

# Reconnaissance Automatique de Parole en Français pour Milieu Difficile: Exemple de Détection de Double Parole pour le Radiotéléphone en Mains Libres.

Hassan Ezzaidi\*, Jean Rouat\* et Ivan Bourmeyster+

\* ERMETIS, Université du Québec à Chicoutimi, Canada

+ Alcatel Mobile Phones, Paris, France

{hezzaidi,jrouat}@uqac.quebec.ca

Ivan.Bourmeyster@alcatel.art.fr

## Résumé

Nous présentons un système de détection de double parole qui utilise l'information de fréquence glottale combinée à différentes stratégies de filtrage inverse et de synthèse de filtres peignes afin de piloter un système d'annulation d'écho. La technique proposée ici permet de savoir quand le correspondant lointain est actif ou non afin de décider des moments au cours desquels le système d'annulation d'écho doit être adapté. Les meilleurs résultats sont obtenus lorsque l'annulation de la contribution glottale du locuteur lointain est effectuée avant et après la suppression des caractéristiques de son conduit vocale. De plus, la méthode de suppression des composantes spectrales qui pondère les harmoniques du fondamental avec une fonction proportionnelle à leur 'éloignement' donne les meilleures performances. Nous donnons les détails des structures et les techniques de suppression.

## Introduction

Il existe des systèmes de reconnaissance de parole indépendants du locuteur pour des vocabulaires limités et qui offrent des performances intéressantes à condition que l'environnement ne soit pas trop corrompu par le bruit et les interférences. Toutefois, l'interrogation de bases de données à distance, (téléphone, radio, etc.) nécessite de concevoir des systèmes polyvalents et peu sensibles aux conditions et aux environnements d'opérations ainsi qu'aux bruits ambiants (téléphone cellulaire, cabine téléphonique, voitures, avions, camions, etc.). Or le traitement de la parole en milieu bruyant est loin d'être résolu et les systèmes de reconnaissance actuels ne peuvent fonctionner dans de tels environnements sans observer de dégradation très significative des performances ce qui les rend à toute fin pratique non utilisables par le grand public.

Devant les difficultés rencontrées en reconnaissance de parole et la nécessité de commercialiser rapidement, la majorité des travaux ont porté sur de la parole "propre" libre de tout bruit. Beaucoup d'énergie a donc été placée dans la réalisation de systèmes pour lesquels la parole a été enregistrée en "laboratoire". Ceci a engendré un espoir un peu trop grand et trop rapide vis-à-vis de cette technologie et de son utilisation en conditions réelles (hors laboratoire). La méthodologie utilisée a donc biaisé la conception des systèmes de reconnaissance en assumant que le signal de parole est libre d'interférences et de bruits. Cependant nous avons conscience de ces limites et nous nous proposons de

travailler à partir de données enregistrées en conditions réelles afin de nous pousser à aborder le problème sous l'angle des conditions réelles et difficiles de fonctionnement. On peut considérer qu'il existe deux écoles de pensée à cet égard. Une qui prône le "nettoyage" et le "débruitage" préalable à la reconnaissance, ce qui a pour avantage de pouvoir ensuite utiliser l'arsenal d'outils déjà développés en reconnaissance de parole sur des données relativement propres. Et l'autre école de pensée qui assume qu'il n'est pas possible de "débruiter" à priori et que le système doit être capable de traiter directement les interférences et le bruit un peu de la même façon que le fait l'être humain. Cette dernière approche, quoique moins mature, a l'avantage de permettre la conception de systèmes plus versatiles. Nous pensons qu'en fait il y a lieu de comparer et d'évaluer les points forts et points faibles des deux tendances afin d'élaborer un système de reconnaissance qui pourra éventuellement être mixte.

Notre laboratoire s'intéresse à la reconnaissance de parole en français du Québec pour l'interrogation de bases de données à distance (par téléphone par exemple) en collaboration avec l'INRS-télécommunications, l'ETS et le CRIM à Montréal. Dans ce contexte, nous avons orienté une partie de nos travaux sur le traitement en milieu difficile et via le réseau téléphonique. Un scénario possible est de nettoyer le signal avant d'effectuer la reconnaissance. Dans ce contexte, le suivi de la fréquence fondamentale peut être intéressant pour améliorer les systèmes de reconnaissance dans les situations difficiles pour lesquelles plusieurs locuteurs interfèrent.

