Modélisation HMM de la variabilité intonative pour la synthèse de parole

Cédric Boidin [1], *Olivier Boeffard* [2]

[1] Orange Labs, Lannion, France
[2] IRISA / Université de Rennes 1 – ENSSAT, Lannion, France
cedric.boidin@orange-ftgroup.com, olivier.boeffard@irisa.fr

ABSTRACT

This paper proposes a statistical intonation model designed to deal with intrinsic variability in speech. In combining the advantages of two well-known statistical algorithms, CART and HMM, the proposed model takes advantage of available linguistic information and successfully tackles the issue of missing para-linguistic information. Promising results of the training process are shown and analyzed.

1. INTRODUCTION

Les systèmes de synthèse par corpus ont amélioré significativement la qualité de la synthèse vocale. Leur succès est basé sur la concaténation d'unités acoustiques extraites de grandes bases de données audio. Ces systèmes fonctionnent très bien sur des corpus homogènes de parole lue ou de parole dite neutre.

Cependant, la qualité de ce type de synthèse se dégrade au fur et à mesure que les corpus diminuent en taille ou augmentent en expressivité et en richesse prosodique. Dans ces contextes, un modèle prosodique devient indispensable pour guider la sélection des unités acoustiques ou spécifier les modifications prosodiques.

Ce modèle prosodique se doit d'être statistique, afin de s'adapter au locuteur, au style prosodique, etc. De tels modèles statistiques ont déjà été utilisés avec succès en sélection des unités [1], en reconnaissance vocale [2], ou en modélisation prosodique [3].

L'apprentissage prosodique reste néanmoins particulièrement compliqué, parce que les facteurs influençant la prosodie sont très complexes ; ce sont par exemple l'attitude ou l'intention du locuteur, ou encore la place de l'accent dans le groupe accentuel. Même dans les descriptions linguistiques les plus avancées utilisées en "Concept-To-Speech", il n'existe pas de relation biunivoque entre description linguistique et prosodie.

Par exemple, la figure 1 représente deux courbes du fondamental extrait du signal acoustique correspondant au mot "d'accord". Les contours intonatifs sont très différents même si les contextes linguistiques sont identiques. Les différences entre les deux contours proviennent probablement de différences d'ordre para-linguistique, i.e. d'ordre sémantique ou pragmatique. Ces différences sont cependant très difficiles à décrire explicitement, et un étiquetage de ce type d'information est actuellement hors de portée.

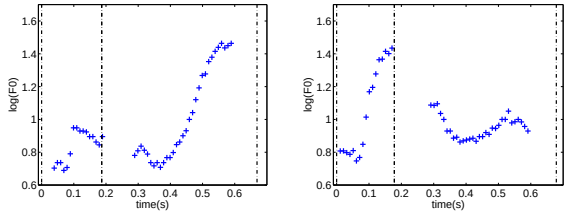

figure 1 : Courbes de fréquence fondamentale (log) du mot "d'accord" dans deux phrase différentes : "D'accord, je le conserve." à gauche, "D'accord, à partir de maintenant vous avez des messages courts." à droite. Les pointillés correspondent aux frontières de syllabes.

Dans cet article, nous avons donc décidé de mettre ces différences sur le compte de la variabilité intrinsèque au signal de parole, et de modéliser cette variabilité comme de l'information non observée par le moyen de variables cachées. Le modèle d'intonation que nous proposons utilise donc à la fois l'information linguistique disponible, au moyen d'un arbre de classification -CART-, et modélise l'information cachée liée à la variabilité au moyen d'un HMM. Pour revenir à l'exemple de la figure 1, le HMM sera censé modéliser distinctement les deux réalisations alors qu'elles ont le même contexte linguistique.

Le reste du document est organisé de la manière suivante : la partie 2 décrit le modèle d'intonation proposé, et avant de conclure, la partie 3 présente une évaluation du modèle à travers des mesures objectives.

2. INTRODUCTION DE LA VARIABILITÉ DANS UN MODÈLE STATISTIQUE

La figure 2 représente un schéma du modèle intonatif proposé. Il se décompose en trois composants fonctionnels : la stylisation intonative, la classification par arbre et enfin la modélisation par HMM.

