Coopération entre méthodes locales et globales pour la segmentation automatique de corpus dédiés à la synthèse vocale *

Safaa Jarifi[1], *Olivier Rosec*[2], *Dominique Pastor*[1]

[1] ENST Bretagne, 29285 Brest Cedex, France
{safaa.jarifi,dominique.pastor}@enst-bretagne.fr

[2] France Télécom, Division R&D TECH/SSTP/VMI,
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
olivier.rosec@francetelecom.com

ABSTRACT

This paper introduces a new approach for the automatic segmentation of corpora dedicated to speech synthesis. The main idea behind this approach is to merge the outputs of three segmentation algorithms. The first one is the standard HMM-based (Hidden Markov Model) approach. The second algorithm uses a phone boundaries model, namely a GMM (Gaussian Mixture Model). The third method is based on Brandt's GLR (Generalized Likelihood Ratio) and aims to detect signal discontinuities in the vicinity of the HMM boundaries. Different fusion strategies are considered for each phonetic class. The experiments presented in this paper show that the proposed approach yields better accuracy than existing methods.

1. Introduction

L'approche de synthèse par corpus (SPC) repose sur la concaténation de segments de parole contenus dans une grande base de données enregistrée par un locuteur professionnel. Le succès de cette technologie tient au fait que, moyennant une couverture acoustico-prosodique suffisante, il devient possible de sélectionner une séquence d'unités acoustiques correspondant au contexte de synthèse. De ce fait, les modifications des unités de synthèse peuvent être sinon évitées, du moins limitées, ce qui permet de préserver le naturel de la parole synthétique ainsi produite. Cependant, avec la SPC, la création de nouvelles voix de synthèse devient extrêmement coûteuse, car, outre l'enregistrement du corpus proprement dit, de nombreux traitements doivent être effectuées pour obtenir un dictionnaire acoustique utilisable par un système de synthèse. Parmi ceux-ci, les tâches de phonétisation mais surtout de segmentation du corpus sont particulièrement critiques. En effet, même lorsque la chaîne phonétique correspondant à l'énoncé enregistré est connue, les méthodes de segmentation automatiques actuelles sont jugées trop peu précises pour pouvoir être utilisées telles quelles dans le processus de création de voix. Par conséquent, une étape de vérification manuelle de la segmentation demeure nécessaire. Cette étape, de loin la plus coûteuse, est un véritable frein à la diversification de voix dans le cadre de la SPC.

L'automatisation du processus de segmentation de la parole revêt donc une importance particulière pour la diversification de voix dans le cadre de la SPC. L'approche la plus répandue en segmentation de la parole et offrant les meilleurs résultats est celle reposant sur l'utilisation de modèles HMM [5]. Cette méthode peut être considérée comme contrainte linguistiquement car elle prend en entrée la chaîne phonétique, supposée exacte et obtenue par étiquetage manuel, correspondant à l'énoncé enregistré pour en déduire une séquence de modèles HMM. Elle consiste alors à effectuer un alignement forcé de cette séquence de modèles HMM sur le signal de parole. La principale limite de cette approche tient au fait que les HMM sont surtout réputés pour leur capacité à modéliser les zones stables des phones et non pas à détecter de manière fine des ruptures dans le signal de parole. D'autres approches telles que l'algorithme de Brandt [6] ont également été proposées pour localiser des ruptures dans le signal de parole. Ces méthodes sont a priori assez bien adaptées pour une tâche de segmentation, mais, n'étant pas contraintes sur le plan linguistique, elles produisent des omissions et des insertions de marques de segmentation.

Dans cet article, nous combinons ces deux types de méthodes. Dans cette optique, trois algorithmes de segmentation sont utilisés. Le premier est un algorithme classique de segmentation par HMM. Le deuxième est un algorithme d'ajustement des marques de segmentation par le biais d'une modélisation des frontières de phones par GMM [3]. Le troisième est une version modifiée de l'algorithme de Brandt de manière à rendre les marques de segmentation produites par cet algorithme compatibles avec la séquence phonétique réalisée. Ces différents algorithmes sont décrits en section 2. En section 3, nous présentons les différentes stratégies de combinaison envisagées pour ces algorithmes. Une expérimentation sur un corpus de parole dédiée à la synthèse vocale est également réalisée pour valider la méthode proposée.

