Débruitage de parole par un filtrage utilisant l'image du locuteur. Une étude de faisabilité

Speech Enhancement with Filters Estimated from the Speaker's Image. A Feasability Study

par Laurent GIRIN, Gang FENG et Jean-Luc SCHWARTZ

Institut de la Communication Parlée, UPRESA 5009
Institut National Polytechnique de Grenoble Université Stendhal
Domaine Universitaire 1180 Av. Centrale - B.P. 25
38040 GRENOBLE Cedex 9
E-mail : girin@icp.grenet.fr

résumé et mots clés

La parole étant à la fois acoustique et visuelle, il est intéressant d'utiliser cette bimodalité pour améliorer les performances des systèmes de télécommunication et de communication homme-machine. Nous proposons dans cet article une méthode originale de réduction de bruit utilisant des filtres estimés à partir de la forme des lèvres du locuteur. Après avoir décrit 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, qui permet une première approche du problème des gestes non-visibles. Les résultats de tests perceptifs sont très encourageants et permettent de valider les principes de base du système, sous réserve d'extensions futures à des stimuli plus complexes.

Débruitage de parole, Parole audiovisuelle, Filtrage, Traitement d'images, Estimation de spectres.

abstract and key words

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 systems. This paper deals with a noise reduction method based on speech enhancement with adaptive filters estimated from the speaker's lip pattern. We first present the two 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 quite promising and allow us to validate the basic principles of our system, which should be tested with more complex stimuli in the future.

Speech enhancement, Audiovisual speech, Filtering, Image processing, Spectral estimation.

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 qu'ont les êtres humains à extraire l'information de leur interlocuteur grâce aux signaux captés par le système visuel : l'image du « visage parlant ». Plusieurs constatations révèlent l'importance de la vision dans la perception de la parole :

– La lecture labiale des sourds comble partiellement les lacunes ou l'absence totale d'audition.

– Plus généralement, la vision des lèvres et du visage du locuteur est un renfort précieux lorsque l'audition est insuffisante : si plusieurs personnes parlent en même temps, regarder un locuteur permet de faire mieux ressortir son message. De même la vision

est appréciée lorsque le discours est bruité ou peu intelligible [2, 5, 12, 26].

– L'effet McGurk [11] révèle la combinaison des informations des deux modalités. Sur proposition de stimuli auditifs et visuels conflictuels, on voit naître un percept mixte : ainsi l'audition du son [aba] en présence du stimulus visuel [aga] aboutit à la perception de [ada].

Un certain nombre d'études permettent de quantifier la contribution de la vision dans la perception de la parole bruitée [2, 5, 12, 26]. Erber [5] rapporte que pour des stimuli présentés avec un rapport signal à bruit (RSB) de -12dB, le pourcentage de reconnaissance correcte passe de 20% avec l'audition seule à 80% en présentation audiovisuelle. Nous reportons en figure 1 la « courbe Erber » (reconnaissance audio seule vs reconnaissance audiovisuelle en fonction du RSB) obtenue par Benoît, Mohamadi et Kandel [2] pour des enchaînements phonétiques du type VCVCV (trois voyelles [i a y], six consonnes [b v z ʒ R l]).

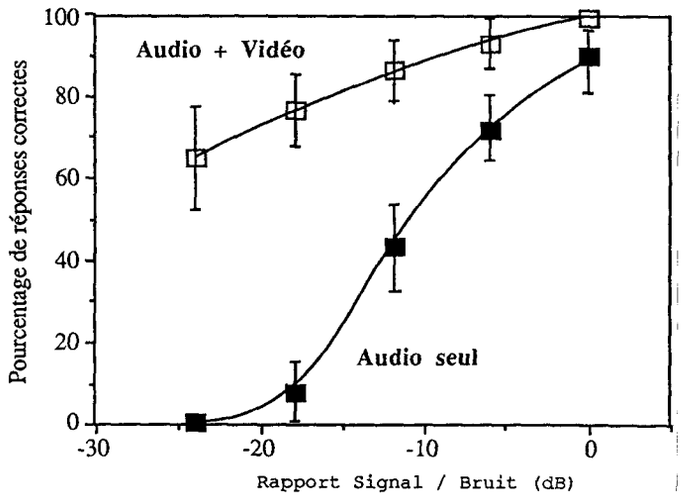


Figure 1. – Reconnaissance auditive et audiovisuelle d'enchaînements VCVCV dans le bruit (d'après Benoît, Mohamadi et Kandel [2]). Le taux de 0% à - 24 dB s'explique par la possibilité d'une non-réponse de type «ne sait pas».

Cet aspect multimodal de la parole est déjà exploité dans un certain nombre de systèmes de traitement de la parole. Les applications concernent notamment la segmentation du signal de parole [13], la reconnaissance visuelle (*lipreading*) [8, 16] et audiovisuelle [4, 6, 21, 25] (voir aussi une série de propositions récentes dans [18]). Dans ce dernier cas les informations visuelles s'ajoutent aux entrées acoustiques. Dans la plupart des applications audiovisuelles, l'apport de la modalité visuelle est particulièrement significatif en milieu bruité.

De même, de nombreux modèles d'intégration audiovisuelle ont été élaborés dans le domaine de la perception de la parole [17, 18, 23, 27, 31]. L'un de ces modèles est dit « à intégration précoce » du fait de la projection et de l'intégration avant catégorisation des informations provenant des deux modalités visuelle et auditive dans un espace de représentation commun qui peut être articulatoire ou spectral. Dans une optique incluant la fonction de reconnaissance de signaux bruités, ce modèle a été testé par Yuhas *et al.* [31], puis par Robert-Ribes [23].

Nous proposons dans cet article une idée nouvelle : le débruitage de signaux de parole par un filtrage utilisant l'image du locuteur. Le principe consiste à estimer, à partir de l'image du locuteur et de la forme de ses lèvres en particulier, un modèle du signal audio prononcé, puis à filtrer le signal audio bruité par un filtre numérique $H(z)$ construit à partir du modèle estimé [fig.2].

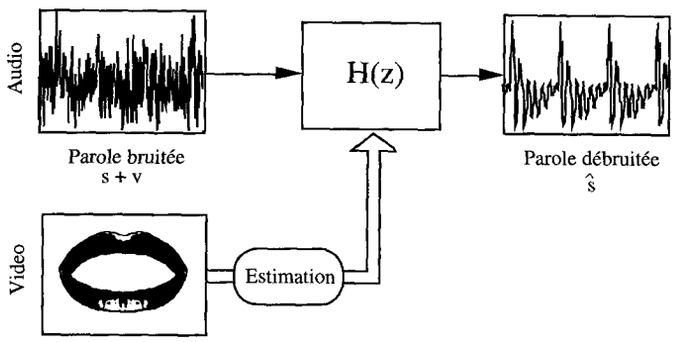


Figure 2. – Schéma de principe du débruitage à l'aide de filtres estimés à partir des lèvres.

En relation avec le modèle d'intégration cité plus haut qui cherche un plan commun pour l'intégration bimodale, on voit que l'idée est ici d'utiliser les informations issues de la modalité visuelle et de les formaliser en filtres pour extraire celles de la modalité auditive. Pour réaliser cet objectif, le système doit contenir deux éléments principaux :

1) Le filtre proprement dit qui doit permettre de débruiter le signal de parole. La question est de trouver un filtre optimal au sens d'un critère adapté à la parole. Le choix n'est pas évident : il pose le problème de la non-unicité du critère d'optimalité et de son adéquation pour l'application spécifique de débruitage de parole. L'optimalité doit être entendue ici en terme de perception de la parole traitée. Nous décrivons dans la section 2 l'élaboration des filtres en liaison avec une revue succincte des techniques classiques de débruitage.

2) L'outil qui doit réaliser l'estimation des paramètres des filtres à partir de la forme des lèvres. L'objectif est d'apprendre l'association entre variables d'espaces différents à partir d'exemples. Il faut souligner que si l'information lue sur les lèvres est effectivement exploitable, elle reste limitée par son caractère partiel : 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) qui conduisent à une multiplicité des sons possibles pour chaque configuration labiale. Cette multiplicité de sons est de plus accrue dans le cas de la parole continue par la coarticulation. Pour valider cette idée de débruitage visuel nous devons commencer par des sons relativement simples. Nous avons choisi de centrer cette première étude

sur le débruitage de voyelles stationnaires en mode monolocuteur. Par là même, nous considérons clairement ce travail comme une étude de faisabilité destinée à confirmer sur des signaux simples l'apport potentiel d'un filtrage labial. Dans ce contexte, nous proposons dans la section 3 une solution simple et efficace pour l'estimation des paramètres des filtres à partir de paramètres descriptifs du contour labial. Outre l'implantation du système complet de débruitage grâce à l'utilisation d'un associateur linéaire simple, ce cadre de voyelles stationnaires a permis une première démarche pour résoudre un problème typique de gestes articulatoires non-visibles.

