces points (en valeur absolue de l'amplitude) par une procédure de recherche exhaustive (« *grid-search* »), avec un pas d'amplitude de 10 *cents* pour l'amplitude globale A_{vib} et un pas temporel de 0.05 s pour les temps d'attaques et de release T_a et T_r.

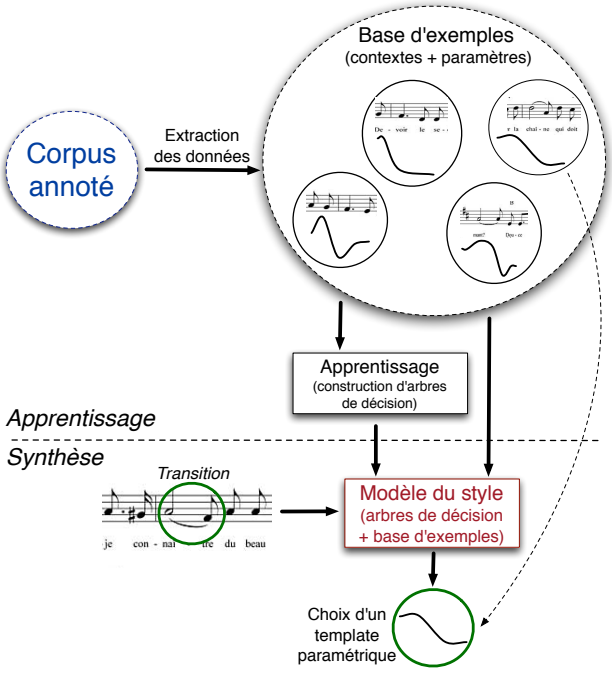

FIGURE 2 – Vue d'ensemble de l'approche proposée pour la modélisation du style. (Exemple du choix d'un template paramétrique de transition)

3.2.3 Transitions

Comme illustré sur la figure 1, les transitions sont séparées en 2 parties. Le milieu de chaque transition est déterminé par l'extremum, au sein du segment, de la dérivée, dont le signe correspond au sens de la transition. A partir de la segmentation, ce point permet de déterminer les durées des parties gauche d_L et droite d_R de la transition. Les amplitudes des préparations et overshoot A_L et A_R sont calculées comme l'écart, en cents, entre les fréquences extremums f_{min} et f_{max} de chaque côté de la transition et les fréquences f_1 et f_2 aux frontières du segment :

$$A_L = 1200 \cdot log_2(\frac{f_{min}}{f_1}) \quad \text{et} \quad A_R = 1200 \cdot log_2(\frac{f_{max}}{f_2})$$

4 Modélisation du style

Dans [13], les auteurs suggèrent que « [le concept de style implique un ensemble de motifs (ou « *features* ») cohérents] », et que « [le style personnel est porté par la répétition de ces motifs survenant dans certaines situations caractéristiques.] ». La modélisation d'un style consisterait donc notamment à pouvoir capturer ces motifs ainsi que les contextes caractéristiques dans lesquels ils sont observés. La figure 2 résume l'approche que nous proposons ici. Celle-ci se compose tout d'abord d'une phase d'apprentissage, qui, à partir d'un corpus annoté pour un style de chant donné, permet de construire un modèle de ce style. Ce modèle peut ensuite être utilisé lors de la synthèse afin de générer une interprétation expressive dans le style souhaité. Dans un premier temps, les contextes considérés dans la modélisation du style sont présentés, puis les approches utilisées dans les phases d'apprentissage et de synthèse sont expliquées.

4.1 Contextes utilisés

Dans les approches actuelles telles que [7] et [8], la portée des contextes utilisés est généralement limitée aux notes adjacentes ou à la mesure. Il nous semble cependant important de prendre en compte certains contextes plus larges, tels que la position temporelle ou mélodique d'une note au sein d'une phrase musicale complète, car ceux-ci peuvent influer sur les choix interprétatifs des chanteurs. Une analyse musicologique du corpus nous permet par exemple de remarquer qu'Edith Piaf use systématiquement d'un vibrato particulièrement marqué sur la dernière note d'une phrase. Dans [9], la position temporelle de la note au sein de la phrase musicale est également considérée.