Nous décrivons un système de prétraitement qui permet de nettoyer le signal avant reconnaissance lorsque associé à des techniques d'annulation d'écho et de bruit dans le cas de double parole. Dans la situation présente, il s'agit de reconnaissance de parole dans le contexte de radiotéléphone mains libres en véhicule et en français. Le locuteur lointain (correspondant lointain) interfère avec le locuteur en véhicule via le haut-parleur du téléphone qui est installé dans le véhicule. De plus, le signal est corrompu par l'écho du à l'habitacle et par les bruits environnants. Un système de suppression d'écho et de bruit doit être adapté (nouveaux coefficients de filtre) à des instants bien précis en fonction de l'activité vocale du correspondant et du locuteur en véhicule. Or, il est très difficile de discriminer les deux locuteurs. Nous avons intégré un algorithme de détection de hauteur tonale basé sur une approche perceptive. Il permet

d'estimer la fréquence fondamentale sur le signal du haut-parleur et de synthétiser un filtre peigne pour diminuer la contribution glottale du locuteur lointain sur le signal du microphone.

À la section suivante nous décrivons le système de suivi de fréquence fondamentale. La section détection de double parole en radiotéléphonie mains libres montre de quelle façon nous utilisons une estimation de fréquence glottale afin de séparer deux locuteurs et de rehausser le signal.

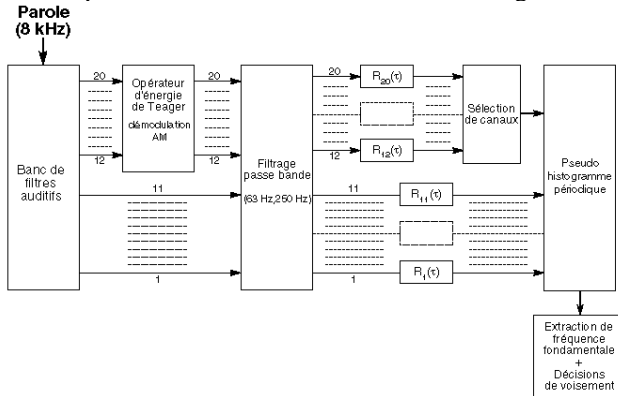


Figure 1. Schéma du système d'estimation de la fréquence glottale et de décision de voisement.

## Système de Détection de Fréquence Glottale et de Décision de Voisement

### Principe Général

Le système de suivi de fréquence glottale et de décisions de voisement est basé sur le fait que les modulations d'amplitude apparaissant à la sortie d'un banc de filtres cochléaires sont caractéristiques de la hauteur tonale. Nous faisons la distinction entre les harmoniques résolus par le système auditif (en basse fréquence) et les harmoniques non résolus qui créent des battements lorsque la largeur de bande des filtres cochléaires est suffisamment grande vis-à-vis de la fréquence glottale. La présence de ces battements permet au système auditif de percevoir la hauteur tonale même lorsque le fondamental est absent ou trop bruité (B. Moore, 1989).

Notre algorithme (J. Rouat et al., 1997) comprend trois modules (Fig. 1.). Le premier module est un banc de vingt filtres cochléaires (fréquences centrales de 330 Hz à 3700 Hz), le second traite les sorties des filtres afin de rehausser la période de modulation du signal glottique et de combiner les informations des canaux sélectionnés en un pseudo-histogramme périodique. Le troisième module estime la fréquence glottale et prend la décision de voisement.