La section 4 fournit des résultats sur chaque élément du système (filtres et associateur) et le système complet est évalué quantitativement. Les résultats se montrent très prometteurs et permettent de discuter en section 5 des apports et des perspectives offertes par cette étude préliminaire par rapport à notre objectif global.

2. élaboration des filtres

2.1. élaboration d'un filtre de type Wiener

Le cadre général théorique du débruitage dans lequel nous nous plaçons est celui de l'estimation adaptative d'un signal $s(k)$ bruité par un bruit additif $v(k)$ [1]. Nous nous intéresserons ici au cas classique de l'estimateur linéaire optimal au sens de la minimisation de l'erreur quadratique moyenne $E[(\hat{s}-s)^2]$ entre le signal ciblé et son estimé. Cet estimateur est fourni en filtrant les observations $x = s + v$ par le filtre de Wiener [fig. 3] dont les coefficients de la réponse impulsionnelle sont donnés par le produit de la matrice d'autocorrélation inverse des observations $\boldsymbol{\Gamma}_{xx}^{-1}$ et du vecteur d'intercorrélation entre les observations et le signal que l'on cherche à estimer $\boldsymbol{r}_{sx}$ [1] soit

$$\boldsymbol{w} = \boldsymbol{\Gamma}_{xx}^{-1}\,\boldsymbol{r}_{sx} \tag{1}$$

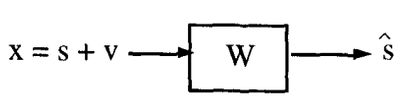


Figure 3. – Filtre de Wiener appliqué à l'observation $x = s + v$.

Dans le cas où s et v sont décorrélés, l'expression du filtre dans le domaine fréquentiel est

$$W(e^{j\omega}) = \frac{\gamma_{sx}(e^{j\omega})}{\gamma_{xx}(e^{j\omega})} = \frac{\gamma_{ss}(e^{j\omega})}{\gamma_{ss}(e^{j\omega}) + \gamma_{vv}(e^{j\omega})} \tag{2}$$

$\gamma_{xx}(e^{j\omega})$, $\gamma_{ss}(e^{j\omega})$ et $\gamma_{vv}(e^{j\omega})$ sont respectivement les densités spectrales de puissance (DSP) du signal bruité, du signal et du bruit. Notons que cette formule sous-entend une propriété essentielle du filtre de Wiener : son adaptation au niveau de bruit en tendant vers le filtre plat unitaire lorsque le bruit (i.e. $\gamma_{vv}(e^{j\omega})$) décroît pour s'annuler.

La détermination du filtre nécessite l'estimation des DSP $\gamma_{ss}(e^{j\omega})$ et $\gamma_{vv}(e^{j\omega})$. Différentes techniques existent pour l'estimation de la DSP du signal dans le bruit à partir des signaux acoustiques. En l'absence d'information a priori sur le signal, on a recours en général aux méthodes de débruitage par soustraction spectrale qui visent à soustraire la contribution du bruit au spectre du signal bruité [10, 29]. Ainsi, le calcul « classique » du filtre de Wiener suppose l'estimation de $\gamma_{xx}(e^{j\omega})$ et de $\gamma_{vv}(e^{j\omega})$, l'estimateur de $\gamma_{ss}(e^{j\omega})$ étant obtenu par la soustraction des deux. L'estimation de $\gamma_{vv}(e^{j\omega})$ traduit la nécessité d'une « référence bruit » supplémentaire. Si l'on ne possède pas de capteur supplémentaire pour le bruit, cette référence peut être obtenue pendant les silences de parole. Elle s'appuie alors sur des hypothèses assez fortes de stationnarité du bruit souvent éloignées des cas réels. C'est pourquoi les méthodes utilisant un second capteur sont en général plus performantes. Ce capteur supplémentaire fournit une « référence bruit » $b(k)$ fortement corrélée à $v(k)$. Le filtre de Wiener s'applique sur cette référence pour obtenir un estimé du bruit $v(k)$ que l'on soustrait à l'observation $x(k)$ [fig. 4].

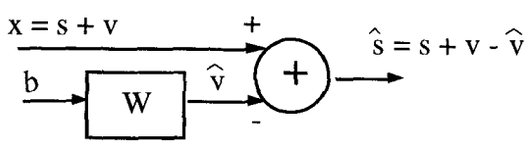


Figure 4. – Filtre de Wiener appliqué à un bruit de référence.

Cette structure dite « soustracteur de bruit » est fondée sur l'équivalence entre $E[(\hat{s}-s)^2]$ et $E[(\hat{v}-v)^2]$. La matrice d'autocorrélation des observations et le vecteur d'intercorrélation entre les observations et le bruit à estimer sont dans ce cas $\boldsymbol{\Gamma}_{bb}$ et $\boldsymbol{r}_{vb}$, et le filtre est donné par :

$$\boldsymbol{w} = \boldsymbol{\Gamma}_{bb}^{-1}\,\boldsymbol{r}_{vb} \tag{3}$$

En pratique, on suppose la décorrélation de s et b pour remplacer $\boldsymbol{r}_{vb}$ par $\boldsymbol{r}_{xb}$.

Dans notre étude, nous ne disposons pas d'une référence bruit, mais d'une référence signal $i(k)$ par le biais de l'image du locuteur, ce qui nous invite à envisager la structure duale de la précédente [fig. 5]. Il faut toutefois souligner une différence majeure entre les structures des figures 4 et 5. Le soustracteur de bruit s'appuie sur la corrélation entre b et v pour construire l'estimateur linéaire optimal et le filtre associé (ce qui revient à identifier le filtre physique entre b et v). Dans notre cas

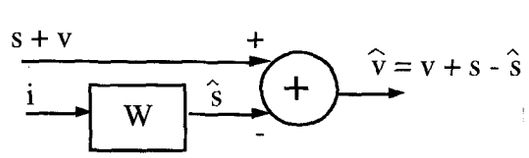


Figure 5. – Filtre de Wiener appliqué à un signal de référence.

le signal de référence est une image i, qui n'est corrélée à s que de façon très partielle. En effet, ces deux signaux ne sont pas de même nature d'une part, et d'autre part l'image ne fournit pas d'information sur l'excitation glottique ce qui signifie l'impossibilité d'accéder à la structure temporelle (i.e. la phase) du signal s. Par conséquent, il n'existe pas de filtre linéaire entre les signaux i et s et l'estimation de W (identification) est impossible. Le principe que nous proposons est donc d'introduire l'entrée supplémentaire image, non pas sous forme de référence supplémentaire, mais sous forme directe de connaissances sur le signal (voir la figure 2). Nous allons voir dans la section suivante comment réaliser cette idée dans le cadre d'un filtre de Wiener.

2.2. formulation du filtre de Wiener Labial

Les signaux s et v étant supposés décorrélés, l'expression du filtre Wiener dans le domaine fréquentiel est celle de la formule (2). Dans notre étude, puisque nous disposons d'une référence image, l'idée centrale est d'utiliser cette référence pour estimer $\gamma_{ss}(e^{j\omega})$. En d'autres termes, la DSP du signal est estimée non plus à partir du signal acoustique mais à partir des lèvres du locuteur. Le principe sous-jacent consiste à déterminer une relation entre descripteurs de l'image et paramètres spectraux modélisat $\gamma_{ss}(e^{j\omega})$. La procédure réalisant cette relation est exposée dans la section 3. En ce qui concerne la modélisation de $\gamma_{ss}(e^{j\omega})$, nous utilisons un modèle auto-régressif (LPC pour *Linear Predictive Coding*) qui a pour propriété de décrire correctement l'enveloppe spectrale du signal de parole [14, 15, 22]. Rappelons qu'un « spectre LPC » est donné par

$$S_{LPC}(z) = \frac{G}{1 + \sum_{i=1}^{p} a_i\, z^{-i}} = \frac{G}{A(z)} \tag{4}$$

$A(z)$ est un polynôme d'ordre p avec $1/A(z)$ analytique pour $|z| \geq 1$. Le gain G du modèle est déterminé tel que G^2 soit l'énergie du résidu sur la fenêtre d'analyse.

En notant N le nombre d'échantillons de la fenêtre d'analyse, la densité spectrale de puissance $\gamma_{ss}(e^{j\omega})$ estimée à partir du modèle LPC est donnée par

$$\gamma_{ss}(e^{j\omega}) = \left.\frac{|S_{LPC}(z)|^2}{N}\right|_{z=e^{j\omega}} = \left.\frac{G^2/N}{A(z)\,A(z^{-1})}\right|_{z=e^{j\omega}} \tag{5}$$

Ce sont ainsi les paramètres de $S_{LPC}(z)$ que nous devons déterminer à partir des lèvres du locuteur (section 3).

