DÉBRUITAGE DE PAROLE PAR UN FILTRAGE UTILISANT L'IMAGE DU LOCUTEUR

L. Girin, G. Feng et J-L. Schwartz

Institut de la Communication Parlée, URA CNRS 368
INPG/ENSERG/Université Stendhal
B.P. 25, 38000 GRENOBLE CEDEX, France
girin@icp.grenet.fr

RÉSUMÉ

La parole étant à la fois acoustique et visuelle, il est intéressant d'utiliser cette bimodalité pour améliorer les performances des outils de télécommunications et de communication homme-machine. Nous proposons dans cet article une technique originale de réduction de bruit utilisant des filtres estimés à partir de la forme des lèvres du locuteur. Après avoir sélectionné deux techniques de filtrage, nous présentons une méthode simple et efficace pour relier la forme des lèvres et ces filtres. Le système complet de débruitage est testé dans le cadre de voyelles stationnaires, ce qui permet notamment une première approche du problème des gestes non-visibles. Les résultats de tests perceptifs sont présentés. Ils ont permis de valider le système.

ABSTRACT

Since speech is both auditory and visual, visual cues could compensate to a certain extent the deficiency of auditory ones, in order to improve man-machine communication and telecommunication tools. This paper deals with a noise reduction technique based on speech enhancement with adaptive filters estimated from the speaker's lip pattern. We first present the selected filtering techniques, and then the tool we used to predict the filter pattern from the lip shape. The whole noise reduction system is tested in the context of stationary vowels including a first kick into the problem of non-visible gestures. The results of perceptual tests are presented in order to quantify the performances of the system. These results are quite promising.

1. INTRODUCTION

Un des principaux problèmes à résoudre pour les systèmes de télécommunication ou de communication homme-machine du futur est celui du débruitage des signaux de parole à transmettre ou à traiter. Or dans ce domaine il existe une compétence humaine importante et pourtant encore largement inexploitée dans les systèmes de traitement de la parole : la capacité d'extraire l'information de l'interlocuteur grâce à l'information captée par le système visuel sur le « visage parlant ». Ainsi, la vision du visage du locuteur et des lèvres en particulier est un renfort précieux lorsque l'audition est insuffisante : on peut citer la lecture labiale des sourds, le rehaussement d'un discours bruité, en langue étrangère ou peu intelligible [2, 4, 9].

De nombreux modèles d'intégration audiovisuelle ont été élaborés dans le domaine de la perception de la parole [9] et certains systèmes de reconnaissance de la parole exploitent déjà cette bimodalité [5]. Mais nous voulons tester dans cette étude une idée nouvelle : le débruitage de signaux de parole par un filtrage utilisant l'image du locuteur. Il s'agit d'estimer, à partir de l'image du locuteur et de sa forme des lèvres en particulier, un modèle du signal audio prononcé, puis à filtrer le signal audio bruité grâce au modèle estimé [Fig. 1].

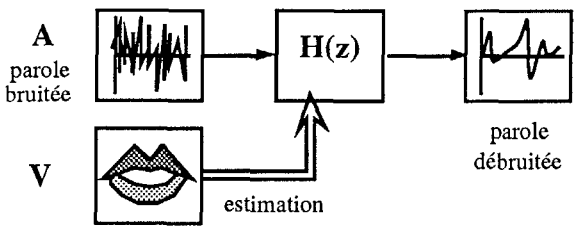

Figure 1 – Principe du système

Une telle architecture implique deux modules principaux :

1) Le filtre proprement dit qui doit permettre de débruiter le signal de parole. Le choix technique des filtres est discuté dans la section 2 de cet article.

2) Un associateur qui doit permettre d'estimer ce filtre à partir de la forme des lèvres. L'information lue sur les lèvres reste partielle : nous n'avons pas accès aux « composantes cachées » de la parole provenant des articulateurs internes du conduit vocal (langue, palais...) et de la source (excitation glottale). Toutefois, une information conséquente peut être lue ou inférée à partir des lèvres [9]. En nous limitant dans cette première étude au cadre de voyelles stationnaires, nous proposons dans la section 3 une solution simple et efficace pour l'estimation des paramètres des filtres et pour la résolution d'un problème typique de gestes articulatoires non-visibles.