2. Méthodes de segmentation

2.1. Segmentation par HMM

Les approches par HMM sont actuellement considérées comme un standard dans le domaine de la segmentation de la parole. Leur mise en œuvre requiert deux étapes : d'une part une phase d'apprentissage visant à estimer les modèles acoustiques et d'autre part une phase d'application de ces modèles à des fins de segmentation. L'étape de segmentation

*Cette étude est soutenue par France Télécom.

proprement dite revient à utiliser l'algorithme de Viterbi pour effectuer un alignement forcé entre la séquence de HMM correspondant à la séquence phonétique d'entrée et le signal de parole.

La phase d'apprentissage est cruciale, car la précision d'une segmentation est étroitement liée à la qualité de l'estimation des modèles. Nous utilisons ici une méthode classique basée sur une estimation initiale des modèles acoustiques par l'algorithme de Baum-Welch et suivie d'une procédure itérative au cours de laquelle les modèles sont réactualisés sur la base d'un alignement forcé [2]. Cette méthode d'apprentissage sera appliquée à l'ensemble du corpus de parole et dénommée dans la suite *HMM1*. En outre, nous nous proposons également d'étudier les performances d'un système de segmentation par HMM, lorsque les modèles acoustiques sont initialisés sur une partie de la base d'apprentissage pour laquelle une segmentation manuelle est disponible. Ces modèles sont ensuite utilisés pour segmenter tout le corpus. Une telle stratégie appelée par la suite *HMM2* offre généralement de meilleurs résultats puisque l'initialisation des modèles acoustiques est *a priori* meilleure [4].

2.2. Post-traitement par modèle de frontière

Dans [3], Wang et al. adjoignent à l'algorithme de segmentation par HMM un post-traitement effectué au voisinage des frontières de phones et utilisant des modèles GMM. Cette méthode de segmentation requiert une estimation préalable des modèles de frontières à partir d'un petit corpus segmenté manuellement. Plus précisément, étant donné une marque de segmentation manuelle, un super-vecteur est construit par la concaténation des N vecteurs acoustiques de taille N_c associés aux N trames de part et d'autre de la trame contenant la frontière, conformément à la figure 1. Lors de cette phase d'apprentissage supervisée, une classification des frontières par arbre de décision est opérée et pour chacune des classes obtenues, un modèle GMM est estimé. Une fois le modèle de frontière appris, le processus de correction consiste à déterminer, au voisinage de la marque de segmentation obtenue par alignement forcé, l'instant pour lequel la vraisemblance du super-vecteur par rapport au modèle de frontière considéré est maximale.

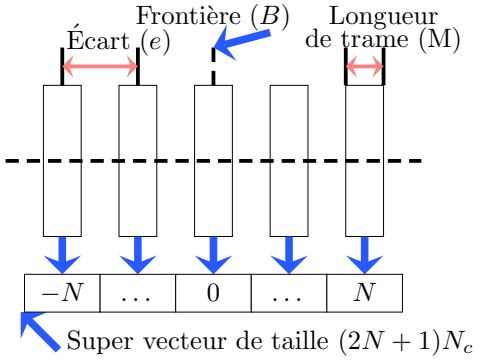

Fig. 1: Constitution d'un super-vecteur

2.3. Algorithme de Brandt

L'algorithme de Brandt [6] permet de détecter des ruptures de stationnarité dans un signal de parole. Il fait l'hypothèse que le signal de parole $y(n)$ est une suite de segments stationnaires et que le signal sur chacun de ces segments est modélisable par un modèle autorégressif (AR) : $y_n = \sum_{i=1}^{p} a_i y_{n-i} + e_n$, où p est l'ordre du modèle supposé constant pour tous les segments et où e_n est un bruit blanc gaussien de moyenne nulle et de variance σ^2. Par conséquent, chaque unité est associée à un vecteur de paramètres $\Theta = (a_1 = \ldots, a_p, \sigma)$.

Soit w_0 une fenêtre d'observation de longueur M. Le principe de l'algorithme de Brandt est de décider si w_0 doit être découpé en deux fenêtres w_1 et w_2 ou non. Cette décision se fait sur les vecteurs de paramètres Θ_1 et Θ_2 associés respectivement aux fenêtres w_1 et w_2. Ainsi, un changement entre Θ_1 et Θ_2 est détecté quand le rapport de vraisemblance généralisé (GLR) dépasse un seuil λ prédéfini. L'instant de ce changement est considéré comme l'instant de rupture de stationnarité. Comme mentionné précédemment, cet algorithme n'est pas contraint linguistiquement et par conséquent engendre des omissions et des insertions de marques de segmentation, dont les taux varient en fonction du seuil de détection choisi.