En ce qui concerne $\gamma_{vv}(e^{j\omega})$, nous abordons uniquement dans notre étude le cas simple d'un bruit blanc. Sa puissance moyenne σ_b^2 est connue car déduite du RSB dont nous considérons la connaissance comme hypothèse de travail. Dans ce cas, en notant $E_b = \sigma_b^2 N$ l'énergie du bruit sur la fenêtre d'analyse, la fonction de transfert numérique du filtre de Wiener utilisant le modèle LPC est donnée par

$$W(z) = \frac{\frac{G^2/N}{A(z)\,A(z^{-1})}}{\frac{G^2/N}{A(z)\,A(z^{-1})} + \sigma_b^2} = \frac{G^2/E_b}{G^2/E_b + A(z)\,A(z^{-1})} \tag{6}$$

Pour déterminer la forme causale de ce filtre, nous utilisons la procédure décrite par Papoulis [20, section 10.3, pp. 337–341] qui aboutit à la formule

$$W(z) = \frac{G^2}{E_b}\,\frac{B_1(z)}{A_1(z)} \tag{7}$$

où

$$A_1(z)\,A_1(z^{-1}) = G^2/E_b + A(z)\,A(z^{-1}) \tag{8}$$

et $B_1(z)$ est un polynôme de degré au plus $p-1$ qui intervient dans l'adaptation du filtre au niveau de bruit (cf. section 4.1).

En résumé, la mise en œuvre du filtre de Wiener « labial » repose essentiellement sur l'estimation du spectre $S_{LPC}(z)$ à partir des paramètres labiaux, la connaissance a priori du RSB, et l'implémentation causale selon la formule (7).

2.3. alternative d'un critère spectral : le filtre LPC

Le critère de minimisation d'erreur quadratique moyenne qui détermine le filtre de Wiener est dual dans le domaine temporel ou fréquentiel, i.e. si $S(z)$ et $\hat{S}(z)$ sont les transformées de z respectives de $s(k)$ et $\hat{s}(k)$ on a

$$E\left[(\hat{s}(k) - s(k))^2\right] \text{ minimale } \Leftrightarrow \oint E\left[\left|\hat{S}(z) - S(z)\right|^2\right] dz \text{ minimale} \tag{9}$$

Il s'agit d'un critère très rigoureux dans la mesure où il s'applique simultanément sur *le module et la phase* de ces spectres. Or on connaît particulièrement la pertinence du module du spectre du signal du point de vue du système auditif dont le fonctionnement repose pour une grande part sur une analyse spectrale du type bancs de filtres ([3] chapitre V). De plus, des études ont montré la prédominance du module du signal sur sa phase dans les domaines de la perception de signaux dégradés et du débruitage [29, 30]. C'est pourquoi nous proposons une alternative au filtre de Wiener en construisant un filtre capable de débruiter la parole par renforcement du spectre (en module) du signal. L'idée est d'utiliser le spectre LPC $S_{LPC}(z)$ comme fonction de transfert numérique d'un filtre rehausseur. L'utilisation de cette technique en tant que filtre correspond bien à une vision physique du débruitage par renforcement spectral. En effet, ce renforcement exige un filtre de module proche du spectre du signal. Or nous avons vu que la LPC fournit un modèle capable de décrire correctement l'enveloppe du spectre du signal et donc les formants de la parole que l'on veut renforcer. Notons que le critère spectral est particulièrement justifié dans le cas d'un bruit blanc prédominant par rapport au signal. En effet, si on note respectivement par $S(z)$ et

$V(z)$ les transformées en z du signal et du bruit et $H(z)$ le filtre débruitant, l'équation du filtrage est donnée par

$$\hat{S}(z) = H(z)\ (S(z) + V(z)) \tag{10}$$

Idéalement, le but du filtrage est de rehausser les formants du signal atténués par le bruit de telle façon que le signal filtré soit proche en module du signal original. Si on cherche pour cela à minimiser l'erreur quadratique en module entre le signal original et le signal filtré, soit

$$E_m = \oint E\left[\left(\left|\hat{S}(z)\right| - |S(z)|\right)^2\right] dz \tag{11}$$

on obtient théoriquement

$$|\hat{S}(z)| - |S(z)| = |H(z)|\,|S(z) + V(z)| - |S(z)| = 0 \tag{12}$$

avec

$$|H(z)| = \frac{|S(z)|}{|S(z) + V(z)|} \tag{13}$$

Dans le cas d'un bruit blanc prédominant d'énergie E_b sur la fenêtre d'analyse, l'équation (13) devient

$$|H(z)| = \frac{|S(z)|}{|V(z)|} = \frac{|S(z)|}{\sqrt{E_b}} \tag{14}$$

La minimisation de E_m impose bien dans ce cas un filtre dont le module est proportionnel à celui du spectre du signal. Notons que la constante multiplicative $1/\sqrt{E_b}$ traduit simplement le besoin de renormaliser le signal d'entrée en énergie pour se caler sur l'énergie du signal original. Cette remarque reste vraie pour tous les niveaux de bruit : on renormalise le signal d'entrée $x = s + v$ par son énergie E_x en adjoignant au filtre la constante $1/\sqrt{E_x}$.

En résumé, le second filtre que nous proposons est donné par

$$H(z) = \frac{S_{LPC}(z)}{\sqrt{E_x}} = \frac{G/\sqrt{E_x}}{1 + \sum_{i=1}^{p} a_i\, z^{-i}} = \frac{G/\sqrt{E_x}}{A(z)} \tag{15}$$

l'estimation des paramètres descripteurs du spectre LPC étant toujours réalisée, comme pour le filtre de Wiener, à partir de l'image du locuteur.

2.4. récapitulation comparative des deux filtres

Résumons les caractéristiques des deux filtres :

Filtre de Wiener

• Critère quadratique en module et phase entre les signaux original et filtré

+ Adaptation au RSB : convergence vers le filtre plat pour un bruit nul

– Nécessité d'hypothèses (contraintes) fortes sur le bruit : décorrélation par rapport au signal, blancheur, estimation du RSB

– Calcul de la forme causale complexe et coûteux.

Filtre LPC

• Critère « perceptif » de renforcement spectral (critère quadratique en module entre les signaux original et filtré dans le cas d'un bruit blanc et faible RSB)

+ Indépendance par rapport à la forme du bruit (renforcement spectral « robuste »)

+ Simplicité de mise en œuvre

– Non adaptation au RSB : traitement non pertinent pour les bruits faibles ou nuls.

Les deux filtres sont donc fondés sur deux critères distincts et possèdent des propriétés visiblement différentes. La différence majeure réside dans l'adaptation du filtre de Wiener en fonction du RSB. A l'inverse, l'utilisation du filtre LPC implique une indépendance par rapport au niveau de bruit. Ceci pose un problème dans le cas d'un bruit faible (et a fortiori nul), le filtrage étant alors plus synonyme de dégradation que de renforcement. Il nous faudra donc tester l'impact perceptif de cette dégradation et comparer l'efficacité de nos deux filtres indépendamment de tout problème d'estimation de leurs paramètres. Ceci fera l'objet de la première phase d'expérimentation de la section 4. Les deux filtres gardent l'avantage d'être estimés à partir de la même information extraite de la forme des lèvres : l'estimation du spectre LPC du signal.

3. l'associateur lèvres-spectre

Pour réaliser l'estimation du spectre LPC du signal à partir de l'image du locuteur, le problème consiste à élaborer un associateur entre deux jeux de paramètres : d'un côté des descripteurs du contour labial qui sont les entrées de notre associateur, de l'autre des paramètres représentatifs du spectre LPC correspondant qui sont les sorties estimées. Dans le cadre du modèle d'intégration audition-vision présenté en introduction, différentes méthodes d'association ont été envisagées telles que les réseaux de neurones, les méthodes de régression linéaire ou polynomiales. Ces méthodes associatives nécessitent toutes une phase préalable d'apprentissage où l'associateur est construit sur la base d'exemples. Notons d'ores et déjà la nécessité de se munir d'un corpus de travail suffisamment important pour se répartir en deux sous-corpus destinés à l'apprentissage de l'associateur d'une part, et aux différents tests menés sur notre système d'autre part. En ce qui concerne le méthode associative, nous avons opté pour une méthode classique de régression linéaire matricielle, les travaux de Robert-Ribes [23] ayant montré que ses performances dans un problème de même

type étaient très acceptables, et même meilleures que celles d'associateurs plus complexes (polynomiaux). Elle présente de plus l'énorme avantage de la simplicité, aussi bien dans sa formulation théorique que dans son implantation.

