# Séparation des sources d'informations caractéristiques du locuteur

Frédéric Chifflet
Institut Eurécom, Département Communications Multimédia
Mémoire de DEA effectué sous la direction d'Henri Méloni et de Philippe Gilles
Laboratoire Informatique d'Avignon

November 9, 1999

## Introduction

Notre travail s'inscrit dans le cadre de la Caractérisation du Locuteur et des enseignements que l'on peut en tirer pour l'Identification et la Vérification du Locuteur. Outre le message purement linguistique, la communication parlée véhicule de nombreuses autres informations, notamment sur l'identité du locuteur, son état de santé, son état émotionnel, etc. Les informations extra-linguistiques relatives à l'identité du locuteur sont celles que nous recherchons pour notre travail de caractérisation du locuteur. Elles proviennent de diverses sources qui sont :

1. les spécificités morphologiques du conduit vocal du locuteur (cordes vocales, conduits oral et nasal, viscosité du milieu...),
2. l'utilisation et le contrôle des articulateurs,
3. la réalisation des séquences de sons d'une langue (cibles phonémiques, coarticulation, intonation...),
4. les choix linguistiques effectués lors de la prononciation d'un message.

## 1 Problématique

Notre problématique consiste à vérifier que la modélisation de la contribution glottique que nous avons choisie d'étudier contient des informations caractéristiques du locuteur. Nous essayons de valider cette hypothèse en localisant et en quantifiant ces informations. Nous nous appuyons sur des connaissances de phonologie et de physiologie de l'appareil de production phonatoire.

## 2 Protocole expérimental

L'idée de base de notre protocole repose sur la comparaison de la pertinence des informations caractéristiques du locuteur présentes dans le signal de parole, à celle des informations présentes dans la contribution glottique à ce signal (i.e. l'onde glottique que l'on aura extraite de ce signal). Nous utilisons comme échelle de comparaison les résultats numériques fournis par le système de Reconnaissance Automatique du Locuteur développé au L.I.A. : le système A.M.I.R.A.L. (Architecture Multi-reconnaisseurs pour l'Indexation et la Reconnaissance Automatique du Locuteur).

Nous avons utilisé le corpus BREF80 élaboré par le L.I.M.S.I. (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur). C'est un corpus de 80 locuteurs prononçant des enregistrements de français lus dans Le Monde. Nous avons également à notre disposition l'étiquetage acoustico-phonétique de tous ces enregistrements. Nous travaillons sur un sous-corpus de 40 locuteurs.

Pour chaque locuteur nous avons deux corpus disjoints de signaux de parole, pour l'apprentissage et le test, et les contributions glottiques correspondantes. Elles sont calculées grâce à l'algorithme I.A.I.F.(Iterative Adaptive Inverse Filtering). Cet algorithme consiste à effectuer une succession de LPC et de filtrage inverse afin d'estimer la contribution du conduit vocal pour mieux l'ôter ensuite au signal de parole.

La phase de test permet d'obtenir une matrice de confusion $M$ où chaque locuteur $L_i$ (représenté par la concaténation de tous les signaux du corpus de test qui lui sont associés) est confronté à tous les modèles de locuteur $L_j$ appris. Chaque coefficient $M_{i,j_t}$ est une mesure de vraisemblance. C'est la vraisemblance que la trame $t$ prononcée par le locuteur $L_i$ soit reconnue par le modèle du locuteur $L_j$. Nous avons deux instances de la matrice $M$ : une matrice $P$ obtenue après la phase de test sur le corpus de parole et une matrice $G$ obtenue sur le corpus glottique, ces deux matrices sont normalisées.

$$\hat{M}_{i,j_t} = \frac{M_{i,j_t}}{\frac{1}{|locuteurs|}\sum_{k=1}^{|locuteurs|} M_{i,k_t}}$$

Nos expériences portent principalement sur les données fournies par le biais de ces matrices de confusion.

# 3 Expériences

Après avoir constaté que notre modélisation de la contribution glottique donnait des résultats similaires au signal de parole sur le système AMIRAL (100% d'identification), nous avons comparé pour chaque trame $t$, le rapport de vraisemblance du signal de parole avec celui de l'onde glottique.

$$Comparaison(t) = \hat{P}_{i,j_t} - \hat{G}_{i,j_t}$$

Ce calcul montre que les informations peuvent être portées avantageusement par l'onde glottique ou par le signal de parole, puisque cette différence est presque partout non nulle. Nous avons localisé trame à trame les informations. Il reste à dégager un critère de classification des trames, nous allons utiliser des critères issus de la phonologie. Ils sont à l'origine des expériences suivantes.

Disposant de l'étiquetage acoustico-phonétique, il nous semble intéressant d'étudier les informations caractéristiques du locuteur via les phonèmes. En effet, nous avons la possibilité d'avoir un lien entre les modes, les lieux articulatoires (principalement sons voisés et non voisés) et les rapports de vraisemblance. Nous allons nous munir d'un critère permettant de quantifier la pertinence des informations sur le locuteur qui sont transportées par un phonème donné.
Pour chaque phonème $x$ ,avec $t_x$ les trames où $x$ est prononcé, nous utilisons sur la matrice de confusion $\hat{M}$ ($\hat{P}$ ou $\hat{G}$) le critère d'émergence

$$E_{\hat{M}}(x) = \frac{\sum_{i=1}^{|locuteurs|}\sum_{t=1}^{|t_x|}\hat{M}_{i,j_t}}{|t_x||locuteurs|}$$

qui donne une mesure de la pertinence des informations caractéristiques du locuteur portées par le phonème $x$.