2.1. Stylisation intonative et annotation

Nous prenons pour hypothèse que le modèle proposé est ancré phonologiquement sur la syllabe. Chaque syllabe est représentée par un vecteur de caractéristiques extraites de l'énoncé acoustique.

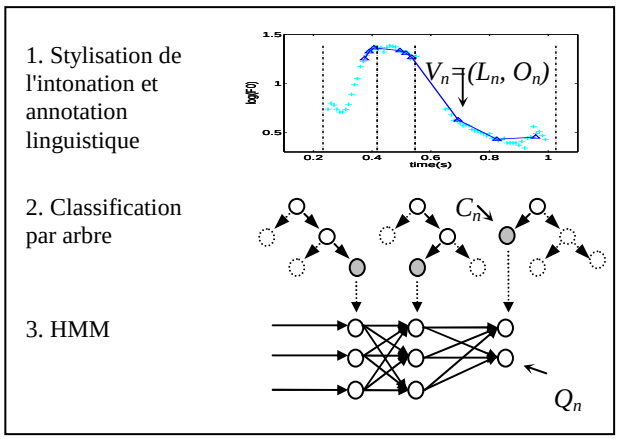

figure 2 : Vue générale du modèle pour la phrase "Je l'appelle.", composée de 3 syllabes. L'étape 1 correspond à la stylisation intonative et à l'annotation linguistique, l'étape 2 à la classification par arbre, et l'étape 3 à la modélisation HMM.

La fréquence fondamentale (F_0) est extraite automatiquement. Comme dans [4], les erreurs grossières et la microprosodie sont supprimées pour donner un contour lissé de F_0, exprimé sur une échelle logarithmique.

Chaque phrase est segmentée en groupes de souffle, syllabes et phonèmes. Pour chaque syllabe, le noyau syllabique est identifié et son contour de F_0 est modélisé par un polynôme d'ordre 2 comme dans [5]. Les points situés à 10%, 50% et 90% de la durée du noyau constituent le vecteur d'intonation de la syllabe.

Une transformation de Karhunen-Loeve est ensuite appliquée à tous les vecteurs d'intonation de façon à les normaliser et à les projeter dans un nouvel espace dans lequel leurs composantes seront orthogonales. Les paramètres de cette transformation sont appris sur un corpus d'apprentissage, puis appliqués à tous les vecteurs d'intonation quels qu'ils soient. Pour chaque phrase le résultat de cette étape est une séquence de vecteurs d'intonation normalisés en dimension 3, $\{O_n\}_{1\leq n\leq N}$, regroupés en groupes de souffle.

Annotation linguistique

Les syllabes sont automatiquement annotées linguistiquement à partir du texte. Les étiquettes sont la catégorie grammaticale du mot, la position de la syllabe dans le mot, la position du mot dans le groupe de souffle et le type de groupe de souffle. A chaque syllabe est attribué un vecteur linguistique, $\{L_n\}_{1\leq n\leq N}$, comprenant les étiquettes susnommées de la syllabe concernée ainsi que celles des syllabes précédente et suivante.

Chaque phrase est donc décrite par une séquence de vecteurs mixtes $\{V_n\}_{1\leq n\leq N}$ associant les vecteurs linguistiques et les vecteurs d'intonation : $V_n = (L_n, O_n)$.

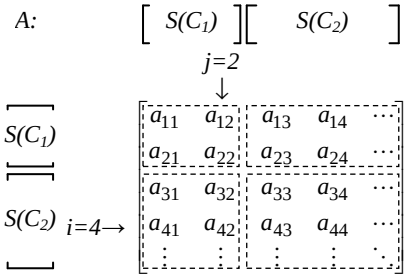

figure 3 : Exemple de matrice de transition *A*. L'arbre donne deux classes C_1 (regroupant 2 états $Q_n \in S(C_1) = \{1,2\}$), et C_2 (regroupant les autres états). a_{42} correspond à $p_{\{C_2, C_1\}}(Q_n=2|Q_{n-1}=4)$.

2.2. Classification par arbre (CART)

Comme pour plusieurs modèles prosodiques existants [6], un arbre de classification non-supervisée est utilisé. Il attribue une classe C_n à chaque syllabe décrite par son vecteur linguistique L_n, et est décrit par une fonction $T: L_n \rightarrow T_{V_{train}}(L_n) = C_n$.

