# Modèles GMM et algorithme de Brandt pour la correction de la segmentation de la parole par HMM

Safaa JARIFI[1], Dominique PASTOR[1], Olivier ROSEC[2]

[1]Département Signal et Communication
ENST Bretagne, Technopôle Brest-Iroise, 29238 Brest, France

[2]France Télécom, Division R&D TECH/SSTP/VMI
2, avenue Pierre Marzin, 22307 Lannion Cedex, France
safaa.jarifi@enst-bretagne.fr, dominique.pastor@enst-bretagne.fr
olivier.rosec@francetelecom.com

**Résumé –** On compare les performances de deux algorithmes de segmentation automatique. Le premier, nommé "HMM amélioré", affine la segmentation produite par les modèles de Markov cachés (HMM). Le deuxième est l'algorithme de Brandt qui vise, quant à lui, à détecter les ruptures de stationnarité. Le premier algorithme requiert la connaissance a priori de la phonétisation, le second non. Étant donné que l'algorithme de Brandt commet des insertions et des omissions, ce qui n'est pas le cas du HMM amélioré, on introduit une généralisation du taux de segmentation correcte (TSC) afin de comparer ces deux algorithmes. Les mesures expérimentales des TSCs permettent d'évaluer une limite supérieure des performances de l'algorithme de Brandt et suggèrent de combiner ces deux méthodes avec d'autres algorithmes adaptés à la séparation des classes acoustico-phonétiques.

**Abstract –** We compare the performance of two automatic segmentation algorithms. The first one is the so-called "refined HMM" and aims at refining the segmentation performed by Hidden Markov Models (HMM). The second is the Brandt's GLR (Generalized Likelihood Ratio) algorithm. It detects speech signal discontinuities. The first method assumes the knowledge of the phonetic sequence whereas in the Brandt'GLR method, this constraint is not used. The refined HMM yields no insertion and no omission in contrast with the Brandt's GLR method. Hence, we generalize the notion of correct segmentation rate (CSR) so as to compare these algorithms. The experimental results given in this paper exhibit an upper limit for Brandt's GLR method CSRs and suggest in combining the two methods with other algorithms adapted to the detection of boundaries between known acoustic classes.

## 1 Introduction

On s'intéresse à la segmentation automatique de grands corpus de parole spontanée et continue dédiés à la synthèse vocale. Dans l'approche classique, cette segmentation est automatisée par l'utilisation de modèles de Markov cachés (HMM) : après apprentissage de ces modèles sur l'ensemble du corpus de parole, la segmentation consiste à apposer les frontières de phones à l'aide d'un alignement forcé à partir d'un étiquetage phonétique supposé exact. Cette technique offre des résultats acceptables mais la précision de la segmentation peut parfois être insuffisante, entraînant des imperfections en sortie du système de synthèse. Des étapes de vérification, fastidieuses et coûteuses, demeurent indispensables. Dans le but de les réduire, nous cherchons à mettre en oeuvre des algorithmes de segmentation automatique dont la qualité approche celle de la segmentation manuelle. Dans cette perspective, ce papier a pour objectif de comparer deux algorithmes de segmentation automatique : le HMM amélioré et l'algorithme de Brandt.

Le premier, introduit dans [7] pour la segmentation d'un corpus chinois, consiste à affiner la segmentation produite par HMM en utilisant des modèles GMM (Gaussian Mixture Model) au voisinage des frontières phonétiques. Cet algorithme, parce qu'il est basé sur le HMM avec alignement forcé, produit un nombre de marques de segmentation égal au nombre de phonèmes présents dans la séquence phonétique.

Le second algorithme, originellement décrit dans [1, 2] et analysé dans [5], détecte les discontinuités locales de la stationnarité dans un signal de parole sans connaissance a priori sur la séquence phonétique de la phrase. Par conséquent, on n'échappe pas à la présence d'insertions et d'omissions dans la segmentation.