Après une description de notre corpus de travail dans la section 3.1, nous décrivons en détail dans la section 3.2 l'associateur lèvres-spectre proprement dit. Enfin la section 3.3 apporte une solution à un problème typique de gestes non-visibles qui intervient dans notre étude.

3.1. le corpus audiovisuel

Rappelons que nous présentons ici une étude de faisabilité du système global restreinte au cadre de voyelles stationnaires monolocuteur. Le corpus disponible est celui réalisé pour la validation des modèles d'intégration audition-vision de Robert-Ribes [23]. Il s'agit de l'enregistrement audiovisuel de 7 voyelles orales du français [a e i ø y o u] prononcées de manière tenue pendant deux secondes environ par notre locuteur. 700 stimuli acoustiques d'une durée de 200 ms ont été extraits de ces enregistrements, soit 100 stimuli par voyelle. Les enregistrements ont été réalisés sur le poste « visage-parole » de l'ICP [9]. Cette station d'acquisition de signaux audiovisuels de parole permet l'extraction automatique de trois paramètres pertinents du contour labial : l'étirement, l'aperture et l'aire intérolabials notés A, B et S [fig. 6]. Les enregistrements visuels sont réalisés avec des contraintes rigoureuses sur le locuteur : tête maintenue par un casque, éclairage idéal et maquillage bleu des lèvres pour l'extraction des paramètres labiaux grâce au procédé « Chroma-Key » [9]. Notre corpus est donc constitué de 700 mesures des paramètres géométriques $[A\,B\,S]$ synchrones au premier échantillon de chacun des 700 stimuli audio.

En ce qui concerne la représentation du spectre LPC de nos voyelles, nous avons choisi les coefficients PARCOR k_i [14, 15, 22] en nous appuyant sur leur robustesse par rapport à d'autres représentations telles que les coefficients de prédiction a_i du polynôme LPC. Cette robustesse au sens d'une grande stabilité du spectre par rapport à des variations des k_i est déjà largement mise en évidence dans le domaine du codage par exemple [14]. Les coefficients k_i sont fournis à partir du signal audio par l'algorithme de Durbin-Levinson appliqué à l'ordre 16 sur les 64 premières ms

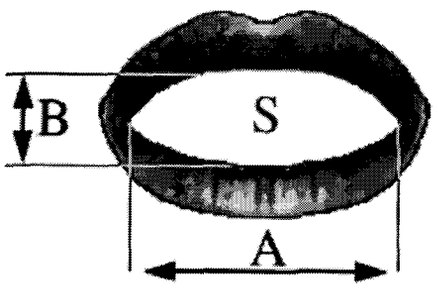


Figure 6. – Paramètres A, B et S du contour labial.

de chaque stimuli, algorithme qui garantit la stabilité du spectre [14, 15, 22]. Des résultats très similaires sur la forme des spectres de voyelles sont obtenus pour un ordre supérieur ou égal à 12 (cinq premiers formants et onde glottale bien modélisés).

Tout au long de l'expérimentation, la première moitié du corpus est utilisée pour la phase d'apprentissage (réglage des paramètres des différents éléments du système) et la seconde moitié pour la phase d'évaluation. Durant la phase d'apprentissage de notre associateur nous utilisons exclusivement les k_i calculés à partir des signaux acoustiques. Durant la phase de test, c'est–à–dire dans l'optique du débruitage labial proprement dit, ces coefficients sont estimés à partir des paramètres $[A\,B\,S]$. Ils peuvent néanmoins être comparés aux coefficients calculés à partir des signaux acoustiques de test.

Notons que lorsque l'on associe une forme labiale à des paramètres spectraux k_i, le gain G du modèle ne peut être estimé. Ceci traduit l'absence au niveau labial d'informations sur l'énergie du signal. Concernant le filtrage des voyelles stationnaires par le filtre LPC, ce problème est contourné en ajustant en sortie l'énergie des signaux filtrés de manière à les rendre écoutables. Par contre le calcul du filtre de Wiener nécessite la connaissance du rapport G^2/E_b. G et E_b ne sont pas directement accessibles, mais on peut calculer aisément G_{LPC} l'énergie de la réponse impulsionnelle du filtre tout–pôles $1/A(z)$ issu de notre associateur. En notant E_s l'énergie du signal modélisé par la LPC, c'est–à–dire celle de la réponse impulsionnelle de $G/A(z)$ on a

$$G_{LPC} = \frac{E_s}{G^2} \tag{16}$$

et alors

$$\frac{G^2}{E_b} = \frac{1}{E_b}\,\frac{E_s}{G_{LPC}} = \frac{RSB}{G_{LPC}} \tag{17}$$

Cette formule nous ramène à notre hypothèse de travail sur la connaissance a priori du RSB pour le calcul du filtre de Wiener.

En résumé notre corpus se compose de 100 répétitions de 7 voyelles tenues monolocuteur. Chaque exemplaire consiste en un triplet de paramètres vidéo $[A\,B\,S]$ associé à un signal audio de 200 ms et à 16 coefficients k_i, ces derniers étant calculés à partir du signal audio pour le corpus d'apprentissage et estimés à partir des paramètres $[A\,B\,S]$ pour le corpus de test.

3.2. l'associateur

L'objectif est donc l'élaboration d'un associateur linéaire entre l'espace d'entrée des paramètres géométriques $[A\,B\,S]$ et l'espace de sortie des coefficients PARCOR du spectre LPC du signal. Pour un associateur linéaire, la phase d'apprentissage consiste à effectuer la régression linéaire entre les deux matrices contenant les données du corpus d'apprentissage de l'un et l'autre espace. Cette régression fournit un associateur sous la forme d'une matrice de passage. Pendant la phase d'estimation, pour tout nouveau vecteur de paramètres d'entrée, on obtient le vecteur de sortie

associé par multiplication matricielle entre le vecteur d'entrée et la matrice de régression.

Rappelons le principe de la régression dite au sens des moindres carrés. Étant données une matrice d'entrée $\boldsymbol{L}$ et une matrice de sortie $\boldsymbol{K}$, de dimensions respectives $m \times n$ et $m \times p$, on cherche la matrice $\boldsymbol{F}$ telle que le produit matriciel $\boldsymbol{LF}$ approche au mieux la matrice $\boldsymbol{K}$ au sens où elle minimise l'erreur quadratique

$$e = \|\boldsymbol{LF} - \boldsymbol{K}\| \tag{18}$$

Dans le cas général, $\boldsymbol{F}$ est une matrice de dimension $n \times p$. Son calcul s'appuie sur des algorithmes de décompositions matricielles (Householder, Cholesky...) que nous ne détaillerons pas ici. Appliquons maintenant cette méthode pour la construction de notre associateur. Rappelons que sur les 700 paires audio-vidéo des sept voyelles, nous utilisons pour cette phase d'apprentissage les 350 premières dont nous avons extraits les coefficients k_i à partir du signal acoustique. La matrice $\boldsymbol{L}$ est ainsi issue de la concaténation des 50 premiers triplets $[A\,B\,S]$ de chaque voyelle, à laquelle on rajoute une colonne de 1 pour rendre compte du terme de translation affine. La matrice $\boldsymbol{K}$ regroupe les coefficients k_i modélisant le spectre de chacun des signaux audio correspondants. $\boldsymbol{K}$ contient donc les 50 lignes $\{k_1\ k_2 \ldots k_{16}\}$ de chaque voyelle. La régression entre $\boldsymbol{K}$ et $\boldsymbol{L}$ fournit la matrice-associateur $\boldsymbol{F}$ de dimension 4×16.