Dans le cadre de la segmentation de corpus dédiés à la synthèse vocale, nous disposons de l'information de phonétisation, voire d'une segmentation en accord avec cette séquence phonétique, obtenue par exemple via une méthode de type HMM. Pour une chaîne phonétique de taille L commençant et se terminant par un silence, notons $U = (U_0, U_1, \ldots, U_L)$ la séquence des marques de segmentation ainsi déterminée. A partir de ces marques de segmentation initiales, nous pouvons alors définir des intervalles temporels sur lesquels sont susceptibles de se produire les transitions entre les différents phones de la séquence phonétique. Ces intervalles sont de la forme $I_i = [V_i, V_{i+1}]$ avec $V_i = \frac{(U_{i-1}+U_i)}{2}$ pour tout i dans $\{1, \ldots, L-1\}$. Nous appliquons alors un algorithme de Brandt modifié sur chacun des intervalles I_i : la détection de rupture par seuillage du GLR est ici remplacée par la maximisation du GLR sur I_i, ce qui permet d'éviter tout risque d'insertion et d'omission.

2.4. Évaluation des algorithmes

Dans cette section, nous évaluons les différents algorithmes présentés ci-dessus sur un corpus de parole dédié à la synthèse vocale pour le français. Ce corpus comporte 7300 phrases prononcées par un sujet féminin et échantillonnées à 16 kHz. L'analyse acoustique et l'apprentissage des HMM sont effectués via l'outil HTK [1]. Les densités de probabilité d'émission, qui sont associées aux états, sont décrites par des lois multigaussiennes. Le nombre de gaussiennes est fixé à 2. Les vecteurs acoustiques sont de dimension $N_c = 39$ et contiennent 12 coefficients MFCC, l'énergie ainsi que les dérivées première et seconde de ces quantités. L'apprentissage de *HMM1* et *HMM2* est obtenu avec 20 itérations de l'algorithme de Baum-Welch. Pour *HMM1*, deux passes de

la procédure itérative ont été utilisées.

L'application des post-traitements par modèle de frontières et par l'algorithme de Brandt a été testée sur les segmentations produites par les deux algorithmes *HMM1* et *HMM2*. Les segmentations produites par post-traitements sur *HMM1* et *HMM2* sont appelées respectivement *Affin1* et *Affin2* tandis que les segmentations obtenues par l'algorithme de Brandt sont nommées *Br1* et *Br2*. Pour *Affin1* et *Affin2*, les paramètres N, M et e présentés sur la figure 1 sont fixés respectivement à 2, 20 ms et 30 ms. Pour *Br1* et *Br2*, l'ordre p des modèles AR est égal à 12 et la longueur minimale des fenêtres w_1 et w_2 est de 10 ms. Pour les algorithmes *HMM2*, *Affin1* et *Affin2* utilisant un apprentissage supervisé, une même partie de la base de données segmentée manuellement et constituée de 100 phrases choisies aléatoirement est utilisée. Afin de pouvoir effectuer une comparaison des différents algorithmes, la base de test considérée est le corpus complet privé de ces 100 phrases. La mesure de qualité choisie est le taux de segmentation correcte pour une tolérance égale à 20 ms, limite jugée acceptable pour garantir une qualité convenable de la voix synthétique.

D'après les résultats du tableau 1, l'algorithme *HMM2* apporte une amélioration significative par rapport à sa version non supervisée HMM1 ; le post-traitement par modèle de frontière semble être le plus performant ; enfin, l'algorithme de Brandt dégrade significativement les résultats des méthodes à base de HMM, qu'il soit appliqué sur les segmentations issues de *HMM1* ou de *HMM2*. Néanmoins, lors de ces expériences, nous avons pu constater que l'algorithme de Brandt parvient à bien localiser certains types de transitions (e.g. parole/silence, non-voisé/voisé). Il nous apparaît donc judicieux de tirer profit des performances de chaque algorithme selon les transitions à traiter. C'est sur la base de cette constatation que nous proposons dans la section suivante une approche par fusion de plusieurs segmentations.

