HAL
archives-ouvertes.fr

Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance

Samir Nefti

▶ To cite this version:

Samir Nefti. Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance. Interface homme-machine [cs.HC]. Université Rennes 1, 2004. Français. tel-00122091

HAL Id: tel-00122091
https://tel.archives-ouvertes.fr/tel-00122091

Submitted on 26 Dec 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre: 2807

THÈSE

Présentée devant

l'Université de Rennes 1

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1
Mention : INFORMATIQUE

par

Samir NEFTI

Équipe d'accueil : Cordial
École Doctorale : Matisse
Composante universitaire : ENSSAT/IRISA

Titre de la thèse :

Segmentation automatique de parole en phones.
Correction d'étiquetage par l'introduction de mesures de confiance

soutenue le 16/12/2004 devant la commission d'examen

M. :	Jean-Paul	HATON	Président
MM. :	Louis-Jean	BOË	Rapporteurs
	Paul	DELEGLISE	
M. :	Olivier	ROSEC	Examinateur
MM. :	Olivier	BOËFFARD	Co-Directeur
	Marc	GUYOMARD	Directeur

Remerciements

Je tiens tout d'abord à remercier Olivier BOËFFARD pour son soutien et sa disponibilité sans failles.

Un grand merci aussi à Marc GUYOMARD.

Merci aux Messieurs les membres du jury d'avoir bien voulu examiner et juger ce travail.

Merci à France Télécom R&D, et plus particulièrement à l'équipe VMI à Lannion, dirigée par Thierry MOUDENC, qui a permis la réalisation de ce travail de thèse.

Merci enfin à tous ceux qui m'ont soutenu dans ce travail et qui n'ont pas été cités ici.

Table des matières

Introduction

Cette thèse concerne la segmentation automatique de parole en phones, en particulier pour les besoins de la synthèse de parole par concaténation d'unités acoustiques.

La synthèse de parole est une technologie dont l'usage connaît un essor important, pour répondre notamment aux besoins des services de télécommunication, tels que les services téléphoniques ou les services de messagerie électronique. Ces mises en service sont dues à une certaine confiance gagnée suite aux améliorations récentes de la qualité vocale qu'a connu cette technologie.

Actuellement, la synthèse de parole par concaténation d'unités acoustiques est la technique la plus efficace pour produire automatiquement de la parole. Il s'agit de générer un signal de parole synthétique par la mise bout à bout de portions de signaux de parole préalablement enregistrés. Ces portions de signaux de parole, appelées unités acoustiques, se trouvent dans une base de données, appelée dictionnaire d'unités acoustiques, qui est construite à partir d'un corpus de parole segmentée en éléments acoustiques ayant une cohérence linguistique, généralement phonétique.

Les diphones étaient, jusqu'à très récemment, les unités acoustiques par excellence de cette technique de synthèse de parole. Depuis peu, le concept des diphones a été élargi à des unités non-uniformes. Il s'agit alors de stocker dans le dictionnaire de synthèse de parole, des unités acoustiques de taille variable avec certaines de leurs variantes contextuelles et prosodiques. En effet, l'accroissement des capacités de stockage de données et le développement d'algorithmes de sélection automatique d'unités ont rendu possible l'exploration de cette technique. Pour atteindre un niveau de qualité suffisant, sa mise en œuvre exige un dictionnaire d'unités acoustiques très fourni, et par voie de conséquence, un corpus comportant plusieurs heures de parole.
L'avantage principal de cette technique est de réduire au minimum les post-traitements liés à la modification de la prosodie et à la concaténation des unités acoustiques.
L'inconvénient majeur de cette technique apparaît par exemple, lorsqu'il est nécessaire de créer une nouvelle voix de synthèse. Les techniques de conversion de voix actuels n'étant pas suffisamment au point pour permettre leur exploitation en situation réelle, la création d'une nouvelle voix de synthèse nécessite donc la répétition du processus d'enregistrement d'un corpus de plusieurs heures de parole, de segmentation de ce corpus en phones et de construction d'un nouveau dictionnaire d'unités acoustiques de synthèse de parole.