### Description des deux Premiers Modules

Les onze premiers canaux (330 Hz - 1270 Hz) sont simplement filtrés passe-bande avant calcul de la corrélation normalisée entre la fenêtre centrée et cette même fenêtre décalée pour chaque canal  $i$  ( $R_i(\tau)$ ,  $i=1, 11$ ). Nous obtenons

ainsi onze représentations différentes du fondamental (lorsque présent) et des premiers harmoniques. Les neuf derniers canaux sont prétraités à l'aide de l'opérateur 'énergie de Teager' (J.F. Kaiser, 1990, 1993). Cette opération est équivalente à estimer le carré de l'enveloppe du signal pondéré par le carré de la pulsation instantanée (J. Rouat, 1993).

Lorsque le signal est très bruité (rapport signal à bruit de 0 dB et moins), l'algorithme peut utiliser une unité de sélection automatique des canaux. Cette unité permet de sélectionner les canaux pour lesquels le caractère harmonique du signal ressort suffisamment bien (J. Rouat et al., 1997). Pour les expériences reportées plus bas, nous n'avons pas utilisé cette technique de sélection automatique car les données enregistrées en véhicule n'étaient pas assez bruitées pour justifier cette augmentation de complexité.

Le pseudo-histogramme périodique (Fig. 1.) est noté PPH et est obtenu en réalisant la somme à travers les canaux sélectionnés des corrélations normalisées.

$$PPH(\tau) = \frac{1}{M} \sum_{i=1}^M R_i(\tau)$$

M est le nombre de canaux qui contribuent à la fréquence glottale. Ici  $M=20$ .

### Décision de Voisement et Estimation de la Fréquence Glottale

Les deux plus grands pics 'éligibles' dans  $PPH(\tau)$  sont sélectionnés. Pour être 'éligible', un pic doit être plus grand qu'un seuil  $S$  prédéterminé. Si on ne trouve pas deux pics vérifiant ces conditions, on déclare que le segment est non voisé. On suppose que les deux pics correspondent à des valeurs de  $\tau_1$  et  $\tau_2$  respectivement à  $\tau_1$  et  $\tau_2$  sont des multiples (ou l'un d'eux peut être égal) de  $T$  qui est la période fondamentale.  $T$  est donc un des sous multiples de  $\tau_1$  et  $\tau_2$ .

On cherche les sous-multiples et  $T$  est associé au plus petit sous multiple qui correspond à un pic vérifiant la relation:

$$PPH(T) \geq S_{pe} \cdot \text{Max} [PPH(\tau_1); PPH(\tau_2)]$$

avec  $S_{pe} = 0.5$ .

Si l'algorithme ne trouve pas  $T$ , le segment est déclaré comme étant possiblement non voisé, sinon il est déclaré voisé avec une fréquence égale à  $1/T$ .

Ces informations sont ensuite traitées afin de prendre la décision finale et définitive de voisement et de valeur de fréquence glottale. (J. Rouat et al., 1997).

L'algorithme a été adapté afin de pouvoir fonctionner en temps réel dans le contexte de la détection de double parole tel que décrite ci-dessous.

## Détection de Double Parole en Radiotéléphonie Mains Libres



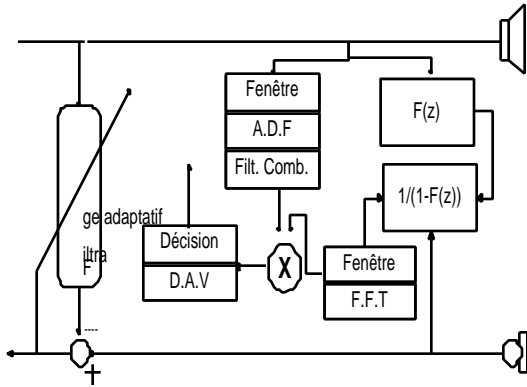


Figure 3: structure proposée

Cette technique est utilisée dans le but de rehausser les performances du système de référence. Elle inclut l'estimation de la fréquence glottale seulement du signal du locuteur lointain (voir figure 3). Dans un premier temps, la contribution vocale du locuteur est paramétrisée par la méthode classique LPC. À partir du signal microphone, on calcule le pseudorésiduel par un filtrage inverse en utilisant les coefficients estimés sur la voix lointaine. Une grande réduction de l'écho est alors réalisée. Ensuite, l'estimation de la fréquence glottale est utilisée pour synthétiser un filtre peigne dans le domaine fréquentiel. En parallèle, nous appliquons une transformée de Fourier au signal pseudorésiduel. Pour chaque trame nous effectuons un produit avec la fonction de transfert du filtre synthétisé peigne toujours dans le domaine fréquentiel. Finalement, un module de D.A.V. à base d'énergie est utilisé pour décider sur l'activité vocale du locuteur local (voir figure 3).