Pour ces raisons, on propose, après une description de chaque algorithme, une généralisation du taux de segmentation correcte (TSC). Cette généralisation a pour but d'évaluer les performances en termes de précision d'un algorithme de segmentation en prenant en compte ses insertions et ses omissions. La section 5 décrit l'ajustement des paramètres du HMM amélioré pour un corpus français et les TSCs des deux algorithmes en comparaison avec le HMM classique. Enfin, sur la base de ces résultats, on propose une approche qui vise à approcher la segmentation obtenue par HMM amélioré de la segmentation manuelle.

## 2 Le HMM amélioré

Le principe de cet algorithme est d'adjoindre à l'algorithme de segmentation par HMM un post-traitement utilisant des modèles GMM au voisinage des frontières phonétiques. Cette méthode s'effectue en deux étapes :

1. On associe un modèle GMM à chaque marque de segmentation d'un petit corpus d'apprentissage segmenté manuellement. En effet, pour accomplir cette association,

on crée d'abord un super-vecteur pour chaque frontière segmentée manuellement. Ce vecteur est obtenu en calculant les vecteurs acoustiques sur les trames autour de la frontière. La figure 1 représente un super-vecteur sur $(2N+1)$ trames. Chaque frontière $B$ dépend du phonème $X$ à sa droite et du phonème $Y$ à sa gauche. On obtient ainsi un pseudo-triphone noté $X-B+Y$ (voir figure 2). Puisque la taille du corpus d'apprentissage est petite, il est alors difficile de bien apprendre un modèle GMM pour chaque frontière. Pour cette raison, une étape intermédiaire avant l'apprentissage consiste à rassembler les pseudo-triphones par classes grâce à la création d'un arbre de régression et de classification (CART). Ensuite, un GMM est appris pour chaque classe.

2. Pour une phrase de test donnée, on cherche une marque de segmentation autour de chaque frontière obtenue par HMM; la marque retenue est alors celle qui maximise la vraisemblance par rapport au modèle GMM associé à cette frontière.

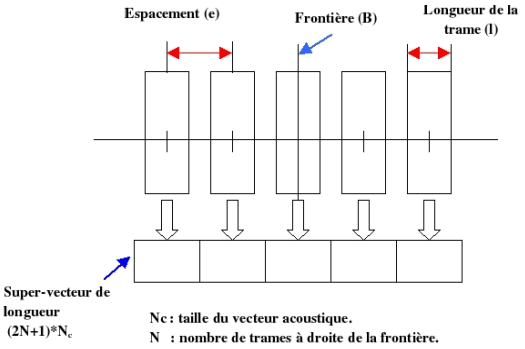


FIG. 1: *Exemple de super-vecteur pour une frontière donnée*

# 3 Algorithme de Brandt

L'algorithme de Brandt [5] est basé sur un critère statistique afin de décider s'il y a ou non une rupture de stationnarité dans le signal de parole. Il considère que le signal est une suite d'unités stationnaires puis modélise chaque unité par un modèle autorégressif (AR).

Soit alors $w = (y_n)$ une de ces unités. Chaque échantillon obéit donc à la loi de la forme : $y_n = \sum_{i=1}^{p} a_i y_{n-i} + e_n$ où $p$ est l'ordre du modèle, supposé constant pour toutes les unités. Le $n^{\text{ième}}$ échantillon de la fenêtre est $y_n$ et $e_n$ est un bruit gaussian de moyenne nulle et de variance égale à $\sigma^2$. Par conséquent, chaque unité est associée à un vecteur de paramètres $\Theta = (a_1 = \ldots, a_p, \sigma)$.

Soit $w_0$ une fenêtre de longueur $n$. Le principe de l'algorithme de Brandt est de décider si $w_0$ doit être découpée en deux fenêtres $w_1$ et $w_2$ ou non. Cette décision se fait sur les vecteurs de paramètres $\Theta_1$ et $\Theta_2$ associés respectivement aux fenêtres $w_1$ et $w_2$. Ainsi, un changement entre $\Theta_1$ et $\Theta_2$ est détecté quand le rapport de vraisemblance généralisé (GLR) dépasse un seuil $\lambda$ prédéfini. L'instant de ce changement est considéré comme l'instant de rupture de stationnarité.