La segmentation phonétique du signal de parole peut être effectuée soit manuellement par un expert humain, soit automatiquement par une méthode programmée.
D'un point de vue qualitatif, l'examen de l'état de l'art de la segmentation de la parole donne la préférence à la segmentation manuelle. En effet, bien qu'il soit difficile d'évaluer la qualité d'une segmentation phonétique, il existe un large consensus sur le fait qu'une segmentation manuelle est plus précise qu'une segmentation automatique. D'autant plus que des logiciels disponibles, dotés d'interfaces graphiques conviviales et interactives, représentant le signal de parole et ses caractéristiques temporelles et fréquentielles, avec une sortie audio pour l'écoute, permettent à l'expert de générer, d'une manière relativement aisée, la séquence phonétique alignée sur le signal de parole.
Lorsqu'il s'agit de segmenter un corpus d'une dizaine ou d'une centaine de phrases, ce processus est envisageable. Mais les besoins permanents en corpus de parole segmentée et la taille grandissante de ces corpus éliminent d'office ce type de segmentation pour son coût exorbitant. Outre cet inconvénient, la segmentation manuelle souffre tout de même d'une variabilité inter et intra-segmenteurs (inconsistante et non-reproductible). De tous ces points découle l'intérêt majeur de la segmentation automatique de la parole.

Problématique

À condition de connaître précisément le contenu phonétique d'un corpus de parole à segmenter, les méthodes automatiques actuelles de segmentation de parole en phones, telles que les méthodes qui utilisent les modèles de Markov cachés, produisent une segmentation dont la précision approche celle d'une segmentation manuelle.
Cependant, la connaissance *a priori* du contenu phonétique d'un corpus de parole enregistré pour les besoins de la synthèse de parole par concaténation d'unités acoustiques, est en pratique une hypothèse peu réaliste pour deux raisons majeures :

- Un locuteur *standard* enregistré pour la constitution de ce corpus de parole, ne lit une phrase qu'à partir de sa représentation graphémique et non phonétique. Contraindre ce locuteur à le faire conduirait à des énoncés qui manqueraient de *naturel*.
- Un locuteur lisant une phrase d'une représentation graphémique, ne produit pas systématiquement un énoncé dont la description phonétique serait exactement celle produite par une phonétisation automatique de cette représentation graphémique.

Les méthodes de segmentation à base de modèles de Markov cachés permettent en effet de segmenter un énoncé de parole en phones, et ce en utilisant une transcription phonétique automatiquement produite par un système de phonétisation à partir de la représentation graphémique de cet énoncé.
Ce système de phonétisation peut fournir une transcription phonétique qui contient plusieurs prononciations possibles d'un texte, mais il ne peut pas fournir à coup sûr la séquence phonétique réellement énoncée. Il faudrait pour cela disposer d'un système de phonétisation qui donnerait toutes les variantes de prononciation de tous les locuteurs d'une langue donnée.

À cause du caractère imprévisible de la parole (variabilités intra et inter-locuteurs), ce système de phonétisation *exhaustif* est difficilement réalisable. Il est en outre difficile de faire respecter les locuteurs les standards de prononciation de la langue.
Par ailleurs, la transcription manuelle du contenu phonétique d'un corpus de plusieurs heures de parole souffre approximativement des mêmes problèmes que la segmentation manuelle d'un tel corpus.
Par conséquent, l'automatisation de la tâche de segmentation de parole en phones est fortement entravée par cette problématique d'indisponibilité des transcriptions phonétiques exactes des énoncés.

Contributions

Cette étude est principalement motivée par la résolution de la problématique posée précédemment. Pour cela, nous nous plaçons dans l'hypothèse réaliste d'un système de segmentation automatique de parole en phones qui traite un énoncé de parole et une transcription phonétique automatique du texte correspondant à cet énoncé. Afin de corriger les éventuelles erreurs d'étiquetage phonétique que peut contenir cette transcription automatique, nous proposons la méthodologie suivante :

- *Détecter les erreurs d'étiquetage phonétique :*
  La mise en œuvre efficace d'un système de segmentation construit sur l'hypothèse précédente passe par la détection des conflits d'association entre la description phonétique produite automatiquement et le signal de parole énoncé. Pour cela, nous proposons d'introduire en aval du processus de segmentation phonétique, un nouveau processus dont la tâche serait de détecter ces erreurs d'étiquetage et ce en utilisant une mesure de confiance acoustico-phonétique.
- *Corriger les erreurs d'étiquetage phonétique :*
  L'objectif étant de corriger ces erreurs d'étiquetage, nous proposons dans cette étude une méthode de correction automatique de ces dernières. Cette méthode consiste à mettre en œuvre un décodage acoustico-phonétique local des segments acoustiques alignés sur les étiquettes phonétiques rejetées par le processus de détection précédent.

