Segmentation automatique de corpus de parole continue dédiés à la synthèse vocale

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

[1]Département SC GET-ENST Bretagne, Brest

[2]France Télécom R&D, Lannion

17 Novembre 2006

Plan

Contexte

- La segmentation est une tâche indispensable dans de nombreux systèmes de communication (reconnaissance vocale, synthèse vocale).
- Technique de segmentation automatique utilisée : chaînes de Markov cachées (HMM).
- Résultats acceptables mais ne garantissent pas une bonne qualité de la voix de synthèse.
- Erreurs ne dépassant pas 20 ms.
- Nécessité de vérification manuelle par des experts : tâche lourde et assez coûteuse.
- Limitation de la création de nouvelles voix.

Objectifs

- Réduire, voire éliminer le temps pris par la tâche de vérification manuelle.
- Mettre en œuvre une méthode de segmentation dont la précision à 20 ms:
  - est proche de la segmentation manuelle;
  - est meilleure que celle obtenue par HMM.

Approches possibles

Première approche

Réaliser une méthode de segmentation automatique indépendante des HMM $\longrightarrow$ segmentation plus précise que la segmentation par HMM.

Approches possibles

Première approche

Réaliser une méthode de segmentation automatique indépendante des HMM $\longrightarrow$ segmentation plus précise que la segmentation par HMM.

Deuxième approche : pré-traitement

Intégrer dans les vecteurs acoustiques mis en entrée du décodage de l'approche par HMM des informations provenant d'autres analyses.

Approches possibles

Première approche

Réaliser une méthode de segmentation automatique indépendante des HMM $\longrightarrow$ segmentation plus précise que la segmentation par HMM.

Deuxième approche : pré-traitement

Intégrer dans les vecteurs acoustiques mis en entrée du décodage de l'approche par HMM des informations provenant d'autres analyses.

Troisième approche : post-traitement

Partir de la segmentation par HMM et améliorer cette segmentation.

Approches possibles

Première approche

Réaliser une méthode de segmentation automatique indépendante des HMM $\longrightarrow$ segmentation plus précise que la segmentation par HMM.

Deuxième approche : pré-traitement

Intégrer dans les vecteurs acoustiques mis en entrée du décodage de l'approche par HMM des informations provenant d'autres analyses.

Troisième approche : post-traitement

Partir de la segmentation par HMM et améliorer cette segmentation.

Quatrième approche: fusion

Fusionner plusieurs méthodes de segmentation automatique.

Approches suivies dans la thèse

- Séquence phonétique connue et supposée correcte.
- Segmentation alignée sur cette séquence.
- Fusion de plusieurs segmentations produites par des méthodes prenant en compte la séquence phonétique.
- Application d'une mesure de confiance pour détecter les erreurs de segmentation.