Tab. 1: Taux de segmentation correcte à 20 ms pour chacun des algorithmes

HMM1	*Affin1*	*Br1*
88.29%	90.70%	84.50%
HMM2	*Affin2*	*Br2*
91.85%	92.70%	83.24%

3. Combinaison de plusieurs segmentations

3.1. Principe

Dans cette section, nous proposons un mécanisme permettant de combiner les marques de segmentation produites par différents algorithmes. Le principe de la méthode est d'analyser les performances de chacun des K algorithmes candidats sur différentes classes de transitions phonémiques, de manière à favoriser, lors de l'étape de segmentation, certains algorithmes par rapport à d'autres en fonction des transitions phonémiques à traiter. Plus précisément, étant donné un ensemble $\{c_1, \ldots, c_T\}$ de T classes, il s'agit d'estimer les taux de segmentation correcte $\alpha_k(c_i, c_j)$ à 20 ms pour $(i, j) \in \{1, \ldots, T\}^2$ et $k \in \{1, \ldots, K\}$. Pour mener cette étape d'apprentissage, il est bien entendu nécessaire de disposer de la segmentation manuelle d'une petite partie du corpus.

Ces taux de segmentation correcte vont permettre de combiner ces algorithmes en fonction des transitions à traiter comme suit. Soit s une transition dont les contextes phonétiques gauche et droit sont respectivement $c_g(s)$ et $c_d(s)$. Notons $t_k(s)$ la marque de segmentation de la transition s obtenue par le $k^{\text{ième}}$ algorithme. Une première solution consiste tout simplement à choisir, pour une transition donnée, l'algorithme fournissant en moyenne la meilleure précision, ce qui revient à effectuer avec la terminologie de [7] une fusion dure de la forme :

$$\hat{t}_{dure}(s) = \frac{\sum_{k \in A} t_k(s)}{Card(A)} \qquad (1)$$

où A est l'ensemble des algorithmes k qui maximise la quantité $\alpha_k(c_g(s), c_d(s))$ avec $k \in \{1, \ldots, K\}$. Précisons que A ne se résume pas à un seul élément. En effet, si les transitions entre les classes i et j ne sont pas observées dans le corpus d'apprentissage, alors les taux $\alpha_k(i, j)$ ne sont pas définis. Dans ce cas, nous posons $\alpha_k(i, j) = 1$ pour tout k ; on a alors $Card(A) = K$ et l'équation (1) devient une simple moyenne des K marques de segmentation produites par les K algorithmes.

Une autre façon de procéder est d'opérer une fusion douce [7] entre les marques de segmentation issues de chacun des K algorithmes. Cela revient à déterminer l'instant de segmentation comme étant le barycentre suivant :

$$\hat{t}_{douce}(s) = \frac{\sum_{k=1}^{K} \alpha_k(c_g(s), c_d(s)) t_k(s)}{\sum_{k=1}^{K} \alpha_k(c_g(s), c_d(s))}.$$

Notons que si les poids $\alpha_k(i, j)$ sont égaux pour tout k, alors les fusions dure et douce sont équivalentes.

3.2. Expériences et résultats

Dans cette section, nous présentons les résultats obtenus en appliquant les deux stratégies de fusion présentées précédemment d'une part sur le triplet $S_1 = (HMM1, Affin1, Br1)$ et d'autre part sur $S_2 = (HMM2, Affin2, Br2)$. La fusion a été effectuée en considérant les 12 classes suivantes : plosives voisées, plosives sourdes, fricatives voisées, fricatives sourdes, voyelles orales, voyelles nasales, diphtongues, consonnes nasales, consonnes liquides, semi-voyelles, pauses et silences.

Tab. 3: Taux de segmentation correcte à 20 ms pour les différentes stratégies de fusion testées

	Fusion dure	Isobary-centre	Fusion douce	Fusion optimale
$S1$	93.10%	92.80%	93.41%	93.68%
$S2$	93.53%	94.11%	94.65%	94.71%

L'estimation des taux de segmentation correcte utiles pour les fusions dure et douce est faite sur un corpus d'apprentissage constitué de 100 phrases différentes de celles utilisées pour l'apprentissage de *HMM2*, *Affin2* et *Affin1*. Les méthodes de fusion sont ensuite

Tab. 2: Pouvoir de correction des 3 stratégies de fusion

Position des marques en cas d'erreur d'au moins un algorithme	Fréquence d'occurrence	Taux de correction après fusion dure	Taux de correction après fusion isobarycentre	Taux de correction après fusion douce
3 marques du même côté	20.35%	51.25%	48.50%	50.57%
2 marques du même côté	8.11%	79.14%	95.07%	95.71%

évaluées en calculant les taux de segmentation correcte à 20 ms sur le corpus de 7300 phrases privé des phrases utilisées d'une part pour l'apprentissage supervisé des algorithmes et d'autre part pour la détermination des fonctions de fusion douce et dure. Nous comparons également les algorithmes de fusion proposés à deux autres méthodes de fusion : la première dénomée isobarycentre consiste à faire une simple moyenne des instants de segmentation fournis par chacun des 3 algorithmes ; la seconde est une fusion douce optimale en ce sens que les taux de segmentation correctes $\alpha_k(c_i, c_j)$ ont été estimés sur l'ensemble du corpus.