Enfin la section 4 fournit des résultats qualitatifs sur chaque élément du système (filtres et associateur) et le système complet est testé quantitativement, ce qui permet d'obtenir des premiers résultats très prometteurs.

2. DEUX TYPES DE FILTRES

Notre étude s'inscrit dans le cadre général de l'estimation adaptative d'un signal s(k) bruité par un bruit additif v(k), s et v étant supposés décorrélés. Nous nous intéresserons ici au cas classique de l'estimateur linéaire optimal en moyenne quadratique [1]. Celui-ci est fourni en filtrant l'observation x=s+v par le filtre de Wiener, dont l'expression dans le domaine fréquentiel est

$$W(\omega) = \frac{\gamma_{sx}(\omega)}{\gamma_{xx}(\omega)} = \frac{\gamma_{ss}(\omega)}{\gamma_{ss}(\omega) + \gamma_{vv}(\omega)}$$

$\gamma_{ss}(\omega)$ et $\gamma_{vv}(\omega)$ sont les densités spectrales de puissance du signal et du bruit, non accessibles à partir du signal acoustique. Comme nous attendons du « signal de référence image » de fournir des informations spectrales sur le signal, nous proposons donc un filtre de Wiener « labial » où la DSP du signal est estimée à partir des lèvres selon la procédure que nous décrivons dans la section 3. Il reste à estimer la DSP du bruit. Dans la suite de l'étude nous considérons uniquement le cas simple d'un bruit blanc et cette estimation revient à l'estimation du RSB ou de l'énergie du bruit (par exemple par la mesure du bruit pendant les silences de parole).

Le critère de minimisation de l'erreur quadratique entre le signal et son estimé propre au filtre de Wiener est un critère très contraignant : dans le domaine fréquentiel, il s'applique simultanément sur *le module et la phase* de ces spectres. Or on connaît la pertinence du module du spectre du signal du point de vue du système auditif ([3] chapitre V). C'est pourquoi on cherche une alternative au filtre de Wiener en proposant un filtre capable de débruiter la parole par renforcement du spectre (en module) du signal bruité. La modélisation par prédiction linéaire (Linear Predictive Coding [7]) s'impose alors en présentant un double avantage : elle décrit correctement l'enveloppe du spectre du signal et donc les formants de la parole tout en fournissant directement un filtre réhausseur sous la forme d'une fonction de transfert numérique tout-pôles H(z) modélisant le conduit vocal et la forme de l'onde glottique.

$$H(z) = S_{LPC}(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{G}{A^+(z)}$$

En général, diverses méthodes sont disponibles pour identifier H(z) à partir d'un signal audio [7]. Mais dans notre application, les paramètres descripteurs du spectre LPC doivent être estimés à partir de l'image du locuteur (section 3).

Nous unifions alors les informations labiales nécessaires aux deux filtres en prenant comme estimateur $\widehat{\gamma}_{ss}(z)$ de la DSP du signal, le module carré du spectre LPC estimé $\widehat{S}_{LPC}(z)$. La fonction de transfert du filtre de Wiener est alors donnée par

$$W(z) = \frac{|S_{LPC}(z)|^2}{|S_{LPC}(z)|^2 + E_b} = \frac{G^2/E_b}{G^2/E_b + A^+(z).A^-(z)}$$

(la notation + et - désigne respectivement la forme causale et anticausale des polynômes ; on a omis la notation estimateur ^). Le calcul de la forme causale du filtre selon la procédure décrite par Papoulis [8] aboutit à la formule

$$W(z) = \frac{G^2}{E_b} \cdot \frac{B_1^+(z)}{A_1^+(z)}$$

où

$$A_1^+(z).A_1^-(z) = G^2/E_b + A^+(z).A^-(z)$$

et $B_1^+(z)$ est un polynôme de degré au plus p-1.