Pendant la phase de génération des « spectres labiaux », pour tout nouveau triplet-vecteur $[A_0\ B_0\ S_0]$ provenant des signaux vidéo du corpus de test, on obtient une estimation $[k_{01}\ k_{02} \ldots k_{016}]$ des k_i du spectre correspondant par

$$[A_0\ B_0\ S_0\ 1]\,\boldsymbol{F} = [k_{01}\ k_{02} \ldots k_{016}] \tag{19}$$

Un problème majeur apparaît dans le caractère non-bijectif de la relation son-image. En effet, certaines voyelles ont des sons (spectres) différents, mais une même forme de lèvres. C'est le cas des voyelles [y] (son de « du ») et [u] (son de « doux ») d'une part, et des voyelles [ø] (son de « deux ») et [o] (son de « dos ») d'autre part. Dans notre application ces ambiguïtés sont ingérables au niveau de la régression linéaire qui ne peut réaliser l'apprentissage de k_i différents avec des entrées semblables. La solution que nous avons choisie consiste à construire deux associateurs traitant respectivement les voyelles [ø y] d'une part et [o u] d'autre part. Nous avons ainsi séparé les sept voyelles en deux groupes de cinq ne comportant respectivement que les voyelles « antérieures », soit [a e i ø y], ou que les voyelles « périphériques », soit [a e i o u] (termes choisis en référence aux catégories phonétiques classiques [3]). Puis nous calculons l'associateur propre à chaque groupe en effectuant deux régressions : entre matrices $\boldsymbol{L}$ et $\boldsymbol{K}$ « antérieures » d'une part, et « périphériques » d'autre part. Notons que les matrices $\boldsymbol{L}$ et $\boldsymbol{K}$ sont de tailles respectives 250×4 et 250×16 ($250 = 5$ voyelles $\times$ 50 échantillons). Nous obtenons donc deux matrices de régression de taille 4×16 qui vont nous fournir deux spectres possibles pour un même jeu de paramètres $[A\,B\,S]$. Il s'agit alors de construire un module destiné à choisir le « bon spectre » parmi les deux proposés. Notons que pour les occurrences de [a e i] aux formes labiales non ambiguës, les spectres estimés par les deux associateurs sont très proches et le choix entre les deux importe peu. Par conséquent, nous allons centrer le problème d'un sélecteur de spectre uniquement sur les couples ambigus [ø o] et [y u].

3.3. le sélecteur antérieures/ périphériques

Le problème de l'ambiguïté entre sosies labiaux provenant justement de la modalité visuelle, il est logique de rechercher dans la modalité auditive les indices susceptibles de nous faire pencher en faveur de l'un des deux filtres. C'est pourquoi nous avons fondé notre sélecteur sur la forme du spectre du signal audio en remarquant que les spectres antérieurs de [ø] et [y] sont « plutôt hautes fréquences » alors que les spectres périphériques de [o] et [u] apparaissent comparativement « plutôt basses fréquences ». Nous sommes restés dans le cadre simple des méthodes linéaires en utilisant une analyse discriminante dans un espace spectral. Chaque spectre est défini sur 20 valeurs en dB, l'échelle fréquentielle choisie étant l'échelle « perceptive » des bark, avec

$$z(bark) = 7 \arg\, sh(f(Hz)/650) \tag{20}$$

d'après la formule de Schroeder *et al.* [24]. Cette échelle est adaptée aux caractéristiques de l'oreille humaine, et elle permet de compacter sur un petit nombre de valeurs (20 points de 1 à 20 bark) un spectre s'étalant en fréquence de 0 à 5000 Hz grâce à une compression quasi-logarithmique des hautes fréquences. Les intensités sont normalisées entre 0 et 1, la valeur 1 correspondant au maximum du spectre en dB et 0 à ce maximum moins 50 dB, toute valeur inférieure à ce seuil étant mise à 0. Les paramètres de l'analyse discriminante sont déterminés, comme toujours, sur la première moitié du corpus, comprenant donc 100 spectres de voyelles antérieures (50 [ø] et 50 [y]) et 100 spectres de voyelles périphériques (50 [o] et 50 [u]). De plus, 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 blanc allant du non bruité à un bruit assez fort mais maintenant une assez bonne séparation des 2 classes. En définissant le RSB comme rapport entre les énergies moyennes de la voyelle et du bruit, les 5 niveaux choisis sont : non bruité, 24, 12, 6 et 0 dB.

Des tests préliminaires ont montré que l'ajout des paramètres labiaux $[A\,B\,S]$, que nous avions éliminés au départ pour la sélection avant-arrière, pouvaient fournir un gain sensible sur les performances du sélecteur, ce dernier étant capable de se servir de l'information distinctive entre les couples [ø o] et [y u] pour mieux « focaliser » sa décision.

En résumé, nous avons implanté une analyse discriminante à 2 classes, avec dans chaque classe 500 vecteurs d'apprentissage (100×5 niveaux de bruit), chacun de dimension 23 (20

paramètres spectraux, 3 paramètres visuels). Le seuil de décision λ est le barycentre des projections sur l'axe discriminant des moyennes de chaque classe, pondérées par les écarts-type des nuages de chaque classe du corpus d'apprentissage par rapport à ces moyennes. Les résultats de l'évaluation de ce sélecteur en termes de taux de classification correcte sont fournis en 4.2.2.

4. évaluation du système

Notre système comporte deux modules de base, le filtre et l'associateur (incluant le sélecteur). Nous avons d'abord testé séparément chaque module avant d'évaluer le système dans son ensemble. Il est important d'insister ici sur la variété des méthodologies mises en jeu lors de ces tests. En particulier, l'évaluation des filtres en section 4.1 passe par l'utilisation d'un corpus plus général que celui des voyelles stationnaires seules qui sert de cadre à notre étude. Par contre, c'est bien sur ce corpus de voyelles que sont évalués l'associateur (section 4.2) et le système complet (section 4.3). Notons que nous n'avons pas mené d'études comparatives avec les méthodes traditionnelles de débruitage fondées sur l'utilisation d'une ou plusieurs références acoustiques. Ce choix est justifié par la simplicité de notre corpus qui impliquerait à coup sûr des scores extrêmement élevés pour les méthodes classiques et dans ce cadre notre système s'avérerait peu compétitif. En revanche, ce type de comparaison sera bienvenu si la poursuite de notre travail sur des séquences de parole continue montre des résultats significatifs.

4.1. comparaison qualitative des deux filtres en terme d'efficacité perceptive

Nous avons évoqué dans la section 2.3, les différentes propriétés des deux filtres et notamment la non-adaptation au RSB du filtre LPC. Celle-ci pose le problème de l'effet du filtrage sur la partie parole du signal bruité et en particulier sur la parole propre. Dans le but de juger qualitativement de cet effet en terme d'intelligibilité et de qualité, et afin de valider de manière générale les deux techniques LPC et Wiener, nous avons mené une phase de tests informels durant laquelle les filtres sont estimés *directement à partir du signal audio propre*. Il s'agit donc ici d'un filtrage idéal au sens où l'on connaît le signal acoustique propre et on utilise directement celui-ci pour générer nos filtres. Il n'est pas encore question dans ces premiers tests d'estimation labiale. En conséquence, nous avons mené ces tests sur de la parole continue, plus apte à fournir un support de jugement qualitatif. Les signaux traités pour cette expérience sont des phrases « phonétiquement équilibrées » et des logatomes (enchaînements phonétiques articulatoirement plausibles mais dépourvus de sens) avec une fréquence d'échantillonnage de 16 kHz. Le processus de construction de filtres et de filtrage est appliqué de façon séquentielle sur des segments adjacents de signaux, la jonction entre les segments filtrés étant assurée par une fenêtre de pondération trapézoïdale avec recouvrement sur le segment linéaire de jonction. La longueur de la fenêtre d'analyse est de 256 échantillons (soit 16 ms) avec un recouvrement de l'ordre de 30 échantillons. Sur chaque segment le spectre LPC du signal est obtenu par l'algorithme de Durbin-Levinson appliqué à l'ordre 16.

4.1.1. *cas des signaux non-bruités*

Les premiers tests ont porté sur de la parole propre non-bruitée. L'écoute nous a permis de découvrir la robustesse de la parole face à ce traitement. Les résultats du filtrage LPC s'avèrent très satisfaisants en terme d'intelligibilité. Le contenu du message est parfaitement conservé ainsi que la mélodie (fréquence fondamentale). En terme de qualité, la déformation due au filtrage se ressent sensiblement dans la sonorité qu'on aurait envie de qualifier de « casserole ». Tout le long de la phrase la sensation qui intervient par rapport à l'écoute de l'original est celle d'un assourdissement conjugué avec l'apparition sur les voyelles d'une sorte de vibrato « dédoublant » le signal. Ce dernier évoque le cas typique d'un signal ayant subi une perturbation sur sa phase, phénomène qui advient doublement pendant notre filtrage : la modélisation LPC ne respecte pas la phase du signal et le filtrage ajoute cette phase perturbante au signal filtré. L'assourdissement provient lui de l'accentuation de la pente des spectres par le filtrage, ceci posant problème surtout pour les spectres descendants (sons voisés qui semblent responsables de la sonorité globale de la phrase) [fig. 7].