Les résultats consignés dans le tableau 3 montrent tout d'abord que, quelle que soit la stratégie de fusion employée, le taux de segmentation correcte après fusion est toujours supérieur à celui fourni par le meilleur des algorithmes impliqués dans la fusion (*Affin1* et *Affin2*). Par exemple, le taux de segmentation correcte passe de 92.70% pour *Affin2* à 94.65% dans le cas d'une fusion douce entre *HMM2* , *Affin2* et *Br2*, soit une réduction du taux d'erreur de 27%. En outre, la fusion douce se révèle être globalement la plus performante. Elle permet notamment d'améliorer significativement les taux obtenus par la méthode de l'isobarycentre, ce qui valide ainsi l'intérêt de l'apprentissage des $\alpha_k(c_i, c_j)$ opéré. Notons enfin qu'un apprentissage réalisé sur l'ensemble du corpus ne conduit pas à une augmentation très sensible des taux de segmentation correcte.

Le tableau 2 permet d'analyser plus finement le comportement des algorithmes de fusion dans deux configurations. La première correspond au cas où au moins un des algorithmes produit une erreur supérieure à 20 ms et que les trois marques de segmentations estimées sont situées du même côté par rapport à la marque de segmentation manuelle. Dans une telle configuration qui concerne 20.35% des transitions, la marque optimale correspond à celle fournie par l'algorithme ayant produit l'erreur la plus faible et par conséquent une stratégie de fusion dure serait *a priori* plus adaptée. Cependant, les pouvoirs de correction des méthodes de fusion dure et douce sont équivalents et légèrement supérieurs à la fusion par l'isobarycentre. Ceci illustre que dans un tel cas il est difficile de déterminer de manière fiable l'algorithme de segmentation le plus adapté. En revanche, lorsqu'une erreur se produit et que les trois marques de segmentation sont situées de part et d'autre de la marque manuelle, le pouvoir de correction des différentes stratégies de fusion est nettement amélioré. Dans cette deuxième configuration qui représente 8.11% des transitions, les taux de correction des stratégies de fusion douce et isobarycentre sont respectivement de 95.71% et de 95.07%, ce qui montre l'intérêt de procéder à un calcul barycentrique. La stratégie de fusion dure, bien que moins adaptée dans ce cas, parvient tout de même à résoudre 79.14% des erreurs contre 51.25% pour la première configuration.

4. Conclusion

Dans cet article nous avons étudié les performances relatives de trois méthodes de segmentation : l'une globale basée sur le formalisme des HMM et les deux autres locales visant à détecter une rupture au voisinage d'une marque. Nous avons de plus proposé deux stratégies permettant de fusionner les marques de segmentation issues des différents algorithmes. Ces stratégies ont été évaluées sur un corpus de parole dédié à la synthèse vocale et conduisent à une amélioration très sensible des taux de segmentation correcte à 20 ms. Ces méthodes semblent donc prometteuses et seront validées prochainement sur d'autres corpus et pour d'autres langues. Des travaux futurs seront également menés pour traiter le cas où seule une chaîne phonétique approchée obtenue de façon automatique est disponible.

5. Remerciements

Nous remercions Toufic Chmayssani pour ses contributions dans la mise en œuvre de l'algorithme de Brandt.

Références

[1] *The HTK book for HTK V3.0.* 2001.

[2] Y.J. Kim and A. Conkie. Automatic segmentation combining an hmm-based approach and spectral boundary correction. *ICSLP 2002,Colorado*, September 2002.

[3] L.Wang, Y. Zhao, M. Chu, J. Zhou, and Z. Cao. Refining segmental boundaries for tts database using fine contextual-dependent boundary models. *ICASSP*, vol.I :641–644, 2004.

[4] J. Matousek, D. Tihelka, and J. Psutka. Automatic segmentation for czech concatenative speech synthesis using statistical approach with boundary-specific correction. *Eurospeech*, 2003.

[5] S. Nefti. *Segmentation automatique de la parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance.* PhD thesis, Université de Rennes I, 2004.

[6] R.A. Obrecht. A new statistical approach for the automatic segmentation of continous speech signals. *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.36 :29–40, January 1988.

[7] S. Pigeon. *Authentification multimodale d'identité.* PhD thesis, l'Université Catholique de Louvain, 1999.