# 4 Critères d'évaluation de performances

Les critères de comparaison des algorithmes de segmentation sont la probabilité d'insertion, la probabilité d'omission et le TSC. En particulier, le taux de segmentation correcte (TSC) mesure la précision d'une segmentation à une tolérance $\epsilon$ près par rapport à la segmentation de référence et ce en prenant en compte le nombre d'insertions et d'omissions.

TAB. 1: *TSCs de la segmentation produite par HMM*

| $\epsilon$ | 5 | 10 | 20 | 30 |
|---|---|---|---|---|
| *TSC* | 33.54 | 59.77 | 85.24 | 92.83 |

Soient $U = \{U_1, U_2, \ldots, U_n\}$ les instants des marques produites par un algorithme de segmentation automatique, et $V = \{V_1, V_2, \ldots, V_p\}$ les instants des marques de la référence. Nous construisons la liste $V_U = (V_{k_1}, \ldots, V_{k_n})$ tels que $k_j$ est l'indice dans $\{1, \ldots, p\}$ et $V_{k_j}$ la marque la plus proche de $U_j$. Donc chaque élément de $U$ est relié à un élément de $V$. Les omissions sont les instants $V_\ell$ qui n'appartiennent pas à la liste $V_U$, où $\ell \in \{1, \ldots, p\}$. Pour les insertions, on regarde les éléments répétés dans $V_U$. Si $V_U$ contient $m$ fois la même marque $V_\ell$ par exemple, le nombre d'insertions autour de $V_\ell$ est $m-1$. Supposons que les $m$ marques de $U$ correspondantes aux $m$ marques $V_\ell$ dans $V_U$ sont $(U_j, \ldots, U_{j+m-1})$. La marque la plus proche, de ce dernier vecteur, de $V_\ell$ est considérée comme la marque non insérée. Les autres $m-1$ marques représentent les insertions autour de $V_\ell$.

On applique cette technique pour tout $\ell$ tel que $V_\ell$ est répété dans $V_U$ afin de localiser toutes les insertions. Ainsi, une segmentation dont le nombre de segments est le même que pour la référence, peut contenir des omissions et des insertions. De plus, nous définissons les probabilités d'insertion et d'omission comme suit : $P_i = \frac{n_i}{p+n_i}$ et $P_o = \frac{n_o}{n+n_o}$ où $n_i$ est le nombre total d'insertions et $n_o$ est le nombre total d'omissions. Pour calculer le TSC, on identifie les marques qui ne sont pas des insertions au sens donné précédemment. Ensuite, on compte le nombre de ces marques qui se situent à moins de $\epsilon$ ms de la marque manuelle correspondante. Enfin, on divise ce nombre par la somme du nombre de marques de la segmentation manuelle et du nombre d'insertions. Autrement dit :

$$TSC = \frac{100}{p+n_i} \sum_{j=1}^{n} I_{[0,\epsilon]}(|V_{k_j} - U_j|)$$

où $I_{[0,\epsilon]}(x)$ est l'indicatrice de l'intervalle $[0, \epsilon]$. Un algorithme est donc considéré comme d'autant plus précis que son TSC est proche de 1.

# 5 Résultats

On utilise un corpus français de 7350 phrases prononcées par un seul locuteur masculin. Les signaux de parole résultants sont échantillonnées à 16 kHz et la phonétisation sur un corpus a été vérifiée manuellement.

## 5.1 HMM amélioré

L'objectif de cette section est d'ajuster les paramètres du HMM amélioré sur un corpus français. Nous partons des valeurs de paramètres ajustés sur un corpus chinois dans [7] et analysons si ces valeurs restent encore valables sur le corpus étudié. Seuls les paramètres d'apprentissage sont ajustés. La zone de recherche de la marque de segmentation autour de celle du HMM est fixée à 60 ms avec un pas de segmentation de 5 ms. Les TSCs dans cette section sont calculés pour