Nous avons étudié cinq structures:

structure 1: L'élimination de la contribution vocale du locuteur lointain est réalisée avant l'annulation du fondamental.

structure 2: L'annulation du fondamental est effectuée avant l'élimination de la contribution vocale.

structure 3: Structure 2 suivie d'une deuxième annulation du fondamental.

Les structures 4 et 5 sont identiques aux 2 et 3 sauf que l'estimation du fondamental est évaluée à partir de la sortie du filtre adaptatif.

De plus, nous avons utilisé trois façons de supprimer la contribution glottique du locuteur lointain. Les composantes spectrales du fondamental ont été supprimées selon les trois considérations suivantes:

H1 :Suppression des composantes spectrales les plus 'proches' des harmoniques du fondamental.

H2: Atténuation des composantes spectrales en limitant l'énergie des harmoniques du fondamental.

H3: Atténuation des composantes spectrales en pondérant les harmoniques du fondamental avec une fonction qui est proportionnelle à leur 'éloignement' .

## Critère d'Évaluations

Nous avons utilisé le critère RLE d'Alcatel. Il définit en dB le rapport entre l'énergie du locuteur local ( $EnLoc$ ) et l'énergie de l'écho résiduel ( $EnEcho$ ) pour chaque trame. Le RLE est évalué à partir du signal microphone  $Tx(n)$ .

On doit discriminer entre 2 situations possibles: le cas de la double parole ( $DbTlk(n)=1$ ) ou de l'écho résiduel seul ( $DbTlk(n)=0$ ). Le  $VadRx(n)$  indique l'activité vocale du locuteur lointain ( $VadRx(n)=1$  s'il parle ou  $VadRx(n)=0$  dans le cas contraire).

La mesure du RLE est déterminée par :

$$RLE = 10 * \log_{10} \frac{EnLoc}{EnEcho}$$

Avec:

$$EnLoc = \frac{\sum_n^N (Tx(n))^2 * VadRx(n) * (1 - DbTlk(n))}{\sum_n^N VadRx(n) * (1 - DbTlk(n))}$$

$$EnEcho = \frac{\sum_n^N ((Tx(n))^2 * VadRx(n) * DbTlk(n))}{\sum_n^N (VadRx(n) * DbTlk(n))}$$

## Résultats

	H1	H2	H3
0 km/h	1.31	1.90	2.05
60 km/h	0.46	0.96	0.82
90 km/h	0.23	0.44	0.95
130 km/h	1.091	0.568	0.24

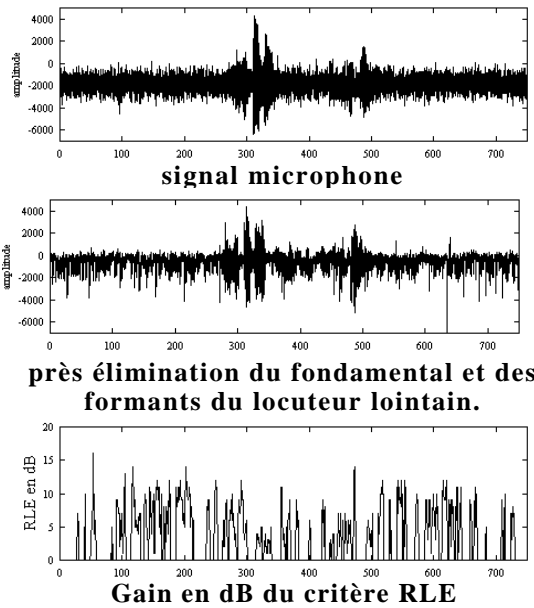
Table 1: résultats de la structure 1

	H1	H2	H3
0 km/h	1.21	1.56	2.02
60 km/h	0.22	0.53	0.77
90 km/h	0.75	0.31	0.95
130 km/h	0.51	0.22	0.24

