EXTRACTION SEMI-AUTOMATIQUE DES MOUVEMENTS DU CONDUIT VOCAL A PARTIR DE DONNEES CINERADIOGRAPHIQUES

Julie Fontecave et Frédéric Berthommier
Institut de la Communication Parlée, INP Grenoble, France
E-mail : fonte, bertho@icp.inpg.fr

Introduction

La radiographie permet l'observation d'une vue sagittale complète des articulateurs de la parole, de la glotte jusqu'aux lèvres. Grâce à la cinéradiographie, les mouvements du conduit vocal peuvent être étudiés avec une résolution temporelle importante. De grandes bases de données cinéradiographiques sont préservées et disponibles pour la communauté scientifique. Face à la quantité de données, l'extraction manuelle d'information géométrique est difficile à partir de telles bases.

Nous proposons une méthode d'extraction semi-automatique d'information géométrique (articulateur par articulateur, séquence par séquence), combinant une étape manuelle limitée puis une extraction automatique des mouvements du conduit vocal sur toute la base traitée.

1. Méthode quasi-automatique d'extraction de mouvements

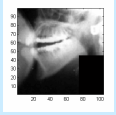
Mise en place de la méthode pour les mouvements de la **langue** à partir de la base de données **Wioland**

Base de données vidéo : Données cinéradiographiques du conduit vocal
5673 images (720*540 pixels) provenant de 64 séquences vidéos enregistrées à 66 im/sec, (phrases prononcées par une locutrice française)

Vidéo

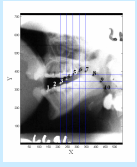
Calcul des 575 (24*24 – 1) **coefficients DCT** basses fréquences pour chaque image, restreinte au préalable à un cadre minimal d'observation (pour limiter les interférences)

DCT



100 Images Clefs choisies aléatoirement

Marquage manuel des images clefs
Description de la position du contour de la langue avec 10 points (12 degrés de liberté, ddl)



INDEX
assigné pour chaque image de la base, correspondant à l'index de l'image clef la plus proche
Mesure de similarité = distance euclidienne entre les coefficients DCT des deux images

Caractéristiques géométriques de la base complète
Marquage de la base entière en associant l'information géométrique disponible uniquement pour les images clefs

Géométrie

+ Traitements postérieurs
- Filtrage temporel
- Moyennage de configurations voisines obtenues par multi-indexation
- Lissage spline appliqué sur les points estimés

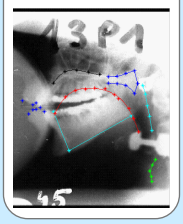
=> Reconstruction complète du mouvement
Erreur moyenne de reconstruction évaluée à 11 pixels (ou 3mm)

2. Extensions à d'autres articulateurs et à d'autres séquences cinéradiographiques

Pour chaque articulateur :
- Découpage des images d'origine pour inclure uniquement l'élément à marquer (= choix du **cadre**)
- Définition indépendante des **paramètres** (nombre d'images clefs, degrés de liberté, nombre de coefficients DCT nécessaires pour l'indexation)

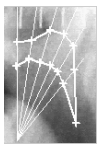
\+ **Adaptations** pour tirer profit des particularités des différentes bases

Wioland : conduit vocal complet

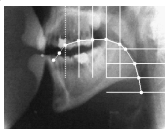


Base de données Flament

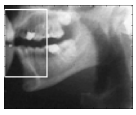
Près de 5000 images (720*540 pixels), enregistrées dans des conditions proches de Wioland (66 im/sec)



Vélum bien visible
Corpus dédié aux nasales du français
13 points / 14 ddl
100 images clefs



Langue
200 images clefs
9 points à 1ddl et 2 points à 2 ddl



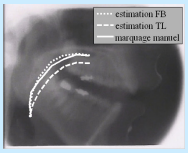
Double marquage pour améliorer la capture des mouvements rapides de la pointe – Fusion par substitution
- 1 marquage global pour le dos et la base
- 1 marquage spécifique pour la pointe (5ddl) à partir d'un cadre spécifique

Erreur moyenne de reconstruction du contour de la langue évaluée à 10 pixels

3. Base cinéradiographique d'ATR, séquence Laval43 : un comparatif sur la langue

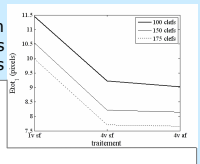
Base de données ATR (Munhall, Bateson, Tohkura)
25 films (Rochette, Perkell, Stevens)
55 minutes, près de 100000 images

Thimm et Luettin (IDIAP, 1999)
Technique de normalisation d'histogrammes + méthode d'extraction de contours (détecteur de Canny)
Données et résultats récupérés sur internet

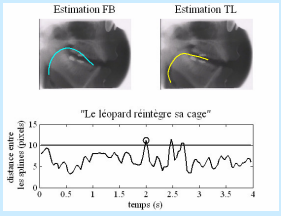


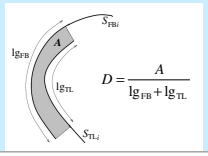
Fontecave et Berthommier (ICP)
Technique de rétro-marquage avec 200 images clefs et 13 ddl pour la langue

Erreur de reconstruction (méthode FB) avec différents traitements postérieurs



Exemple de décrochage entre les 2 méthodes
(D >10 pixels)





Différence moyenne ***D*** entre les 2 estimations => **6.8 pixels**

Comparaison limitée au dos et à la base de la langue
Pointe négligée car souvent manquante
(difficulté d'estimation par une approche contour)

Erreurs de reconstruction, mesurées sur un jeu d'images tests et sur 8 degrés de liberté définis par la méthode FB, entre les marques manuelles et celles estimées par chacune des 2 méthodes
TL => 20 pixels **FB => 8 pixels**

Conclusions et Perspectives

Méthode mise en œuvre avec succès sur plusieurs séquences => Extraction des mouvements du tractus vocal
Calcul de la fonction d'aire et mise en correspondance de nos mesures avec les caractéristiques temporelles et spectrales de la parole
Possibilité d'extension à toute la base de données ATR (méthode applicable par séquence de 2-3 minutes)

Remerciements : Nous remercions Pascal Perrier pour les films cinéradiographiques numérisés Wioland et Flament et Kevin Munhall pour la version de la base ATR sur DVD