Voici la liste des contextes qu'il nous a semblé utiles de considérer dans notre approche pour la modélisation de la *F0* :

- Note : hauteur midi ; durée
- Notes voisines (précédente et suivante) : hauteur midi ; durée ; intervalle, et différence de durée avec la note courante
- Phrase : positions temporelle des notes dans la phrase (première, dernière, ou avant-dernière note) ; position mélodique des notes dans la phrase (note la plus haute, note la plus basse, sommet ou vallée mélodique)
- Phonétique : caducité des notes courante et suivante (e muet prononcé en fin de phrase)

Cette liste a été établie avec l'aide d'une musicologue, par l'analyse des morceaux du corpus. L'ensemble des contextes utilisés sont extraits à partir des annotations effectuées. Nous définissons ici une phrase musicale comme un ensemble de notes situées entre 2 silences. D'autres contextes, tels que la position d'une note au sein de la mesure, ou bien la macro-structure de la partition pourraient également être utiles, mais ne sont actuellement pas pris en compte car ceux-ci ne peuvent être déduis des annotations du corpus.

Pour la modélisation des durées de phonèmes, certains contextes supplémentaires relatifs au texte doivent également être pris en compte, à savoir :

- Le nombre de consonnes successives
- La position par rapport aux autres consonnes de la syllabe (première, dernière, ou centrale)
- Les types des phonèmes précédents et suivants

4.2 Apprentissage

Etant donné la quantité réduites de données disponibles dans le corpus utilisé (une ou deux chanson seulement par style), l'utilisation d'une approche par HMM n'est pas envisageable. Dans [7], les auteurs utilisent une base d'apprentissage composée de 25 chansons, et 60 dans [11]. Une approche par sélection d'unités ou de templates paramétriques telles que [8] et [9] semble donc plus appropriée dans notre cas. Avec une approche de ce type, l'unité, ou le template, est choisi en fonction de son contexte d'observation qui doit correspondre au mieux au contexte cible. Or, la notion de « plus proche contexte » implique une relation de distance qui semble compliquée à définir, en raison des dimensions et importances variables des différents contextes considérés. [8] et [9] contournent ce problème par l'utilisation de fonctions de coûts et de règles déterminées empiriquement. Le défaut de telles approches est qu'elles sont peu flexibles, car elles ne permettent pas d'inclure de nouveaux contextes sans modifier les règles utilisées, ou de prendre en compte l'importance variable de certains contextes d'un style à l'autre. Une autre solution envisageable, dans le cas d'une modélisation paramétrique, est de classer automatiquement ces contextes par ordre d'importance selon leur influence sur le choix des paramètres expressifs. Pour cela, nous proposons ici d'utiliser une méthode de sélection et regroupement des contextes (« *context-clustering* ») basée sur l'utilisation d'arbres de décision, comme c'est le cas dans les approches par HMM telles que [14] pour la parole ou [7] pour le chant.

A partir de notre corpus annoté, et suite à l'extraction des paramètres expressifs, tel qu'expliqué dans la section 3, une base d'exemples est constituée. Celle-ci associe les paramètres extraits aux contextes dans lesquels ils ont été observés. A partir de cette base, des arbres de décisions binaires sont construits au moyen de l'algorithme CART. Dans notre cas, les valeurs des paramètres modélisés étant continues, il s'agit alors d'arbres de régression. Le critère utilisé pour le choix des questions à chaque noeud dans la construction de ces arbres est la minimisation de la variance, basée sur le calcul de l'erreur quadratique moyenne (MSE) [15].

Ainsi, la base d'exemples constituée pour un style donné, associée aux arbres construits à partir de cette base, forme notre modèle pour ce style de chant. Cette approche a été appliquée pour la modélisation de la *F0*, mais également pour prédire les durées des phonèmes pour la synthèse. Les paragraphes suivants détaillent les spécificités liées à chacun de ces 2 paramètres.