L'arbre est appris sur un corpus d'apprentissage $V_{train} = \{V_1 \ldots V_N\}$. La phase d'apprentissage consiste à construire de façon récursive les nœuds de l'arbre, c'est-à-dire à choisir les étiquettes linguistiques des L_n ainsi que les partitions binaires des valeurs associées à ces étiquettes, qui maximisent la variance inter-groupes.

Les feuilles de l'arbre constituent les classes C_n. Le nombre de feuilles, donc de classes C_n augmente avec la complexité (cp) de l'arbre, fixée manuellement par le choix d'un critère d'arrêt.

2.3. Modélisation HMM

L'étape de modélisation HMM est ajoutée afin de modéliser la variabilité intonative. Elle introduit des états cachés qui autorisent plusieurs réalisations intonatives pour chaque contexte linguistique : chaque classe est subdivisée en plusieurs états cachés, chaque état caché est associé à une unique classe.

La séquence des états cachés est décrite par $\{Q_n\}_{1\leq n\leq N}$. L'association classe/état est définie par la fonction $S: C_n \rightarrow S(C_n)$, $S(C_n)$ étant un sous-ensemble des états cachés Q_n.

Nous définissons alors un HMM standard :

$$P(O,Q) = P(O_1,\ldots O_N,Q_1,\ldots Q_N)$$

$$= P_{T(L_1)}(Q_1)P(O_1|Q_1)\prod_{n=2}^{N} P_{\{T(L_n),T(L_{n-1})\}}(Q_n|Q_{n-1})P(O_n|Q_n)$$

La matrice de transition $A=\{a_{ij}\}$ est définie par :

$$p_{\{C_k, C_l\}}(Q_n=j|Q_{n-1}=i) = a_{ij} \text{ tel que } (i, j) \in (S(C_k), S(C_l))$$

avec la propriété : $\forall (i, l), \sum_{j\in S(C_l)} a_{ij}=1.$

La figure 3 illustre une telle matrice de transition *A*. C'est une matrice pleine subdivisée en blocs indépendants représentant les sous-matrices de transition d'une classe vers l'autre. Chaque bloc est ainsi normalisé comme une matrice de transition habituelle, i.e. la somme des coefficients sur une ligne vaut 1.

Identiquement nous définissons les probabilités initiales $\Pi = \{\pi_i\}_{1 \le i \le N}$ comme :

$$p_{C_k}(Q_1 = i) = \pi_i \text{ tel que } i \in S(C_k)$$

avec la propriété : $\forall k, \sum_{i \in S(C_k)} \pi_i = 1$.

Nous définissons les probabilités d'observation $B = \{b_j\}_{1 \le j \le N}$, la probabilité d'un vecteur d'observation à un instant n pour un état j étant modélisée par une loi gaussienne : $b_j(o_n) = p(O_n = o_n | Q_n = j) = \mathcal{N}(o_n | \mu_j; \Sigma_j)$.

Les paramètres du modèle $\lambda = \{B; A; \Pi\}$ sont appris grâce à un algorithme EM, afin d'obtenir le jeu de paramètres optimal λ^* qui maximise la vraisemblance $P(O_{train} | \lambda)$. Le modèle a été entièrement réimplémenté pour prendre en compte la nouvelle formulation des probabilités initiales et des probabilités de transition.

Puisque les composantes des vecteurs d'intonation ont été globalement décorrélées sur un corpus d'apprentissage, nous faisons l'hypothèse qu'elles sont également décorrelées pour chacun des états et choisissons donc des matrices de covariance diagonales. Les paramètres initiaux de l'EM sont obtenus grâce à une Quantification Vectorielle de chaque classe.

3. EXPERIENCES

3.1. Protocole expérimental

Nous avons testé notre modèle sur un corpus du français composé de phrases d'un service vocal interactif de France Télécom. Les phrases ont été enregistrées par une locutrice professionnelle pour le service opérationnel, la prosodie des phrases est donc naturelle et expressive. Le corpus contient 582 phrases, 1033 groupes de souffle et 6439 syllabes.