La force du filtre de Wiener réside dans son adaptation au RSB qui repose sur deux phénomènes : le déplacement des pôles du filtre par rapport aux pôles LPC à l'intérieur du cercle unité sous l'influence de la constante G^2/E_b et la compensation des résonances correspondants à ces pôles par les racines de $B_1^+(z)$. Dans le cas d'un bruit faible, pôles et zéros du filtre se compensent tout en se dirigeant vers l'origine pour fournir un filtre aplani tendant vers le filtre plat unitaire. Quand le bruit augmente et devient prédominant, les pôles du filtre tendent vers les pôles LPC, ce qui explique que l'on retrouve des caractéristiques spectrales du signal.

L'adaptation au RSB est exclusive à Wiener, mais elle s'accompagne d'une complexité accrue et d'hypothèses fortes sur le bruit que ne connaît pas le filtre LPC. En revanche ce dernier n'est pas adapté au cas d'un bruit faible (et a fortiori nul), le filtrage étant alors plus synonyme de dégradation que de renforcement. Finalement l'oreille est seule apte à « départager » les deux filtres et à juger de la validité de leurs critères d'optimalité dans notre application (section 4). Rappelons que les deux filtres gardent l'avantage d'être synthétisés à partir de la même information : l'estimation du spectre LPC du signal à partir de la forme des lèvres que nous allons décrire maintenant.

3. L'ASSOCIATEUR LÈVRES-SPECTRE

Le problème consiste donc à élaborer un associateur entre deux jeux de paramètres : d'un côté des descripteurs du contour labial, de l'autre des paramètres représentatifs du spectre LPC correspondant.

En ce qui concerne la forme des lèvres, le poste « visage-parole » de l'ICP [4] permet l'extraction automatique de trois paramètres pertinents du contour interne du vermillon : la largeur A, la hauteur B et l'aire S. Les enregistrements audiovisuels sont réalisés avec des contraintes rigoureuses sur le locuteur : tête maintenue par un casque, éclairage idéal et maquillage bleu des lèvres pour l'extraction du contour grâce au procédé « Chroma-Key ».

Pour la représentation du spectre LPC, nous avons choisi les coefficients PARCOR k_i. Ce choix est motivé par leur robustesse (au sens d'une grande stabilité du spectre de H(z) par rapport à des variations des k_i) comparée à celle d'autres représentations testées, notamment les pôles de H(z).

Devant la difficulté posée par le traitement de la parole continue (multiplicité des configurations labiales, gestion de la coarticulation et des gestes non-visibles...) nous avons limité, dans cette étude de faisabilité, l'implantation de notre système au débruitage de voyelles stationnaires en mode monolocuteur. Le corpus contient 700 enregistrements audio-vidéo pour 7 voyelles du français [a e i ø y o u] prononcées de manière tenue, soit 100 séquences par voyelle. Chaque élément est composé d'un segment audio de 200 ms et d'une mesure [A B S] synchrone au premier échantillon audio.

L'absence au niveau labial d'informations concernant l'énergie du signal interdit l'accès au gain G. Ceci ne pose pas de problème en terme de gain pour le filtrage de voyelles stationnaires : l'énergie en sortie est simplement ajustée pour les rendre écoutables. En ce qui concerne le calcul du filtre de Wiener, on montre aisément que si G_{LPC} est l'énergie de la réponse impulsionnelle du filtre tout-pôles 1/A(z) issu de notre associateur, le rapport G^2/E_b est égal au rapport RSB/G_{LPC}, ce qui nous ramène à l'hypothèse de contrôle du RSB.