4.2.1 *F0*

Contrairement aux approches par HMMs, nous n'effectuons pas de modélisation statistique des paramètres considérés, et basons notre approche sur la sélection de templates paramétriques, similairement à [9], ce qui permet de modéliser de façon plus explicite les variations expressives de la *F0* qu'avec les paramètres bas niveau utilisés dans les approches par HMM, tout en nécessitant peu de données. Pour cela, les différents paramètres du modèle de *F0* présentés dans la section 2 ne sont pas considérés de façon indépendante, mais de façon liée, pour chaque type de segment. Ainsi, un arbre de décision « multi-variable » est construit pour chaque type de segment. Ces arbres sont construit de manière à minimiser la variance, à chaque noeud, de manière simultanée sur l'ensemble des paramètres considérés.

L'ensemble des contextes énoncés ci-dessus pour la *F0* est utilisé pour la construction des arbres, excepté pour les segments de type « attaque » et « release », pour lesquels nous avons seulement considéré la hauteur et la durée de la note courante (respectivement première et dernière notes de la phrase). Pour les segments de transition, la note considérée comme courante est la note située à droite de la transition. De plus, les transition non voisées, ne présentant pas de courbe de F0 continue permettant d'extraire des paramètres de manière fiable, ne sont pas utilisées. Enfin, étant donné que les valeurs des différents paramètres appartiennent à des dimensions très variables (durées en secondes, amplitudes

en cents, et fréquence en Hz), celles-ci sont tout d'abord normalisées par leur écart-type, de sorte que toutes les valeurs soient du même ordre de grandeur, afin de ne pas privilégier un paramètre sur un autre.

Afin de permettre une certaine variabilité dans le choix des paramètres pour la synthèse, un critère d'arrêt est fixé de manière à avoir, pour chaque feuille de l'arbre, un nombre minimum de templates disponibles.

4.2.2 Durées des phonèmes

Après la *F0*, la durée des phonèmes est également un vecteur important d'expressivité dans le chant. Un chanteur peut par exemple transmettre une intention musicale en appuyant particulièrement une syllabe par l'allongement de certaines consonnes. Afin de modéliser ces effets, la même procédure est appliquée pour générer des durées propres à chaque chanteur, selon les contextes, pour la synthèse. Pour cela, les durées sont directement extraites à partir des segmentations phonétiques des chansons du corpus. Comme on ne peut s'attendre à rencontrer chaque phonème dans une variété de contextes importante au sein d'une ou deux chansons seulement, ceux-ci sont rassemblés par classe avec les autres phonèmes présentant des caractéristiques articulatoires (et donc des durées) proches. Lors de l'apprentissage, un arbre est construit pour chaque classe, tel que décrit précédemment. Les classes de phonèmes considérées sont (avec les notations phonétiques SAMPA des phonemes correspondants) : fricatives voisées (v,z,Z) ; fricatives non voisées (f,s,S) ; plosives voisées (b,d,g) ; plosives non voisées (p,t,k) ; nasales (m,n) ; semi-voyelles (w,j,H) ; R ; et l. Le phonème lui-même devient alors un contexte utilisé dans la construction de l'arbre pour chacune de ces classes.

4.3 Synthèse

4.3.1 Choix des paramètres

Une fois la phase d'apprentissage terminée, un modèle du style peut être construit, associant les arbres de décisions à la base d'exemples du style. Ce modèle peut ensuite être utilisé lors de la synthèse afin de sélectionner les paramètres selon les contextes liés à la partition à synthétiser. Pour cela, le séquencement des différents segments du modèle de *F0* est tout d'abord déterminé en fonction des notes de la partitions, comme expliqué dans [5]. Pour chaque segment, rattaché à une note de la partition, il suffit ensuite de parcourir l'arbre depuis la racine en fonction des contextes cibles liés à ce segment, jusqu'à arriver sur une feuille de l'arbre. Un template est alors sélectionné aléatoirement dans la base d'exemple parmis ceux associés à la même feuille de l'arbre, et ses paramètres sont utilisés pour la synthèse. Cette sélection alétoire permet de conserver une part de variabilité dans l'interprétation générée. Contrairement à [8], aucune règle spécifique n'est nécessaire pour sélectionner dans la base des transitions de même direction que la cible, car cela est pris en compte dans les contextes utilisés (d'après l'intervalle entre 2 notes consécutives) et cette règle est donc implicitement modélisée lors de la construction de l'arbre.