Un autre objectif de cette étude concerne la réalisation d'un système de segmentation phonétique de parole le plus efficace possible en termes de précision temporelle des segmentations produites.
Pour atteindre cet objectif, nous formulerons l'hypothèse selon laquelle les transcriptions phonétiques exactes des énoncés sont connues et nous effectuerons quelques expériences qui permettront d'évaluer une technique d'analyse acoustique susceptible d'améliorer qualitativement les performances des méthodes de segmentation qui utilisent les modèles de Markov cachés.

Organisation du document

Dans une première partie, *"État de l'art de la segmentation phonétique de parole"*, nous présentons le cadre scientifique de notre étude ainsi qu'un état de l'art sur la segmentation de parole en phones.
Le chapitre 1, *"La synthèse de la parole"*, expose le cadre technologique ainsi qu'un état de l'art de la synthèse de la parole par concaténation d'unités acoustiques.
Le chapitre 2, *"Segmentation de la parole"*, développe une analyse des systèmes automatiques d'étiquetage et de segmentation de parole.
Enfin, le chapitre 3, *"État de l'art des mesures de confiance phonétiques et décodage acoustico-phonétique"*, présente principalement une synthèse des travaux sur le problème de la confiance dans l'annotation d'un signal de parole.

Dans une seconde partie, *"Mises en œuvre et évaluations"*, nous développons nos contributions directes à ce problème de segmentation automatique.
Le chapitre 4, *"Réalisation d'un système de segmentation markovienne de parole en phones"*, décrit la réalisation d'un système de segmentation par modèles de Markov cachés aux performances conformes à l'état de l'art.
Dans un chapitre 5, *"Détection des erreurs d'étiquetage phonétique"*, nous présentons une mesure de confiance efficace qui permet de caractériser les erreurs de transcription phonétique.
Pour finir, le chapitre 6, *"Correction des erreurs d'étiquetage phonétique"*, intègre cette mesure de confiance dans un système de segmentation complet et propose une solution pour la correction automatique des erreurs détectées.

Première partie

État de l'art de la segmentation phonétique de parole

Chapitre 1

La synthèse de la parole

1.1 Introduction

Afin de manipuler le signal de parole, pour la segmentation phonétique par exemple, nous devons maîtriser les concepts essentiels liés à la production et à la perception de la parole, à la phonétique et à la phonologie.
Ce chapitre fournit un rappel des connaissances essentielles qui décrivent les natures physique et phonétique de la parole, permettant par la suite de cerner la problématique de la segmentation de parole en phones.

Dans ce même chapitre, nous aborderons aussi le cadre technologique de notre étude. Nous présenterons ainsi un état de l'art de la synthèse de parole par concaténation d'unités acoustiques, actuellement la plus performante technique de synthèse de parole, et nous décrirons la fonction fondamentale qu'occupe la segmentation de parole dans le processus de préparation des bases de données d'unités de cette technique de synthèse.

1.2 Le signal de parole : de la production à la perception

Pour F. de Saussure [Duchet 1981], la parole s'oppose à la langue par son caractère concret, individuel et créatif. La parole est en effet la réalisation phonétique de la langue résultant d'un acte psychophysiologique et volontaire de la part d'un individu.
Aucun mode de communication animal ne peut égaler la complexité du langage parlé. Celui-ci est rendu possible par une anatomie particulière et par l'existence de régions spécialisées dans le cerveau.

1.2.1 La production de la parole

La parole est un phénomène acoustique qui se distingue des autres sons par des caractéristiques liées aux mécanismes de sa production par l'appareil phonatoire. Ce dernier fait intervenir divers éléments : l'air, comme source d'énergie ; les cordes vocales, comme principal organe vibratoire ; la langue et les lèvres, comme organes vibratoires accessoires ; les cavités buccale et nasale, comme caisses de résonance ; et le système