À la lumière de ces précisions, décrivons à présent notre associateur. Nous avons opté pour une méthode classique de régression linéaire matricielle, les travaux de Robert-Ribes [9] ayant montré qu'elle suffisait à fournir de bonnes performances dans le problème à traiter. Son principe est le suivant : étant données deux matrices L et K de dimensions respectives m×n et m×p, on cherche la matrice F telle que le produit LF approche au mieux la matrice K au sens où elle minimise l'erreur quadratique $e = \| LF - K \|$. Dans le cas général (L non inversible et $F \neq L^{-1}K$), F est une matrice de dimension n×p réalisant la régression linéaire entre K et L au sens des moindres carrés. Le calcul de F s'appuie sur des algorithmes de décomposition matricielle classiques (Householder, Cholesky...) que nous ne détaillerons pas ici. Appliquons maintenant cette méthode pour la construction de notre associateur. Sur les 700 paires audio-vidéo des sept voyelles citées, 350 vont être utilisés lors de la phase d'apprentissage et 350 seront disponibles pour tester les performances de

l'associateur et du filtrage. La matrice K contient les coefficients k_i modélisant le spectre de chacun des signaux audio d'apprentissage. Ils sont obtenus par la méthode d'autocorrélation qui garantit la stabilité du spectre [7] (algorithme de Durbin-Levinson appliqué à l'ordre 16 pour une bonne modélisation des cinq premiers formants et de l'onde glottale). La matrice L est issue de la concaténation des 50 premiers triplets [A B S] de chaque voyelle. On lui rajoute une colonne de 1 de manière à donner à la régression une forme affine qui tient compte d'une position origine optimale des k_i, calculée de manière intrinsèque (c'est la dernière ligne de F de dimension 4x16). Pour tout vecteur $[A_0\ B_0\ S_0]$ provenant du signal vidéo on obtient une estimation $[k_{01}\ k_{02}\ \ldots\ k_{0p}]$ des k_i du spectre correspondant par

$$[A_0\ B_0\ S_0\ 1] \,.\, F = [k_{01}\ k_{02}\ \ldots\ k_{0p}]$$

Sur notre corpus de test, le pourcentage de k_i ne garantissant pas la stabilité du filtre (extérieurs au segment]-1,1[) est de moins de 1%, ce qui nous a permis de valider notre associateur. Le problème des filtres instables résiduels est résolu par une procédure de renormalisation des modules des k_i incriminés à 0,99, procédure qui fait de nouveau appel à l'appréciable robustesse de ces coefficients.

Un problème majeur apparaît dans le caractère surjectif de l'associateur dû à la non-distinction entre certaines voyelles à partir des paramètres [A B S]. Ainsi les couples [ø o] et [y u] impliquent chacun des voyelles de formes labiales très proches, ne différant quasiment que par la position antérieure ou postéreure de la langue [3]. Dans notre application ces ambiguïtés sont ingérables au niveau de la régression qui ne peut réaliser l'apprentissage de k_i différents avec des entrées semblables. Nous avons donc choisi de séparer les sept voyelles en deux groupes de cinq ne comportant respectivement que les voyelles antérieures ou périphériques (incluant les postérieures), soit [a e i ø y] et [a e i o u]. Pour construire un outil capable de choisir le bon spectre parmi les deux proposés par les deux associateurs, nous sommes restés dans le cadre, simple et manifestement fonctionnel pour le présent problème, des méthodes linéaires. Nous avons utilisé une analyse discriminante portant uniquement sur les couples ambigus [ø o] et [y u] puisque le choix parmi les spectres conjoints pour les [a], [e] et [i] importe peu (ils sont d'ailleurs très semblables en pratique). Cette analyse intervient dans un espace spectral qui vient suppléer l'espace visuel porteur de l'ambiguïté. Elle repose sur le contraste entre la forme « plutôt hautes fréquences » des spectres antérieurs de [ø] et [y] et la forme « plutôt basses fréquences » des spectres postérieurs de [o] et [u]. Chaque spectre est défini sur 20 valeurs en dB, l'échelle fréquentielle choisie étant l'échelle perceptive des bark adaptée aux caractéristiques de l'oreille humaine (z(bark) = 7 ArgSh (f(Hz)/650 [3]).