Table 2: résultats de la structure 2

Les tableaux résultats 1, 2 et 3 donnent la mesure du gain RLE pour différentes structures. On a également pris en compte différentes conditions de conduites pour le véhicule. À 0 km/h le moteur du véhicule est en marche. La vitesse de 60 km/h caractérise surtout les conditions réelles en ville. Alors que 130 km/h caractérise les conditions extérieures de la ville. Les résultats présentés ici, se réfèrent au cas où les fenêtres du véhicule sont entièrement fermées.

Les notations H1, H2 et H3 caractérisent les stratégies de suppression des harmoniques telles que décrites précédemment.



**Figure 4. Traitement du signal microphone** (voie=homme, vitesse= 130 km/h, fenêtres fermées) avec la structure 1 et la stratégie H1.(temps cs).

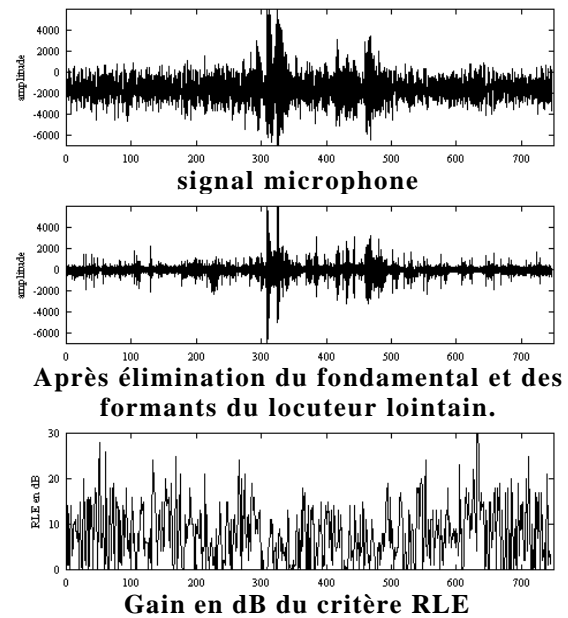
La structure 1 est meilleure que la structure 2 pour les conditions de 0 km/h et 130 km/h (voir fig.4). Ceci correspond à la situation pour laquelle la fonction de transfert caractérisant le couplage et le bruit à l'intérieur du véhicule varie légèrement. Au contraire, la structure 2 n'est pas aussi bonne que la structure 1 à 90 Km/h et 60 Km/h.

	H1	H2	H3
0 km/h	2.90	3.53	3.5
60 km/h	3.33	4.13	4.5
90 km/h	4.00	5.31	5.0
130 km/h	4.50	4.7	4.0

**Table 3 : résultats de la structure 3**

La structure 3 semble donner les meilleurs résultats avec le critère RLE pour les 3 structures proposées dans les différentes conditions de conduite (voir fig.5). La méthode H3 semble être la meilleure technique de suppression du fondamental et de ses harmoniques. En fait, la structure 3 et la structure 2 sont supposées être comparables puisque la seule différence est qu'on a ajouté une annulation du fondamental dans la structure 3. La différence importante en terme du gain RLE est peut être due au processus de suppression des formants qui probablement, fait ressortir les harmoniques du locuteur lointain. D'après les graphiques, il

semble que la structure 3 introduit moins de distorsions au signal traité que les autres structures.



**Figure 5. Traitement du signal microphone** (voie=homme, vitesse= 130 km/h, fenêtres fermées) avec la structure 3 et la stratégie H3. (temps cs)

### Conclusion

Différentes approches ont été présentées, introduisant la notion de fréquence glottale pour résoudre le problème de détection de double parole dans le contexte radio mobile.