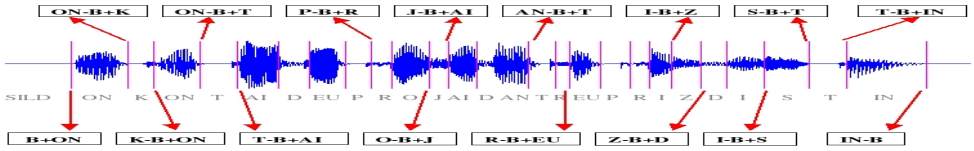


FIG. 2: *Exemples de pseudo-triphones*

TAB. 2: *TSC vs* $(T, MTI)$

| $T$ | $\epsilon$ | $MTI = 10$ | $MTI = 20$ | $MTI = 40$ | $MTI = 100$ |
|---|---|---|---|---|---|
| 20 | 5 | 38.49 | 38.08 | 37.48 | 35.13 |
| | 10 | 64.55 | 64.29 | 62.94 | 60.13 |
| | 20 | 87.91 | 87.74 | 87.05 | 85.16 |
| | 30 | 94.46 | 94.44 | 93.93 | 93.45 |
| 100 | 5 | 38.55 | 38.12 | 37.48 | 35.13 |
| | 10 | 64.63 | 64.28 | 62.94 | 60.13 |
| | 20 | 87.98 | 87.73 | 87.04 | 85.16 |
| | 30 | 94.50 | 94.46 | 93.93 | 93.45 |
| 350 | 5 | 38.38 | 37.90 | 37.49 | 35.13 |
| | 10 | 64.32 | 64.01 | 62.94 | 60.13 |
| | 20 | 87.93 | 87.51 | 87.11 | 85.16 |
| | 30 | 94.35 | 94.42 | 93.93 | 93.45 |

TAB. 3: *TSC vs taille du corpus d'apprentissage*

| *Taille* | $\epsilon = 5$ | $\epsilon = 10$ | $\epsilon = 20$ | $\epsilon = 30$ |
|---|---|---|---|---|
| 200 | 38.27 | 63.97 | 87.63 | 94.31 |
| 300 | 38.55 | 64.63 | 87.98 | 94.49 |
| 600 | 41.26 | 67.10 | 88.78 | 95.03 |
| 800 | 41.42 | 67.76 | 88.87 | 95.15 |

TAB. 4: *TSC pour différentes valeurs du* $(N, e)$

| $e$ | $N$ | $\epsilon = 5$ | $\epsilon = 10$ | $\epsilon = 20$ | $\epsilon = 30$ |
|---|---|---|---|---|---|
| 0 | 0 | 33.95 | 58.70 | 83.70 | 92.27 |
| 10 | 2 | 39.01 | 64.42 | 86.13 | 93.41 |
| 30 | 2 | 38.55 | 64.63 | 87.98 | 94.49 |
| 30 | 3 | 37.16 | 63.83 | 88.06 | 94.58 |

des tolérances $\epsilon \in \{5, 10, 20, 30\}$ ms, puis ils sont comparés à ceux du HMM classique (voir tableau 1).

Les paramètres d'apprentissage sont : la taille du corpus d'apprentissage, le nombre de trames sur lesquelles on calcule le super-vecteur $(2N + 1)$, l'espacement entre les trames $e$, la taille de la trame (20 ms), le nombre de gaussiennes pour le GMM (ici égal à 1), le nombre de coefficients par vecteur acoustique (égal à 39), et les critères d'arrêt dans le CART. Ces derniers sont : le nombre minimum d'éléments par nœud final $(MTI)$ et le seuil du log de vraisemblance $(T)$ pour associer une question à un nœud du CART.

Dans le tableau 2, le couple $(T, MTI)$ qui maximise le TSC est $(100, 10)$. Ce résultat est obtenu en fixant $(N, e)$ à $(2, 30)$ et la taille du corpus d'apprentissage à 300.

Le tableau 3 montre l'influence de la taille du corpus sur les TSCs. Le couple $(N, e)$ est encore fixé à $(2, 30)$ et $(T, MTI)$ est égal à $(100, 10)$ d'après ce qui précède. Ce tableau montre que 300 phrases sont suffisantes pour avoir un bon TSC.

Enfin, grâce au tableau 4 nous vérifions que la valeur égale à $(2, 30)$ du couple $(N, e)$ reste valable aussi pour le corpus français étudié. Ce choix semble être un bon compromis entre deux contraintes : la longueur du super-vecteur doit être suffisamment longue pour prendre en compte le plus d'information possible sur la transition entre les deux phonèmes ; cette longueur ne doit pas être trop longue afin de ne pas incorporer des informations qui ne sont pas liées directement à la frontière elle-même.