Les paramètres de l'analyse discriminante sont déterminés à nouveau sur la moitié du corpus (50 spectres pour chacune des 4 voyelles). L'objectif étant une classification de spectres bruités, nous avons élargi l'ensemble d'apprentissage en présentant chacun des 200 spectres à 5 niveaux de bruit allant du non bruité à un bruit assez fort mais maintenant une séparation suffisante des 2 classes, soit : non bruité, 24, 12, 6 et 0 dB. De plus, des tests préliminaires ont montré que les paramètres labiaux (A, B, S) éliminés au départ pour la sélection avant-arrière pouvaient fournir un gain sensible, le classifieur étant capable d'utiliser l'information distinctive entre les couples [ø o] et [y u] pour mieux « focaliser » sa décision. Nous avons donc 1000 vecteurs d'apprentissage de dimension 23 (20 paramètres spectraux, 3 paramètres visuels A, B, S). Le seuil de décision λ est naturellement le barycentre des moyennes des nuages de chaque classe du corpus d'apprentissage projetés sur l'axe discriminant, pondéré par les écarts-types de ces nuages par rapport à ces moyennes. Les performances du sélecteur ont été testées avec la moitié du corpus non utilisée pour l'apprentissage, et ce sur 9 niveaux de bruit. Les pourcentages de choix correct sont exposés dans le tableau 1. Ils nous ont permis de valider ce sélecteur et d'implanter notre système de débruitage pour lequel la partie suivante fournit quelques résultats formels et informels.

RSB	SB	24	12	6	0	-6	-12	-18	-24
%	99	99,5	100	97,5	93	77,5	68	64	65

Tableau 1 – Performances du sélecteur antérieure/périphérique

4. ÉVALUATIONS DU SYSTÈME

Nous avons observé le bon comportement global de l'associateur lèvres-spectre. Les spectres ont une allure correcte et il ne commet aucune erreur grossière du type création de « monstres spectraux ». D'une manière générale, le principal problème est que la répartition des paramètres [A B S] ne suit pas une relation parfaitement linéaire avec l'espace acoustique [9]. Il existe trois regroupements autour des distinctions ouvertes [a], étirées [e i] et arrondies [ø y] ou [o u] qui ne correspondent pas vraiment à la couverture de l'espace acoustique. Dans ces conditions il est difficile pour la régression de fournir une solution linéaire qui satisfasse chaque groupe, et les regroupements provoquent un phénomène de « moyennage » de spectres issus de phonèmes distincts mais labialement proches. Les deux associateurs s'en tirent plutôt bien pour les distinctions [o/u] ou [ø/y] avec cependant quelques problèmes pour le délicat couple [e i]. On aboutit dans cet exemple à la synthèse d'un filtre souvent hybride dont l'allure spectrale est un mélange des formants des deux phonèmes [Fig. 2a].

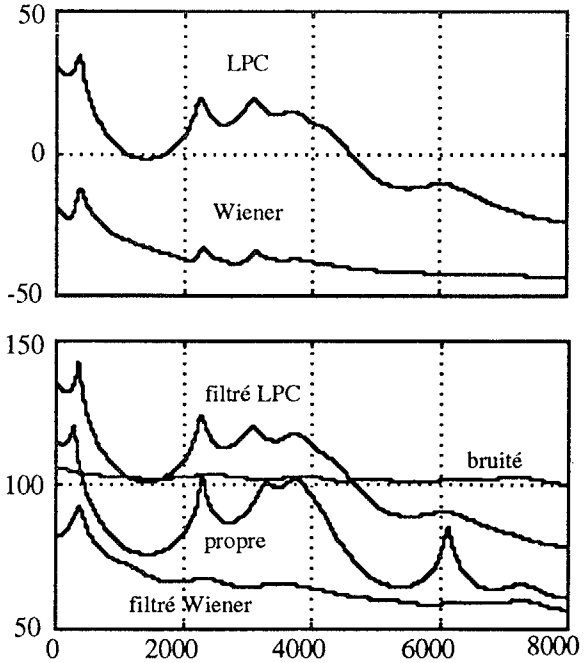

Figure 2 – Exemple de la voyelle [i] – RSB=-18dB
a) Filtre LPC et filtre de Wiener
b) Signal original propre et bruité, signal bruité filtré LPC et Wiener