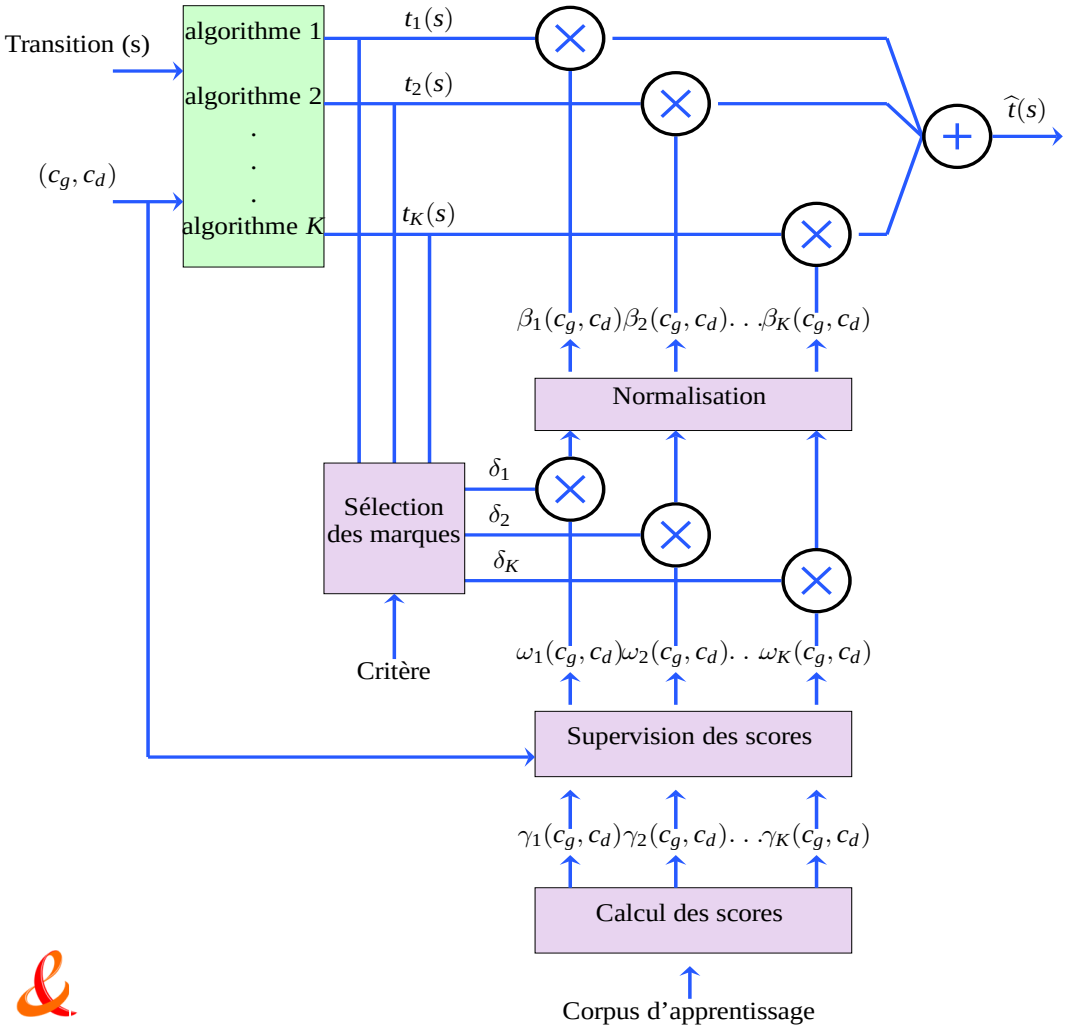
Fusion de plusieurs segmentations

- Fusionner des segmentations produites par des algorithmes différents.
- Tirer parti du meilleur de chaque algorithme.
- Soient K segmentations contenant le même nombre de marques.
- Il s'agit de faire une nouvelle estimation de la $n^{\text{ième}}$ marque entre deux classes de phonèmes (c_g, c_d) à partir des K marques obtenues par les K algorithmes.
- Combinaison linéaire des K marques :

$$\widehat{t}(s) = \sum_{k \in A} \beta_k(c_g, c_d) t_k(s), \tag{1}$$

où s est la transition entre c_g et c_d et $\beta_k(c_g, c_d)$ sont les poids.

Calcul des poids



- Scores : doivent évaluer les performances de chaque algorithme pour chaque classe de transition.
- Supervision des scores : fonction de pondération transformant les scores en poids.
- Sélection des marques : quelles marques seront fusionnées ?

Algorithmes choisis

Trois algorithmes

- Segmentation par HMM.
- Post-traitement par modèle de frontière.
- Algorithme de Brandt.

Particularités

- Méthodes complémentaires pour optimiser l'estimation : sont adaptées à détecter des types de transition différents.
- Une méthode globale, deux méthodes locales.
- Deux utilisant la connaissance *a priori* sur la phonétisation.
- Algorithme de Brandt ne prend pas en compte la phonétisation : omissions et insertions.

Évaluation du taux de segmentation correcte

- Deux corpus : français (corpusFR) et anglais (corpusEN).
- 12 classes de phonèmes pour le français.
- 11 classes de phonèmes pour l'anglais.

		Supervision			
		uniforme	*dure*	*douce*	
				$f(x) = x$	$f(x) = \frac{1}{1-x}$
100	*corpusFR*	93.68%	92.50%	94.08%	93.77%
	corpusEN	93.67%	92.35%	93.92%	93.77%
300	*corpusFR*	94.39%	93.83%	94.87%	94.92%
	corpusEN	94.36%	93.10%	94.67%	94.77%
700	*corpusFR*	94.59%	94.31%	95.09%	95.26%
	corpusEN	94.58%	93.81%	94.93%	95.17%

Résultats en taux de segmentation correcte

- Meilleurs résultats : supervision douce, inverse du taux d'erreur.
- Toutes les méthodes de fusion améliorent par rapport aux 3 algorithmes fusionnés.
- Réduction de 30% des erreurs (post-traitement par modèle de frontière).
- Réduction de 60% des erreurs (approche par HMM 88%).

Évaluation de la qualité de la parole

- Utilisation de tests subjectifs pour évaluer la qualité de la parole après synthèse vocale.
- Phase importante : satisfaction des utilisateurs.
- Test choisi (MOS) : Très mauvais (1) $\rightarrow$ Mauvais (2) $\rightarrow$ Satisfaisant (3) $\rightarrow$ Bien (4) $\rightarrow$ Très bien (5).
- Phase d'apprentissage pour les sujets naïfs.
- Évaluation sur un ensemble de textes contenant le plus d'erreurs de la segmentation par HMM.

Tests subjectifs : résultats

Amélioration de 50% la qualité de la parole grâce à la fusion en comparaison avec la segmentation par HMM pour le français.

		Nombre de sujets	*Score*	*Écart type*
Français	Segmentation par HMM	16	2.86	0.41
	Fusion douce		3.15	0.37
	Segmentation manuelle		3.35	0.4
anglais	segmentation par HMM	11	3.04	0.37
	Fusion douce		3.13	0.41
	Segmentation manuelle		3.06	0.44

Conclusions

- Proposition d'une approche générique pour la segmentation de grands corpus.
- Approche simple et efficace.
- Avec le choix de trois algorithmes complémentaires :
  - réduction de 60% des erreurs par rapport à l'approche standard par HMM;
  - amélioration de la qualité de la parole.

Perspectives

- Tests avec d'autres méthodes de sélection, d'autres fonctions de pondération.
- Tests subjectifs plus complets.
- Si la phonétisation est erronée, comment fait-on ?
- Détecter les erreurs de segmentation qui restent après la fusion.
- Tests sur d'autres corpus (multi-locuteurs, bruités...) $\rightarrow$ extension au domaine de la reconnaissance vocale.

Merci de votre attention...