## 5.2 Algorithme de Brandt

Pour l'algorithme de Brandt, on choisit un ordre de modèle égal à 16 et un seuil égal à 30 comme dans [5]. En calculant les moyennes des probabilités d'insertion et d'omission de 4 tests utilisant 1200 phrases choisies aléatoirement, on trouve $P_i = 0.6$ et $P_o = 0.1$. la valeur élevée de $P_i$ est due aux oscillations du GLR. En fusionnant cet algorithme avec un détecteur d'activité vocale (DAV) [3], la probabilité d'insertion diminue significativement (de 0.6 à 0.3). En effet, la plupart des insertions sont localisées dans les silences (voir figure 3).

## 5.3 Comparaison

Les résultats que nous obtenons en comparant ces algorithmes à l'aide des critères ci-dessus sont synthétisés par la figure 4. Cette figure est obtenue en moyennant 4 tests croisés sur un corpus de 1200 phrases. Chaque phrase contient en moyenne une vingtaine de phonèmes. Les 5 courbes représentées sur cette figure sont les suivantes :

- **La courbe "Brandt" :** les TSCs obtenus par l'algorithme de Brandt sont très faibles en raison de la valeur élevée de $P_i$.
- **La courbe "Brandt+DAV" :** elle est obtenue en calculant les TSCs de la segmentation de Brandt après avoir éliminé les insertions dans les silences grâce au DAV suggéré dans [3].
- **La courbe "HMM" :** elle est obtenue par HMM avec alignement forcé et considérée comme la courbe de référence dans notre comparaison.
- **La courbe "HMM amélioré" :** elle résulte de l'utilisation du HMM amélioré ; les TSCs sont meilleurs que ceux du HMM standard.
- **La courbe "Brandt idéal" :** elle correspond aux TSCs calculés à partir de l'algorithme de Brandt après avoir éliminé les insertions grâce au traitement décrit dans 4 et remplacé les omissions par les marques correspondantes de la segmentation du HMM amélioré ; les TSCs sont meilleurs que ceux du HMM amélioré, ce qui signifie que les marques de segmentation de Brandt qui ne sont pas des insertions sont plus proches des marques manuelles que celles des autres algorithmes étudiés ici. En éliminant les insertions de l'algorithme de Brandt, il est théoriquement possible d'atteindre les TSCs de "Brandt idéal". Cependant, nous pensons qu'il est a priori très

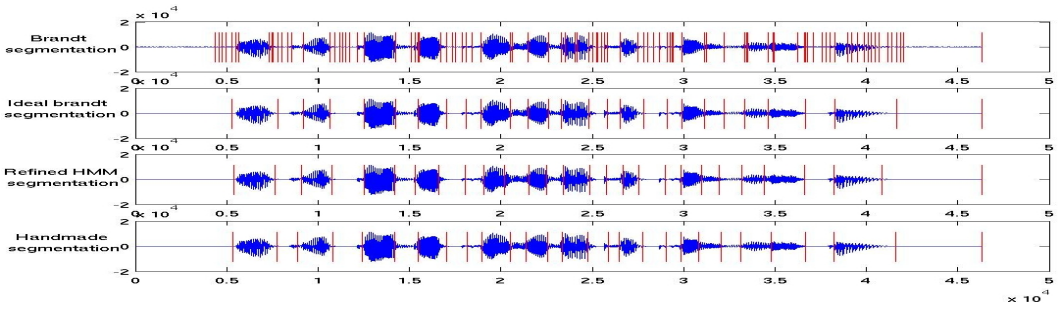


FIG. 3: *Segmentations manuelle, semi-automatique avec Brandt idéal, et automatiques avec l'algorithme de Brandt et HMM amélioré.*

difficile d'enlever les insertions dans les parties "parole" car ces insertions ne sont pas suffisamment bien localisées autour des marques de segmentation manuelle (figure 3).