En ce qui concerne le filtrage, l'amélioration est particulièrement spectaculaire pour le filtre LPC dans le cas fortement bruité. Le filtrage transforme des signaux inaudibles (bruit noyant complètement le signal dès -12 dB) en voyelle,

au mieux identifiable sans ambiguïté, au pire classifiable de manière erronée. Dans le cas fortement bruité, la sonorité fricative de la voyelle, qui n'est autre que le résultat de l'excitation d'un filtre formantique par un bruit, reste perceptuellement préjudiciable. Dans le cas faiblement bruité, la déformation induite sur les signaux dénote une faiblesse du filtre LPC. À l'inverse le filtre de Wiener se caractérise par sa pertinence pour les RSB fort (filtrage quasi plat) et son aspect « moins violent » que le filtrage LPC pour les RSB faibles.

Pour quantifier les résultats du débruitage, nous avons conduit un test perceptif d'identification à choix forcé sur les échantillons de nos sept voyelles non utilisés pendant la phase d'apprentissage. Le test a consisté à soumettre à 20 sujets en chambre sourde des séries équilibrées de 35 stimuli (7 voyelles, 5 échantillons par voyelle) ordonnés aléatoirement, et ceci pour les signaux originaux bruités et les mêmes signaux filtrés par nos deux filtres d'une part, et pour 7 RSB différents d'autre part (sans bruit, 12, 6, 0, -6, -12, et -18 dB). Chaque point de la courbe des résultats [Fig. 3] représente ainsi 700 stimuli différents (5 par voyelle et par série, 7 voyelles, 20 auditeurs).

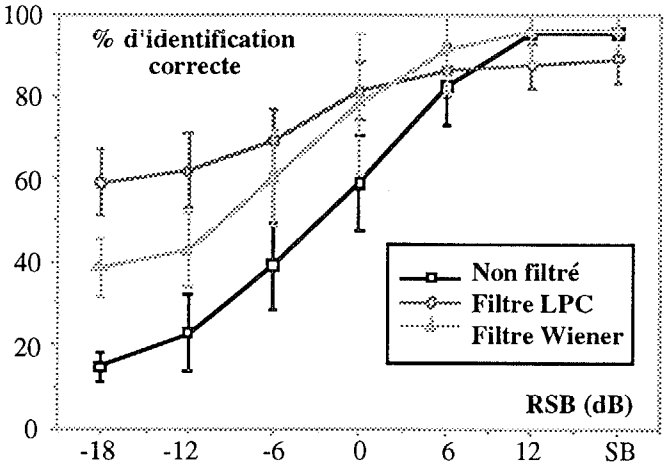

Figure 3 – Résultats du test perceptif

Les résultats sont excellents. Le filtrage LPC permet 59% d'identification correcte pour un RSB de -18 dB et 62% pour -12 dB représentant des gains d'intelligibilité respectifs de 44% et 39% par rapport aux scores sur signaux bruités (proches de l'aléatoire à -18 dB). Pour ces mêmes RSB faibles les scores du filtrage de Wiener sont sensiblement en retrait (38,5% et 43% pour des gains de 23,5% et 20%), alors que les scores relatifs aux faibles niveaux de bruits sont meilleurs : ils dépassent la LPC à partir de 6 dB et s'égalisent ensuite quasiment avec les scores sur signaux non filtrés, alors que la LPC induit une dégradation d'environ 6% par rapport aux signaux propres. On retrouve bien la répartition des performances des deux techniques le long de l'axe faible-fort bruit.