De même, pour les durées des phonèmes, une valeur au hasard est prise sur la feuille de l'arbre correspondant au contexte cible, pour la classe de phonème considéré.

4.3.2 Règles et corrections additionnelles

Une fois les paramètres sélectionnés, certaines règles peuvent également être appliquées, afin d'améliorer la cohérence du résultat. En particulier, l'observation de nombreux enregistrements nous ont permis de remarquer que les inflexions des transitions (préparation et overshoot) se confondent généralement avec les inflexions phonétiques dues à la prononciation des consonnes. Des règles sur le placement des transitions ont donc été établies et sont appliquées ici :

- Les transitions montantes commencent au temps de la première consonne d'une syllabe, si celle-ci ne contient pas de semi-voyelle. S'il y a une semi-voyelle, la transition commence sur le début de celle-ci.
- Les transitions descendantes se terminent sur l'onset de la voyelle.

Ces règles permettent d'éviter un mauvais placement des transitions par rapport au texte qui pourrait sonner peu naturel. Dans le cas de transition entre deux voyelles, aucune règle particulière n'est appliquée.

De plus, afin d'éviter le choix de valeurs trop faibles ou trop importantes pour les paramètres du modèle (par exemple dûes à des erreurs d'annotations ou dans l'extraction automatique des paramètres), il est également possible d'imposer des limites sur ces valeurs.

Une fois les paramètres séléctionnés et ces règles appliquées, il est possibles que certains segments du modèle de F0 se recouvrent partiellement (par exemple 2 transitions de part et d'autre d'une note courte). Les paramètres de durées de ces segments sont alors réduits de telle sorte que la fin du 1er segment coïncide avec le début du 2nd. De même les durées des consonnes contenues dans une note sont réduites si nécessaire à une durée totale maximale (dans notre cas 60% de la durée de la note).

5 Expérience et résultats

L'approche proposée dans cet article a été intégrée au sein du système de synthèse concaténative présenté dans [5]. Des modèles de styles ont été générés à partir des différentes chansons du corpus (excepté « *les feuilles mortes* »), et ceux-ci ont été utilisés pour synthétiser différents extraits des « *feuilles mortes* » dans ces différents styles. D'après des écoutes informelles sur les premiers résultats obtenus, il semble que la procédure utilisée permet de produire une synthèse plus expressive qu'avec un paramétrage par défaut, sans nécessiter d'intervention supplémentaire de l'utilisateur. Les résultats obtenus nous permettent également de dire que l'approche proposée permet bien de capturer certaines caractéristiques propres aux styles des différents chanteurs, perceptibles en particulier dans la conduite du vibrato et les durées des transitions générées, ce qui est donc encourageant.

L'une des difficulté dans l'évaluation de l'expressivité et de la modélisation d'un style est l'absence de critères quantitatifs, permettant de juger objectivement des résultats obtenus. La validation de ces résultats doit donc passer par une évaluation subjective, à l'aide de tests d'écoute qui devront être menés prochainement.

6 Conclusion

Nous avons présenté dans cette article une méthode de modélisation des paramètres stylistiques pour la synthèse de chant expressif. Cette méthode, basée sur la sélection de templates paramétriques extraits d'enregistrements, nous permet de choisir les paramètres d'un modèle de *F0*, ainsi que les durées des phonemes, en fonction des contextes liés à une partition cible, par l'utilisation d'arbres de décision. Un avantage d'une telle méthode est qu'elle nécessite peu de données. De plus, le modèle de *F0* utilisé propose une paramétrisation permettant à un utilisateur de corriger si besoin les résultats obtenus de manière simple et intuitive.