Les phrases ont été segmentées manuellement en phonèmes, les syllabes automatiquement déduites du découpage en phonèmes et étiquetées linguistiquement. La fréquence fondamentale a été extraite automatiquement toutes les 10 ms et stylisée selon la méthode décrite en 2.1.

Nous avons réalisé des expériences de manière à mesurer l'aptitude du modèle à modéliser des données réelles et à déterminer le jeu de paramètres optimal.

Le processus expérimental consiste à effectuer un apprentissage (normalisation, CART et HMM) sur un corpus d'apprentissage et à utiliser un corpus de test pour vérifier la qualité du modèle obtenu. Les mesures effectuées sont la log-vraisemblance du corpus de test et du corpus d'apprentissage, ainsi que le nombre total d'états du HMM.

Les paramètres du modèle sont le taux de complexité de l'arbre (HMM correspond à un arbre à une seule feuille ; low-cp est un arbre avec peu de feuilles ; middle-cp ; high-cp est un arbre complexe avec de nombreuses feuilles), ainsi que le nombre d'états associés à chaque classe (avec un maximum de *int(M/100) + 1* états pour une classe regroupant *M* syllabes). Le nombre total d'états du HMM est égal à la somme du nombre d'états de toutes les classes, il est donc égal au nombre de classes multiplié par le nombre d'états par classe, tant qu'aucune classe n'a atteint son maximum d'états.

La figure 4 correspond à des modèles allant de 1 à 8 états par classe, tandis que la figure 5 montre des expériences supplémentaires menées avec un plus grand nombre d'états par classe. De manière à avoir des résultats significatifs, chaque expérience a été reproduite 10 fois avec une répartition aléatoire entre corpus d'apprentissage (80% des phrases) et corpus de test (20% restants).

3.2. Résultats : comparaison des log-vraisemblances

La figure 4 montre l'évolution de la log-vraisemblance moyenne (par syllabe) en fonction du nombre d'états par classe, pour les corpus d'apprentissage (trait plein) et de test (trait pointillé) et pour les différents taux de complexité de l'arbre, ainsi que leurs intervalles de confiance à 95%. Les intervalles de confiance peuvent ne pas être visibles s'ils sont plus petits que les symboles.

On retrouve sur cette figure les composants -CART et HMM- pris isolément. Ainsi, la courbe bleue aux symboles carrés correspond à un HMM ergodique, toutes transitions entre états étant autorisées ; cela correspond dans notre formalisme à une seule classe C_n issue de l'arbre. De même, les premiers points de chaque courbe (les plus à gauche) correspondent à un état par classe, ce sont donc les résultats obtenus uniquement avec la classification par arbre. Le nombre moyen de classes obtenus par la construction de l'arbre est respectivement 4.0, 7.8 et 12.3 pour les arbres de complexité low, middle et high.

Sur la figure 4, les différences de vraisemblance entre corpus d'apprentissage et corpus de test sont faibles, signe d'une bonne qualité de l'apprentissage. La vraisemblance augmente avec le nombre d'états par classe, cela montre que le modèle HMM capture bien la variabilité intonative que le CART seul ne peut pas modéliser.

Les modèles de plus grande complexité semblent donner de meilleurs résultats que les modèles de faible

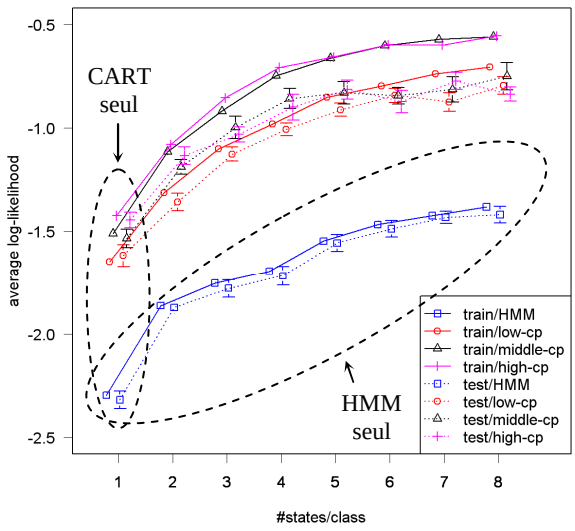

figure 4: Evolution de la log-vraisemblance en fonction du nombre d'états par classe et de la complexité (cp) de l'arbre sur les corpus de test et d'apprentissage.