Une détection sur chacun des canaux de transmission a été étudiée. Malheureusement, l'insensibilité du fondamental aux distorsions introduites par l'habitacle n'est pas toujours vérifiée. De plus, la similarité du fondamental entre les locuteurs peut soulever d'autres problèmes.

Une détection mono-canal a été étudiée. Elle consiste à rehausser le signal du locuteur local (voie microphone) vis-à-vis de l'écho. L'introduction du détecteur de fréquence glottale dans ce cas, s'est avérée plus intéressante avec un gain en dB considérable pour toutes les structures proposées. L'introduction d'un réducteur de bruit, n'a pas été étudiée. Mais, il semble être un facteur important pour améliorer davantage les performances atteintes. En effet, l'annulation du fondamental et l'élimination de la contribution vocale du locuteur lointain sur le microphone donne naissance à un écho de nature aléatoire (bruit). Cet écho pourrait être complètement supprimé en lui appliquant un algorithme de réduction de bruit.

Nous pensons qu'il est donc possible d'intégrer cette technique à un algorithme de réduction de bruit afin d'obtenir des performances supérieures à celles décrites ici. Nous rappelons que l'évaluation a été faite après suppression de la contribution du locuteur lointain sans réduction a priori ou à

posteriori de bruit. L'intégration de systèmes mixtes (intégration de connaissances perceptives avec les algorithmes standards de traitement des signaux) devrait donc permettre d'améliorer les systèmes actuels de reconnaissance de parole.

Microélectronique, par la fondation de l'UQAC et par Alcatel Mobile Phone.

### Références

- S.F. Boll (1979), Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. on ASSP*, Vol. 27, No. 2, pp. 113-120, April.
- N. D. Degan and C. Prati (1988), Acoustic noise analysis and speech enhancement techniques for mobile radio applications, *Signal Processing* 15, pp 43-56.
- Y. Ephraim, D. Malah (1984), Speech enhancement using a minimum mean square error short time amplitude estimator, *IEEE Trans. on ASSP*, Vol. 32, No. 6, pp. 1109-1121, December.
- S. Haykin (1991), Adaptive filter theory, *Second Edition*, Prentice-Hall, Englewood Cliffs, New Jersey.
- J. F. Kaiser (1990). On a simple algorithm to calculate the 'energy' of a signal. *Actes de IEEE-ICASSP'90*, Albuquerque, pp.381-384.
- J. F. Kaiser (1993). Some Useful properties of Teager's energy operators. *Actes de IEEE-ICASSP'93* vol. 3, pp.149 -152.
- J.S. Lim, A.V. Oppenheim (1979), Enhancement and bandwidth compression of noisy speech, *Proc. of the IEEE*, Vol. 37, No. 12, pp. 1586-1604, December.
- O. Macchi, M. Bellanger (1988), *Le filtrage adaptatif transverse*, *Traitement du Signal*, Vol. 5, No. 3, pp. 115-132.
- B.C.J. Moore (1989). An Introduction to the Psychology of Hearing . Academic Press, London.
- K. Ozeki, T. Umeda (1984), An adaptative algorithm using an orthogonal projection to an affine subspace and its properties, *Electronics and Communications in Japan*, Vol. 67-A, No. 5, pp. 19-27.
- J. Prado, E. Moulines (1993), Frequency-domain adaptive filtering with application to echo cancellation, *Proc. of the Third International Workshop on Acoustic Echo Control*, pp. 249-258.
- J. Rouat (1993). Nonlinear operators for speech analysis. dans *Visual representations of speech signals*, ed. by Martin Cooke, Steve Beet and Malcom Crawford. John. Wiley&Sons, London. pp.335-340.
- J. Rouat, Y.C. Liu and D. Morissette (1997). A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication*, Vol. 20, mars 1997, 17 pages.
- B. Widrow et al. (1975), Adaptive noise cancelling: principles and applications, *Proc. of the IEEE*, Vol. 63, No. 12, pp. 1692-1716.

### Remerciements

Ces travaux ont été financés par le Conseil National de la Recherche Canadienne en Sciences Naturelles et en Génie, par le fonds FCAR, par la Société Canadienne de