Analyse de la Parole avec Contraintes de Production

Résumé de Thèse de Doctorat, par **Sacha Krstulović**

Décembre 2001

Les techniques de l'état de l'art en analyse de la parole et extraction de paramètres caractéristiques reposent principalement sur des modèles simplifiés de l'appareil auditif (par ex. avec l'échelle Mel ou les paramètres PLP). Ces sytèmes peuvent modéliser n'importe quel son, et ne sont pas particulièrement spécialisés à la modélisation de la parole. Ainsi, ils échouent à refléter des caractéristiques spécifiques à la parole, telles que la co-articulation.

Pour combler cette lacune, nous proposons d'intégrer certains paradigmes de production de la parole aux technologies du Traitement Automatique de la Parole (TAP). Cette intégration est basée *(1)* sur une analogie entre la Prédiction Linéaire (LP) et les modèles acoustiques de tubes sans pertes, et *(2)* sur le fait que plusieurs des modèles de l'état de l'art en production de parole emploient un tube comme interface entre le niveau acoustique et celui de la stratégie de production. Dans ce cadre, nous développons deux méthodes innovantes pour l'extraction de paramètres caractéristiques : l'analyse à topologie non-uniforme (NUT en anglais), et une méthode reliant l'acoustique à un modèle linéaire de forme de conduit vocal (ReALiSM en anglais).

Pour établir l'**analyse NUT**, nous commençons par généraliser l'équivalence traditionnellement établie entre la prédiction linéaire et les modèles de tubes sans pertes au cas où les tubes sont discrétisés en sections de longueurs inégales, comme dans le cas du modèle de production à régions distinctives (DRM).

Nous montrons que l'imposition de sections inégales est équivalente à contraindre certains des coefficients de réflection du filtre associé à garder une valeur nulle. Cette contrainte de "Topologie Non-Uniforme" permet de dé-coupler le nombre de degrés de liberté (DdLs) du modèle de la dimension de sa contrepartie acoustique (donnée par le nombre de pôles). Pour utiliser ce nouveau modèle en analyse, nous dérivons des estimateurs paramétriques adéquats, basés sur la minimisation analytique d'un critère d'erreur bien défini. Enfin, en remarquant qu'une topologie non uniforme fixe (telle que celle du modèle DRM) peut ne pas être optimale pour tous les constituants de la parole, nous proposons une méthode pour optimiser la répartition des longueurs/des retards de filtrage.

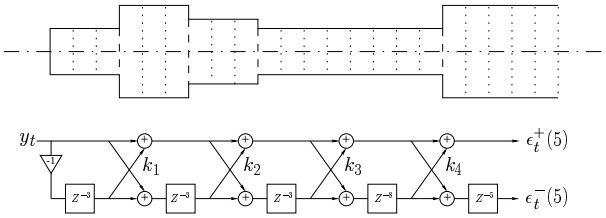
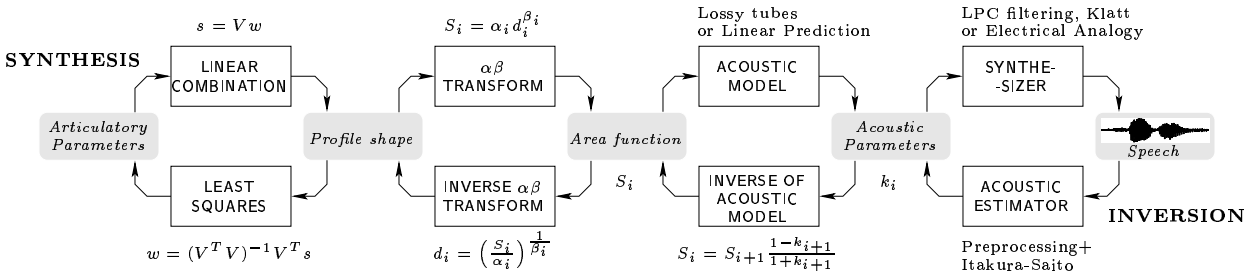


Fig.1 : Un tube NUT et le filtre en treillis équivalent.

La phase de validation montre que les modèles NUT permettent de réduire significativement le nombre de paramètres nécessaires à la description d'un spectre de parole, tout en gardant un haut degré de précision. Il est vérifié qu'ils produisent de manière consistante une erreur résiduelle moindre que les filtres non contraints à nombre de DdLs égal. Il est aussi vérifié que les modèles NUT sont en accord avec l'analyse spectrale grâce à l'optimisation de la topologie, qui permet de minimiser la distortion spectrale relative à la réduction du nombre de DdLs. En outre, les topologies peuvent en elles-même être utilisées comme un outil d'analyse.

Par ailleurs, pour établir la **méthode ReALiSM**, nous réalisons une projection de la solution du filtrage linéaire inverse vers l'espace des paramètres du modèle de conduit vocal de Maeda, à travers une séries de transformations linéaires et non-linéaires :



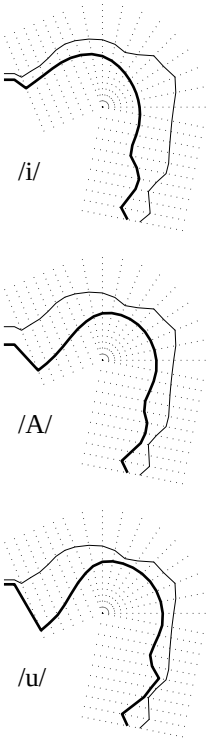


Fig.2 : Inversion de voyelles humaines.

La validation de ce système est réalisée en deux phases. Premièrement, il est vérifié que les pertes d'information relatives au lissage par moindres carrés, au ré-échantillonnage de la fonction d'aire et à l'estimation des coefficients de réflection permettent de retrouver des formes de conduit vocaux synthétiques imposées. Pour cela, un ensemble de voyelles synthétiques est produit (voyelles cardinales françaises de la base UPSID), et des tests informels d'écoute assurent qu'elles sont acceptables malgré l'approximation à longueur fixe et malgré le synthétiseur LPC sans pertes. Les résultats consécutifs montrent que l'inversion de ces voyelles produit des formes proches des formes synthétiques originales.

Dans une seconde phase, le système est utilisé pour inverser de la parole réelle enregistrée dans un environnement silencieux par un locuteur français. Plusieurs séquences de voyelles et de VCV sont testées. Par exemple, les résultats correspondants aux voyelles /i A u/ sont donnés dans la figure 2. Le système localise les cavités au niveau de lieux d'articulation raisonnables d'un point de vue phonétique (par ex., en avant pour le /A/, en arrière pour le /i/). Les ouvertures aux lèvres sont également réalistes, par ex., dans une séquence /A b i/, la fermeture consonnantique du /b/ est détectée. Ce système offre des avantages significatifs par rapport aux systèmes d'inversion acoustico-articulatoire existants, comme le calcul en temps réel, la modularité et des liens avec les techniques du Traitement Numérique du Signal.

Comme l'analyse par Prédiction Linéaire est à la base de l'état de l'art des techniques d'extraction de paramètres utilisées par la plupart technologies du TAP (telles que le codage, la reconnaissance de parole ou de locuteur, la synthèse, le débruitage etc.), les méthodes d'analyse que nous proposons, basées sur la production, créent une nouvelle passerelle pour l'**intégration de contraintes de production dans les principales applications du TAP**. Pour évaluer les bénéfices que nos nouvelles méthodes introduisent, nous proposons plusieurs manières d'exploiter les systèmes NUT et ReALiSM dans diverses branches du TAP. En particulier, nous fournissons et discutons des résultats préliminaires encourageants en reconnaissance de la parole.