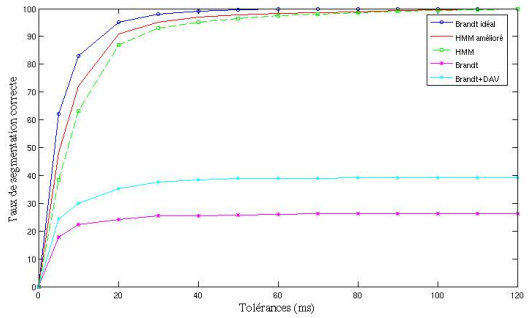


FIG. 4: *TSCs pour diffèrents algorithmes*

# 6 Approche dédiée à la synthèse vocale

D'après ce qui précède, on peut songer compléter l'algorithme de Brandt par des algorithmes permettant de réduire le nombre d'insertions et d'omissions et espérer ainsi approcher les performances de "Brandt idéal". La détection de voisement, l'analyse multi-résolution, la détection parole/non parole sont autant d'approches susceptibles de détecter des insertions et des omissions de Brandt. La difficulté à la laquelle on se heurte alors est que le contrôle des insertions et des omissions de la segmentation de Brandt est reporté sur les algorithmes complémentaires qui, eux aussi, commettent des erreurs de segmentation.

Il est donc plus raisonnable de partir d'une segmentation sans insertion ni omission (comme le HMM amélioré, voire le HMM classique) et d'essayer d'affiner cette segmentation à l'aide d'algorithmes dédiés. La méthode que nous proposons d'étudier est alors la suivante. On part de la segmentation obtenue par HMM amélioré. Pour chaque marque de cette segmentation, on calcule le GLR. Si ce GLR dépasse un certain seuil, alors on conserve la marque de segmentation du HMM amélioré. Sinon, on recourt à des algorithmes de segmentation acoustique adaptés aux classes acoustico-phonétiques des segments à identifier. Ces classes sont connues grâce à la phonétisation dont on dispose. De nombreux exemples d'algorithmes de segmentation acoustique sont décrits dans la littérature [6]. Parmi ces exemples, on s'intéressera particulièrement à la méthode dite de "Mallat-Modulation" décrite dans [4]. Cette méthode est une extension de l'algorithme de Mallat, adaptée à la détection d'énergie autour de fréquences arbitraires.

# 7 Conclusion

Après avoir décrit et comparé expérimentalement deux algorithmes de segmentation automatique, nous avons déduit une méthode de segmentation qui conjugue l'absence d'insertion et d'omission du HMM amélioré avec la précision des marques de la segmentation de Brandt, lorsque celles-ci ne sont pas des insertions. Notre travail futur consistera alors à mesurer les performances de cette méthode par rapport au HMM amélioré.

# 8 Remerciements

Nous remercions les reviewers pour leurs suggestions constructives que nous allons en prendre compte dans nos travaux.

# Références

[1] A.V. Brandt, *Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test*, Proc.ICASSP, pp.1017-1020, Boston, MA, 1983.

[2] A.V. Brandt, *Modellierung von Signalen mit Sprunghaft Veränderlichem Leistungsspektrum durch Adaptive Segmentierung*, Doctor-Engineer Dissertation, 1984, München, RFA (in German).

[3] S. Jarifi, D. Pastor, O. Rosec, *Jump and silence/speech detection for automatic continuous speech segmentation*, International Symposium on Image/Video Communications, Brest, France, 2004.

[4] C. Lemoine *Recherche de traits acoustiques de la parole bruité par analyse multi-résolution*, Thèse, Université de Bordeaux I, 1998.

[5] R.A. Obrecht, *A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals* , IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.36(1), pp.29-40, 1988.

[6] D.T. Toledano and L.A. Hernández Gómez and L. Villarubia Grande, *Automatic Phonetic Segmentation*, IEEE Transactions on Speech and Audio Processing, Vol.11(6), pp.617-625, 2003.

[7] L. Wang and Y. Zhao and M. Chu and J. Zhou and Z. Cao, *Refining Segmental Boundaries for TTS Database Using Fine Contextual-Dependent Boudary Models*, Proc. ICASSP, pp.641-644, Montreal, Canada, 2004.