Les résultats pour chaque RSB ont été rassemblés dans des matrices de confusions (non présentées ici) qui montrent que l'identification des signaux filtrés reste efficace au-delà de la reconnaissance exacte. La concentration des scores autour de la diagonale en trois blocs [a], [e i] et [ø y o u] reflète bien la répartition des voyelles dans l'espace des paramètres [A B S]. On note la double confusion au sein du groupe [ø y o u], à la fois à l'intérieur des couples [ø y] et [o u] qui représentent chacun une famille spectrale (aiguë ou grave), et à l'intérieur des couples [ø o] et [y u] que le sélecteur antérieur/périphérique cherche à départager. Les performances de ce dernier restent limitées en fort bruit ce qui soulève la question de l'allure perceptive des voyelles d'une catégorie traitées par un filtre provenant d'une voyelle de l'autre catégorie.

Enfin, il est intéressant de comparer nos résultats à ceux de Robert-Ribes qui a mené un test perceptif audiovisuel sur le même corpus de voyelles pour obtenir une mesure du gain entre les scores en condition audiovisuelle par rapport à la condition auditive seule [9]. Les résultats sont très proches, ce qui montre que notre filtrage peut atteindre les performances de l'intégration audiovisuelle sur ce corpus très limité, et ceci en connaissant des limites tout-à-fait similaires aux limites humaines (confusion pour des formes de lèvres proches, confusion antérieure/périphérique pour une information auditive déficiente).

5. Conclusion – perspectives

Nous avons présenté dans cet article une méthode inédite de débruitage de parole par un filtrage utilisant l'image du locuteur. L'utilisation de deux types de filtres, LPC et Wiener voit chaque méthode se spécialiser dans un extrême de l'axe faible-fort bruit tout en gardant des performances globales remarquables, ce qui laisse la voie ouverte vers l'élaboration d'un filtre hybride pondéré en fonction du RSB.

L'implantation du système dans le cadre limité de voyelles stationnaires à l'aide d'un associateur linéaire simple a montré qu'une bonne estimation de spectres peut être effectuée par le système à partir de formes de lèvres assez distinctes. De plus le corpus a permis une première approche du problème complexe des gestes non-visibles en nécessitant l'élaboration d'un classifieur linéaire optimal entre voyelle d'avant et d'arrière. Les performances globales du débruitage à ce niveau de l'implantation semblent dépendre plus des qualités de nos outils utilisés pour extraire l'information des lèvres que des techniques de filtrage proprement dites. On peut donc envisager l'élaboration d'autres types d'associateurs et de sélecteurs plus complexes (par exemple des réseaux de neurones).

La réalisation dynamique du filtre des lèvres pour la parole continue qui reste l'objectif essentiel de nos travaux à venir peut donc s'appuyer sur des résultats préliminaires encourageants et prometteurs.

Références bibliographiques

[1] D. Baudois, C. Servière, A. Silvent, Soustraction de bruit, Traitement du Signal, vol. 6 , n°5, 1989.

[2] C. Benoît, T. Mohamadi, S. Kandel, Effects of Phonetic Context on Audio-Visual Intelligibility of French, *Journal of Speech and Hearing Research* , vol. 37, 1994, pp. 1195-1203.

[3] Calliope, La parole et son traitement automatique, J.P. Tubach (Ed.), Masson, Paris, 1989.

[4] N.P. Erber, Interaction of audition and vision in the recognition of oral speech stimuli, *Journal of Speech and Hearing Research*, vol. 12, 1969, pp. 423-425.

[5] A.J. Goldschen, Continuous automatic speech recognition by lipreading, Doctoral dissertation, George Washington University, 1993.

[6] T. Lallouache, Un poste « Visage-Parole » couleur : acquisition et traitement automatique des contours des lèvres », Thèse doctorale, INPG, Grenoble, 1990.

[7] J.D. Markel & A.H.Jr. Gray, Linear Prediction of Speech, Springer-Verlag, New-York, 1976.

[8] A. Papoulis, Signal Analysis, McGraw-Hill, New-York, 1977.

[9] J. Robert-Ribes, Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles, Thèse doctorale, INPG, Signal-Image-Parole, Grenoble, 1995.