Une compensation hautes fréquences sur toute la phrase par un filtre du premier ordre représente une solution simple et efficace pour atténuer cet assourdissement. On améliore ainsi le confort

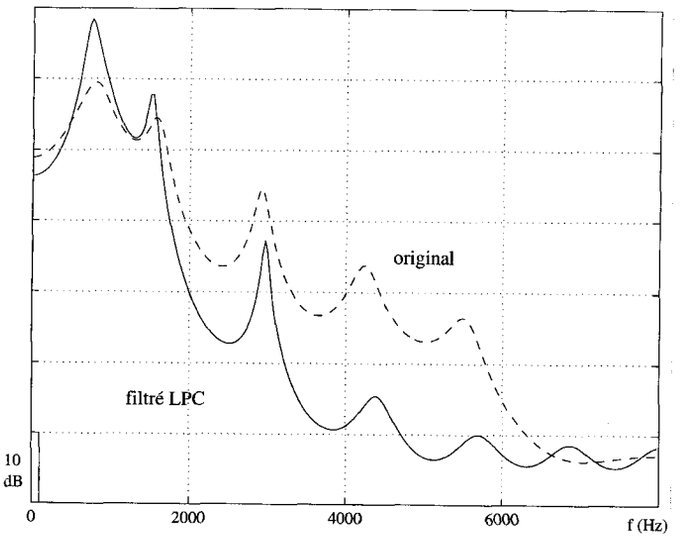


Figure 7. – Premier [a] de [baba] : spectres du signal original et du signal original renormalisé en énergie et filtré LPC.

d'écoute (la phrase paraît plus claire) sans changer l'aspect « casserole ».

En ce qui concerne le filtre de Wiener, les résultats sont sans surprise : le filtrage est plat et le signal est totalement conservé, ce qui représentait un des avantages de ce filtre.

4.1.2. cas des signaux bruités

Dans le cas de signaux bruités, rappelons que le filtre de Wiener est calculé à partir du spectre LPC selon la procédure abordée en 2.2. Celle-ci sous-entend les hypothèses de travail d'un bruit blanc additif et de la connaissance a priori du RSB sur chaque fenêtre d'estimation (c'est-à-dire du rapport G^2/E_b). En pratique, les bruits générés pour ces tests sont d'énergie constante sur toute la phrase, leurs niveaux étant déterminés par un RSB moyen global défini par rapport à l'énergie moyenne de la phrase. C'est le RSB global qui est mentionné dans les résultats de ces premiers tests. En revanche le RSB sur chaque fenêtre d'estimation est donné par les connaissances a priori extraites directement sur les signaux audio.

Dans le cas de signaux bruités, le signal filtré linéairement peut être considéré comme la somme du filtrage du signal propre original et du filtrage du bruit blanc additif. Ceci est particulièrement observable pour le filtre LPC, qui, pour des RSB raisonnables (supérieurs à -12 dB) permet à l'oreille de séparer la contribution de chacune des entrées : on sent la superposition d'un signal « casserole » semblable à celui décrit précédemment et d'un signal « fortement chuchoté » issu de l'excitation du filtre par le bruit. L'avantage de la LPC réside ainsi dans la conversion de la contribution du bruit en entrée en un signal utile sous forme de parole chuchotée (effet « vocodeur »). Au fur et à mesure de l'augmentation du bruit, sa contribution devient prédominante : un indice pertinent pour s'en rendre compte est la disparition progressive de la mélodie de la phrase provenant du fondamental. La voix chuchotée devient de plus en plus râpeuse et monocorde. Pour un RSB de l'ordre de -12 dB le filtrage conserve tant bien que mal quelques indices du signal propre original (notamment la mélodie). Enfin dans le cas très fortement bruité (RSB inférieur à -18 dB), la LPC trouve logiquement tout son intérêt : pour un signal d'entrée totalement inintelligible, le signal filtré révèle le contenu complet du message. Il s'agit dans ce cas du signal résultant de l'excitation d'une fonction de transfert modélisant la source glottique et le conduit vocal par un bruit blanc. Si les informations du type voisement, phase ou timbre du locuteur sont sérieusement altérées voire perdues, l'objectif de gain d'intelligibilité est largement atteint.

Pour le filtre de Wiener, les résultats sont aussi différemment appréciés suivant le niveau de bruit, mais dans des proportions inverses. L'amélioration s'est située ici essentiellement dans le cas faiblement bruité ce qui s'explique en détaillant le mécanisme de l'adaptation du filtre au RSB . Celle-ci se traduit en effet par deux phénomènes simultanés : le déplacement des pôles du filtre (racines de $A_1(z)$) par rapport aux pôles LPC (racines de $A(z)$) à l'intérieur du cercle unité sous l'influence du rapport G^2/E_b et la compensation des résonances correspondant à ces pôles par les racines de $B_1(z)$ (voir formule (7)). 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 aplati tendant vers le filtre plat unitaire. Quand le bruit augmente et devient prédominant, les pôles du filtre tendent vers ceux de la LPC du signal ($A_1(z)$ tend vers $A(z)$), ce qui explique que l'on retrouve des caractéristiques spectrales du signal. Cependant, lorsque le bruit augmente, la phase du signal filtré n'est plus contrôlable (fort bruit de phase aléatoire en entrée). La soustraction entre le signal et son estimé filtré dans le calcul de l'erreur quadratique devient dans ce cas une « addition » vectorielle. La minimisation de cette erreur impose la minimisation en module du signal filtré, et ceci à toutes les fréquences. Le comportement du filtre est alors double : le gain G^2/E_b devient très petit et les zéros du numérateur continuent à compenser les pôles. Cette dernière remarque est particulièrement cruciale dans le cas d'un RSB très faible, où les pôles du filtre rejoignent ceux du spectre LPC (sur la figure 8, la forme limite du filtre est quasiment atteinte pour un RSB de -20 dB, et le gain du filtre est inclus dans le numérateur).

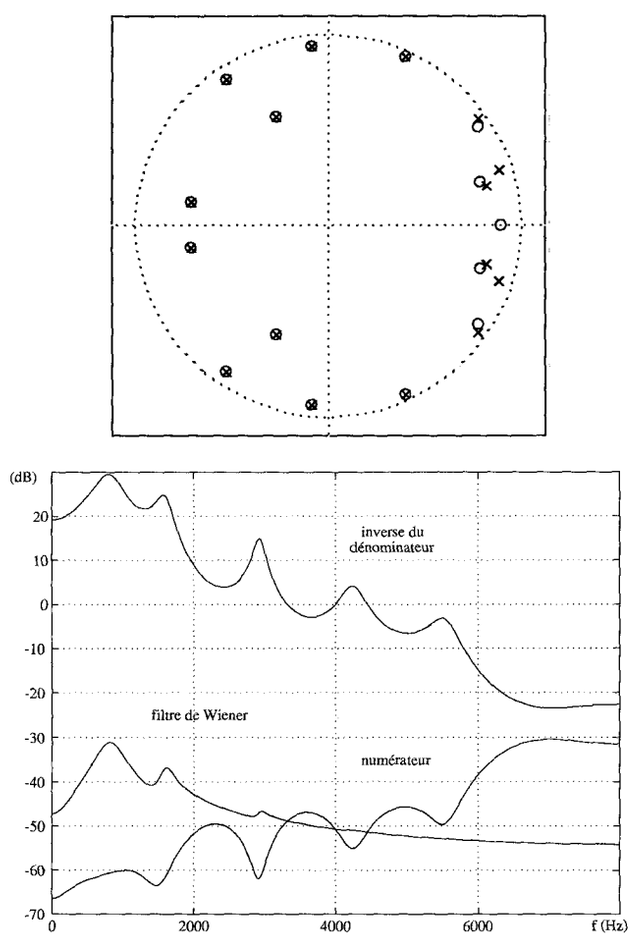


Figure 8. – Premier [a] de [baba] : filtre de Wiener causal pour RSB = -20 dB
a) Pôles (X) et zéros (O) du filtre
b) Numérateur, inverse du dénominateur et produit des deux.

La première conséquence du filtre de Wiener plus plat que le filtre LPC est de conserver davantage les caractéristiques de la parole originale. En contrepartie le débruitage perd de son efficacité : bien qu'atténué le bruit est toujours présent en sortie, comme greffé au signal utile. A l'inverse de la LPC qui transformait le bruit (utilisé comme excitation) en parole chuchotée plus ou moins prédominante selon le RSB d'entrée, le filtre de Wiener réalise une sorte de compromis dans la conservation du bruit et de la parole originale en tentant d'avantager cette dernière et de la restituer plus fidèlement que le filtre LPC. Malgré ce déficit d'efficacité par rapport au filtre LPC, le gain d'intelligibilité est notable puisque le filtrage de Wiener parvient à dégager l'information utile d'un signal totalement inintelligible, ce qui, compte tenu des résultats probants du cas faiblement bruité, nous permet de valider cette technique de filtrage.