Nous constatons que si de l'information subsiste dans notre modélisation de l'onde glottique, il n'y a pas de phonème privilégié où cette information serait plus pertinente que dans le signal de parole. De plus, les écart-types sont trop importants pour pouvoir réellement comparer les moyennes entre elles.

La classification des occurences des trames $t_x$ en prenant un ou plusieurs critères de classification (locuteur, lieux et modes d'articulation, voisement) ne permet pas d'exhiber de classe de phonèmes dont l'information caractéristiques du locuteur serait significativement plus pertinente dans la contribution glottique que dans le signal de parole. Le reproduction de ces expériences avec le critère d'émergence

$$E_{\breve{M}}(x) = \frac{\sum_{i=1}^{|locuteurs|}\sum_{t=1}^{|t_x|}\breve{M}_{i,j_t}}{|t_x||locuteurs|}$$

| Phn | $E_{\hat{P}}(x)$ | $\sigma_{E_{\hat{P}}(x)}$ | $E_{\hat{G}}(x)$ | $\sigma_{E_{\hat{G}}(x)}$ | Phn | $E_{\hat{P}}(x)$ | $\sigma_{E_{\hat{P}}(x)}$ | $E_{\hat{G}}(x)$ | $\sigma_{E_{\hat{G}}(x)}$ |
|---|---|---|---|---|---|---|---|---|---|
| mm | 7.467 | 10.214 | 5.236 | 7.800 | yy | 5.658 | 9.391 | 5.045 | 8.610 |
| ai | 14.251 | 13.744 | 13.197 | 13.065 | Ott | 3.345 | 7.038 | 3.051 | 6.618 |
| ss | 3.271 | 5.584 | 3.807 | 6.700 | Btt | 1.727 | 3.569 | 1.707 | 3.696 |
| on | 7.569 | 10.601 | 5.688 | 8.425 | rr | 5.473 | 9.201 | 4.965 | 8.447 |
| dd | 4.480 | 8.080 | 3.763 | 7.063 | an | 9.339 | 12.137 | 6.966 | 9.672 |
| eu | 9.990 | 12.533 | 8.311 | 11.160 | uu | 4.718 | 8.063 | 4.003 | 7.005 |
| Okk | 3.356 | 7.250 | 3.171 | 6.836 | nn | 8.080 | 10.653 | 6.173 | 8.824 |
| Bkk | 1.518 | 3.357 | 1.449 | 3.325 | oe | 13.235 | 12.617 | 11.934 | 11.439 |
| ll | 6.345 | 10.212 | 5.838 | 9.636 | ff | 3.082 | 6.147 | 2.859 | 5.442 |
| oo | 11.649 | 13.343 | 9.438 | 11.564 | vv | 4.369 | 7.668 | 3.845 | 6.849 |
| ch | 2.301 | 4.941 | 2.332 | 4.981 | ou | 4.631 | 8.222 | 3.351 | 6.003 |
| ei | 13.254 | 13.593 | 12.130 | 12.888 | jj | 3.301 | 7.022 | 3.011 | 6.352 |
| ii | 8.705 | 11.546 | 7.689 | 10.597 | gg | 4.798 | 8.997 | 4.349 | 8.352 |
| in | 12.336 | 12.749 | 10.530 | 11.585 | un | 16.546 | 15.160 | 14.617 | 14.832 |
| Opp | 3.541 | 6.866 | 2.999 | 5.618 | bb | 4.289 | 7.589 | 3.453 | 6.172 |
| Bpp | 1.200 | 1.376 | 1.240 | 1.436 | au | 8.380 | 11.285 | 6.700 | 9.553 |
| uy | 3.532 | 6.703 | 3.034 | 6.093 | zz | 4.870 | 8.086 | 4.698 | 7.936 |
| aa | 14.192 | 13.548 | 12.991 | 12.931 | ww | 7.934 | 10.911 | 6.394 | 9.328 |

Table 1: Pertinence et écart-type de la pertinence des phonèmes pour la contribution glottique et le signal de parole

ne modifie pas les résultats, pas plus que l'utilisation de la stabilité cepstrale (on privilégie les trames où la phonation de $x$ est optimale).

## Conclusion et perspectives

Ce travail a constitué une première approche du problème de la séparation des sources d'informations caractéristiques du locuteur. Les résultats que nous obtenons montrent que de l'information subsiste dans l'onde glottique, cela confirme la redondance de informations contenues dans le signal de parole complet.

Outre le signal de parole et une modélisation de l'onde glottique, le protocole que nous avons mis en place pourrait être étendu à la réponse du conduit calculée à partir des deux signaux précédents. C'est un calcul qui pose des problèmes de synchronisation du signal et de recomposition du signal à partir de coefficients cepstraux ou LPC.

La prochaine étape de ce travail serait de chercher une meilleure modélisation de la contribution glottique. Dans cette optique, il nous paraît intéressant de mettre en place une étude utilisant un électroglottographe. Globalement, le principe serait d'enregistrer le signal de parole des locuteurs tout en mesurant l'onde glottique avec l'électroglottographe. Cela permettrait d'avoir un corpus de signaux glottiques de référence auxquels nous pourrions comparer les estimations de la contribution glottique que nous serons amenés à étudier.