Une utilisation intéressante de cette approche serait également de pouvoir apprendre le style d'un utilisateur du synthétiseur, au fur et à mesure de son utilisation. Plutôt que de se baser sur des enregistrements réels, celui-ci pourrait alors commencer avec un modèle de style par défaut, puis apporter des corrections selon ses envies. Le système mettrait alors à jour le modèle selon ces paramètres modifiés pour construire un modèle correspondant mieux aux goûts de l'utilisateur, dans un processus itératif.

Un autre avantage d'utiliser des arbres de décision est que, contrairement à d'autres méthodes d'apprentissage automatique, ceux-ci sont facilement lisibles et interprétables par un humain. Au-delà de leur utilisation pour la synthèse, leur lecture peut donc également éventuellement nous permettre de vérifier certaines hypothèses émises par l'analyse musicologique sur l'importance de certains facteurs contextuels dans les choix interprétatifs d'un chanteur.

Les paramètres modélisés actuellement ne permettent de capturer qu'une partie des caractéristiques stylistiques des chanteurs, ne suffisant pas à caractériser complètement un style. Nos travaux devront donc être étendus à la modélisation d'autres paramètres expressifs, en commançant par l'intensité. Enfin, les résultats obtenus devront être évalués à l'aide de tests d'écoute subjectifs.

Références

[1] Hideki Kenmochi et al. Singing synthesis as a new musical instrument. In *ICASSP*, volume 2012, pages 5385–5388, 2012.

[2] Céline Chabot-Canet. *Interprétation, phrasé et rhétorique vocale dans la chanson française depuis 1950 : expliciter l ?indicible de la voix*. PhD thesis, Université Louis Lumière-Lyon II, 2013.

[3] Takeshi Saitou and Masataka Goto. Acoustic and perceptual effects of vocal training in amateur male singing. In *INTERSPEECH*, pages 832–835. Citeseer, 2009.

[4] Tatsuya Kako, Yasunori Ohishi, Hirokazu Kameoka, Kunio Kashino, and Kazuya Takeda. Automatic identification for singing style based on sung melodic contour characterized in phase plane. In *ISMIR*, pages 393–398. Citeseer, 2009.

[5] Luc Ardaillon, Gilles Degottex, and Axel Roebel. A multi-layer f0 model for singing voice synthesis using a b-spline representation with intuitive controls. In *Interspeech 2015*, 2015.

[6] Gunilla Berndtsson. The kth rule system for singing synthesis. *Computer Music Journal*, 20(1) :76–91, 1996.

[7] Takashi Nose, Misa Kanemoto, Tomoki Koriyama, and Takao Kobayashi. Hmm-based expressive singing voice synthesis with singing style control and robust pitch modeling. *Computer Speech & Language*, 34(1) :308–322, 2015.

[8] Martı Umbert, Jordi Bonada, and Merlijn Blaauw. Generating singing voice expression contours based on unit selection. In *Proc. SMAC*, 2013.

[9] Yukara Ikemiya, Katsutoshi Itoyama, and Hiroshi G Okuno. Transferring vocal expression of f0 contour using singing voice synthesizer. In *Modern Advances in Applied Intelligence*, pages 250–259. Springer, 2014.

[10] Takayoshi YoshimuraÝ, Keiichi TokudaÝ, Takashi MasukoÝÝ, Takao KobayashiÝÝ, and Tadashi KitamuraÝ. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In *Proc. Eurospeech*, pages 2347–2350, 1999.

[11] Keijiro Saino, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, and Keiichi Tokuda. An hmm-based singing voice synthesis system. In *INTERSPEECH*, 2006.

[12] Tin Lay Nwe and Haizhou Li. Exploring vibrato-motivated acoustic features for singer identification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(2) :519–530, 2007.

[13] Chilin Shih and Greg Kochanski. Prosody control for speaking and singing styles. In *INTERSPEECH*, pages 669–672, 2001.

[14] Nicolas Obin. *MeLos : Analysis and modelling of speech prosody and speaking style*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2011.

[15] sklearn decision trees documentation. http ://scikit-learn.org/stable/modules/tree.html, 2016.