En résumé ces tests ont montré la répartition des performances des deux méthodes aux extrémités de l'axe faiblement-fortement bruité telle qu'on pouvait la pressentir à l'issue de l'étude théorique de ces filtres : un avantage pour la LPC dans le cas fortement bruité (dû au critère spectral) et préférence pour Wiener pour des bruits faibles (due à l'adaptation au RSB). Tout en validant ces deux méthodes, nous ne devons pas oublier que dans ces tests pilotes, la qualité du débruitage provient principalement des conditions optimales d'obtention des filtres (à partir du signal propre). Nous allons maintenant passer au cas plus délicat qui nous préoccupe principalement, c'est-à-dire celui où l'estimation des filtres est réalisée à partir des lèvres.

4.2. comportement de l'associateur

Nous reprenons à présent notre corpus de voyelles stationnaires de la section 3.1 en rappelant que les différents éléments de l'associateur (deux régressions « antérieure » et « périphérique » + un sélecteur) ont été déterminés à partir de la première moitié de ce corpus (les 50 premiers échantillons de chaque classe vocalique). Les tests que nous décrivons ici portent naturellement sur l'autre moitié.

4.2.1. *normalisation des coefficients k_i*

Lorsqu'on applique l'une ou l'autre des deux matrices de régression sur les valeurs de $[A\ B\ S]$ des voyelles de test, la contrainte nécessaire à la stabilité du spectre estimé d'obtenir des valeurs de k_i comprises dans l'intervalle]-1, 1[n'est pas forcément assurée. Cependant, on constate que le pourcentage de k_i extérieurs à cet intervalle est inférieur à 1% sur l'ensemble du corpus de test. Dans ces cas rares, nous renormalisons la valeur absolue des k_i incriminés à la valeur 0.99. La robustesse de ces coefficients, qui avait guidé notre choix, permet d'obtenir un spectre estimé tout à fait acceptable en comparaison avec le spectre estimé à partir du signal audio. Ce résultat ne peut être obtenu avec d'autres représentations LPC moins robustes. En particulier la représentation en pôles que nous avons testée dans une phase initiale de ce travail se montre beaucoup trop sensible par rapport au module de ces pôles et fournit de moins bons résultats que les coefficients PARCOR. Nous avons également utilisé les coefficients de rapports logarithmiques de sections d'aires (Log Area Ratio) qui évitent intrinsèquement ce problème de renormalisation par un redéploiement des ki sur l'intervalle $]-\infty\ +\infty[$ grâce à la formule

$$LAR_i = \log \frac{1+k_i}{1-k_i} \tag{21}$$

Cependant nous avons constaté que les spectres issus de la prédiction des k_i avec ou sans renormalisation et ceux issus de l'interface LAR sont extrêmement similaires comparés au spectre estimé à partir du signal acoustique correspondant. Ceci montre que notre associateur linéaire n'exploite que très peu la transformation (pas du tout sur le domaine où celle-ci est linéaire) d'une part, et d'autre part que la procédure de renormalisation fondée sur la robustesse des k_i est largement satisfaisante surtout compte tenu de la très faible proportion des coefficients singuliers. Même si la représentation LAR reste potentiellement envisageable pour de futurs développements dans un cadre plus complexe, la représentation PARCOR reste bien adaptée à ce stade de notre application.

4.2.2. *validation du sélecteur antérieures/périphériques*

L'analyse discriminante à 2 classes ayant permis de déterminer un axe discriminant entre voyelles antérieures et voyelles périphériques selon la procédure décrite en section 3.3, nous présentons en figure 9 la projection des stimuli de l'ensemble d'apprentissage sur cet axe discriminant. Les stimuli sont présentés pour les 9 niveaux de bruit utilisés pour l'apprentissage du seuil de décision λ. Ceux-ci comprennent les 5 niveaux d'apprentissage du sélecteur auxquels on ajoute 4 niveaux plus bruités afin de tenir compte pour λ du regroupement des nuages en condition très bruitée.

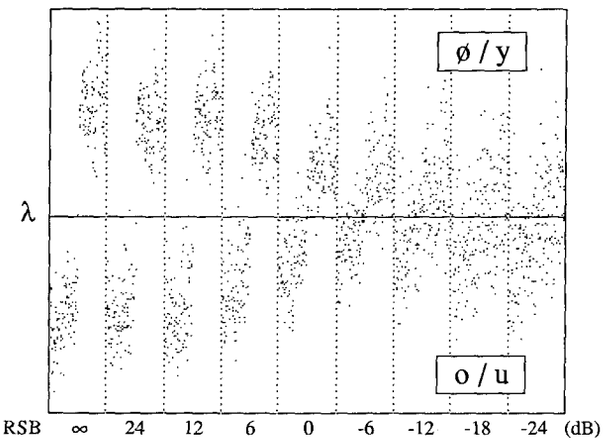


Figure 9. – Répartition des projections du corpus d'apprentissage sur l'axe discriminant pour 9 niveaux de bruit.

Les performances du sélecteur en terme de taux de classification correcte ont été évaluées avec la deuxième moitié du corpus non utilisée pour l'apprentissage. On utilise les mêmes 9 niveaux de bruit que ci-dessus, les 4 niveaux plus forts étant destinés à juger des qualités de notre sélecteur dans des conditions plus dégradées et non-apprises. Les pourcentages de choix correct globaux et relatifs à chaque voyelle sont résumés dans le tableau 1. Les scores s'avèrent très satisfaisants : ils sont notamment supérieurs à 90% pour RSB $\geq$ 0 dB et dépassent largement l'aléatoire (50%) pour les signaux très bruités (environ 65% à -18 et -24 dB).

Tableau 1. – Taux de classification correcte du sélecteur antérieures / périphériques.

RSB	∞	24	12	6	0	-6	-12	-18	-24
o	98	98	100	98	94	52	36	34	48
u	98	100	100	94	84	74	84	90	94
ø	100	100	100	98	96	94	96	90	86
y	100	100	100	100	98	90	56	42	32
moy	99	99,5	100	97,5	93	77,5	68	64	65

4.2.3. *comportement d'ensemble de l'associateur*

Le comportement de l'associateur sur le corpus de test est globalement bon. En effet, les spectres estimés ont une allure correcte et l'associateur 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. Pour notre locuteur et parmi le corpus total de nos sept voyelles, il existe trois regroupements visuels autour des distinctions ouvertes [a], étirées [e i] et arrondies [ø y] ou [o u] qui ne correspondent pas exactement à la répartition des formants dans l'espace acoustique comme le montre la figure 10 pour le groupe des voyelles périphériques.

Même si le choix d'un associateur linéaire est globalement justifié, ces regroupements provoquent un phénomène de mélange et de dispersion des spectres estimés associés à des phonèmes auditivement distincts mais labialement proches [fig.11].

Les deux régressions (« antérieure » et « périphérique ») réalisent plutôt correctement la distinction entre [ø] et [y] ou [o] et [u] mais tendent à confondre les voyelles [e] et [i]. On aboutit souvent dans cet exemple à l'estimation d'un spectre hybride dont l'allure est un mélange des formants des deux phonèmes [fig.12].

4.3. résultat d'évaluation du système complet

Ayant évalué les deux modules indépendamment, nous avons testé le système complet de débruitage de voyelles, qui inclut l'estimation des filtres par les lèvres. Nous rappelons que les bruits utilisés sont additifs et blancs. Le RSB est défini comme rapport entre les énergies moyennes de la voyelle et du bruit, et il est supposé connu pour le calcul du filtre de Wiener.

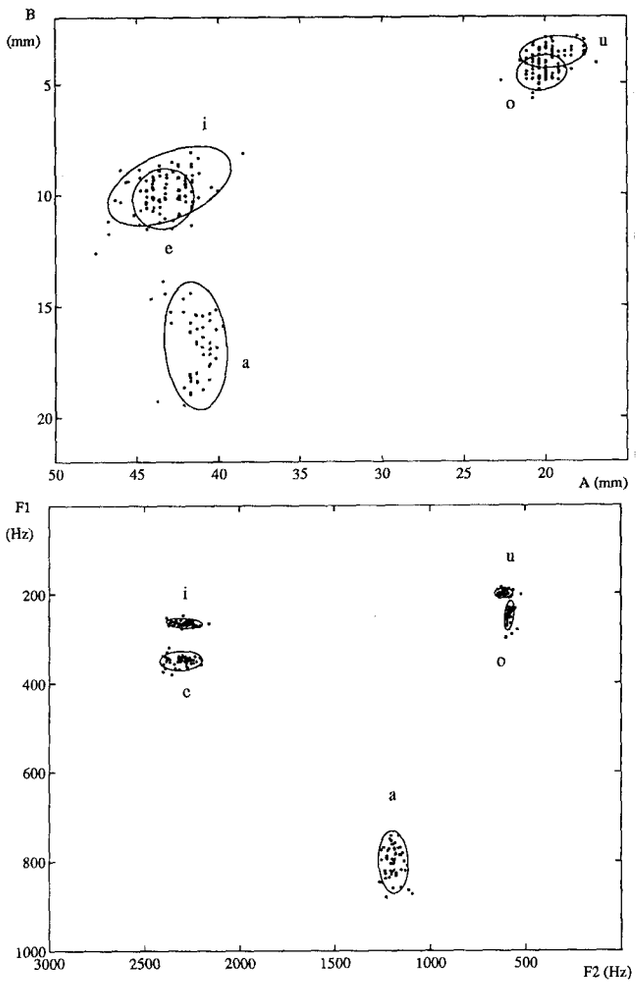


Figure 10. – Répartition des voyelles d'apprentissage
a) dans l'espace visuel A/B
b) dans l'espace acoustique F1/F2 : les formants sont détectés sur les spectres LPC audio.