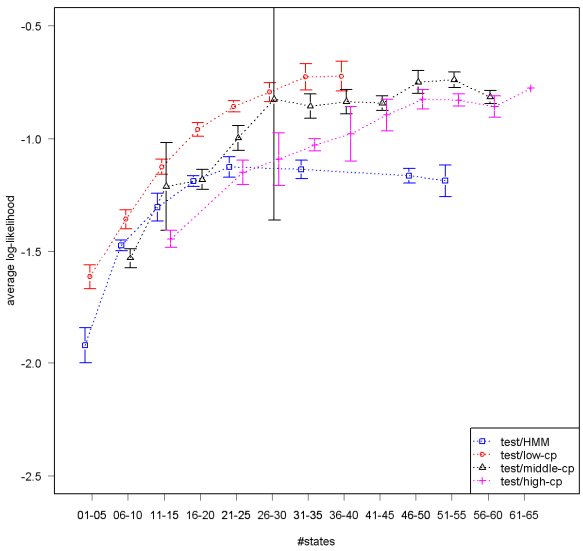

figure 5 : Evolution de la log-vraisemblance en fonction du nombre total d'états et de la complexité (cp) sur le corpus de test.

complexité, la différence est particulièrement frappante entre le HMM seul et les autres modèles. Cette comparaison est bien sûr critiquable car, étant donné le nombre d'états par classe, le nombre total d'états est plus faible pour les modèles à arbres peu complexes que pour les modèles à arbres plus complexes.

La figure 5 montre les même courbes, mais en fonction du nombre total d'états du HMM, et uniquement sur le corpus de test. Les nombres totaux d'états ont été groupés dans des intervalles de largeur 5 pour simplifier les figures et calculer des intervalles de confiance à 95%.

Cette fois-ci, les modèles à arbre complexe obtiennent des vraisemblances inférieures aux modèles plus simples à nombre total d'états équivalent. Le modèle HMM standard fait exception : bien que son arbre soit le moins complexe (une seule classe), sa vraisemblance n'est pas aussi bonne que d'autres modèles plus complexes. De plus, il souffre de sur-apprentissage quand le nombre total d'états augmente. Sur cette figure, le modèle low-cp apparaît comme le meilleur compromis entre nombre total d'états et vraisemblance.

4. Conclusion

Nous avons proposé dans cet article un modèle d'intonation statistique conçu pour capter au mieux la variabilité prosodique. Le comportement de cette variabilité est enfoui dans un modèle de type HMM. Les états non observés correspondent à un apprentissage non supervisé de classes mélodiques sous la contrainte d'étiquettes linguistiques.

Les premiers résultats montrent un bon comportement à l'apprentissage sur des données expressives. Le modèle surpasse chacun de ses sous-composants -CART et HMM- pris indépendamment. Les résultats expérimentaux montrent aussi qu'un arbre peu complexe associé à un nombre élevé d'états par classe (8 ou plus) semble être le meilleur compromis entre vraisemblance et complexité.

Les résultats sont prometteurs mais ce ne sont que des mesures de vraisemblance. Il faut maintenant tester ce modèle en synthèse de parole, soit pour générer une courbe d'intonation, soit pour assister l'étape de sélection des unités, et tester subjectivement la prosodie obtenue.

Bibliographie

[1] A. Hunt and A. Black, Unit selection in a concatenative speech synthesis system using a large speech database, *ICASSP-96*.

[2] L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, February 1989

[3] C. Traber, *Talking machines: theories, models and designs*, chapter F0 generation with a database of natural F0 patterns and with a neural network, pages 287-304, 1992.

[4] S. Narusawa, H. Fujisaki and S. Ohno, A method for automatic extraction of parameters of the fundamental frequency contour, *ICSLP 2000*.

[5] G. Bailly and B. Holm, SFC: a trainable prosodic model, *Speech Communication 46*, 2005.

[6] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, M. Viswanathan, Recent improvements to the IBM trainable speech synthesis system, *ICASSP 2003*.