Les résultats du filtrage des voyelles du corpus de test confirment ceux des phases de test précédentes. En ce qui concerne le filtre LPC, l'amélioration est très sensible dans le cas fortement bruité (RSB < -12 dB) pour lequel le filtrage transforme des signaux inaudibles en voyelles. On retrouve dans ce cas une voyelle chuchotée qui n'est autre que le résultat de l'excitation d'un filtre formantique par un bruit. Cet aspect reste un problème du point de vue perceptif, ce caractère non voisé peu habituel pouvant porter préjudice à l'appréciation de la voyelle. Quand au filtrage de Wiener, il se caractérise à nouveau par sa pertinence pour les RSB forts (filtrage quasi plat) et son aspect « moins violent » que le filtrage LPC pour les RSB faibles. A titre d'exemple, on remarque sur la figure 13 le faible rehaussement spectral du signal filtré par Wiener par rapport au même signal filtré par le filtre LPC (signaux non renormalisés en énergie). On retrouve par ailleurs le déplacement de formants par rapport au signal original (cf. 4.2.3 et fig. 12).

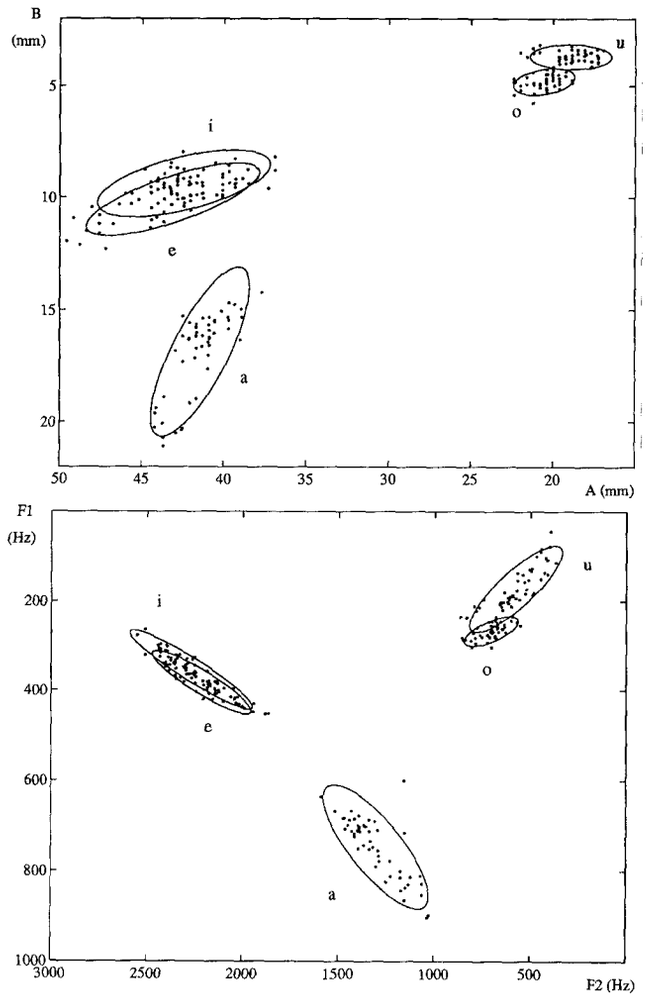


Figure 11. – Répartition des voyelles de test
a) dans l'espace visuel A/B
b) dans l'espace acoustique F1/F2 pour les spectres « vidéo » : les formants sont détectés sur les spectres LPC estimés à partir de [A B S].

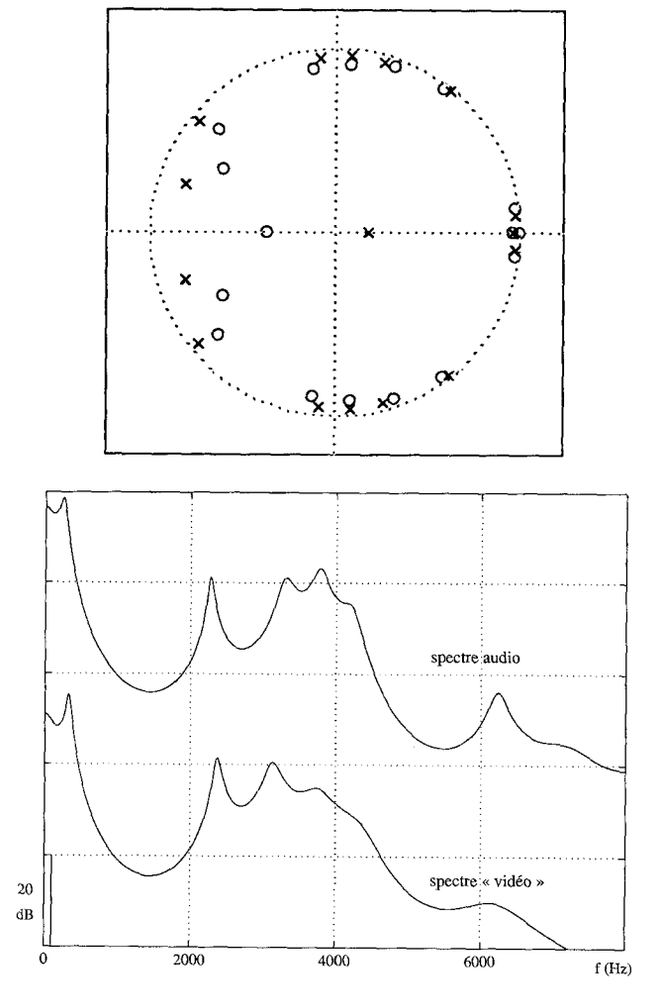


Figure 12. – Exemple de la voyelle [i]
a) Pôles des spectres LPC *audio* (X) et estimé à partir des lèvres (O)
b) Spectres correspondants : audio et « vidéo » (le décalage énergétique est dû à l'absence du gain G pour le spectre « vidéo »).

Pour quantifier les résultats du débruitage, nous avons mené un test perceptif d'identification sur les échantillons de nos sept voyelles non utilisés pendant la phase d'apprentissage. Le test a consisté à soumettre à 20 sujets 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). On demande aux auditeurs de classifier chaque stimulus entendu dans l'une des 7 catégories vocaliques [a e i ø y o u]. Ainsi, le test est dit « à choix forcé » (pas de possibilité « ne sait pas »). Les auditeurs sont de langue maternelle française et ont tous réalisé le test dans les mêmes conditions matérielles (chambre sourde, notice, choix par clavier, prétest d'accoutumance...). La figure 14 présente les résultats de ce test en terme de pourcentage d'identification correcte. Notons que chaque point des courbes de résultats représente 700 stimuli différents (5 par voyelle et par série, 7 voyelles, 20 auditeurs).

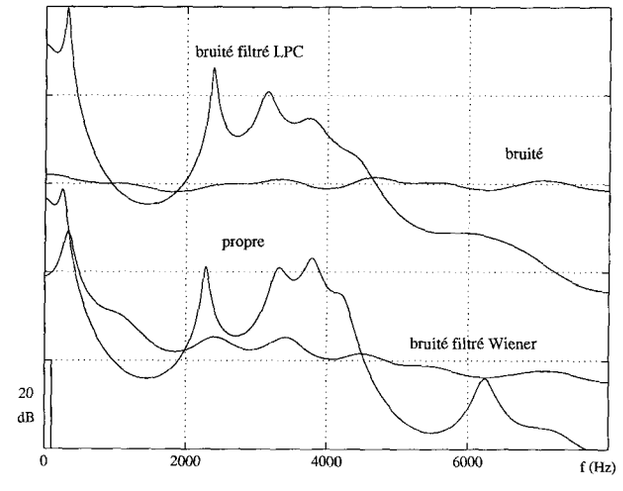


Figure 13. – Exemple de la voyelle [i] – RSB = -18 dB
Spectres du signal original propre et bruité et du signal bruité filtré LPC et Wiener.

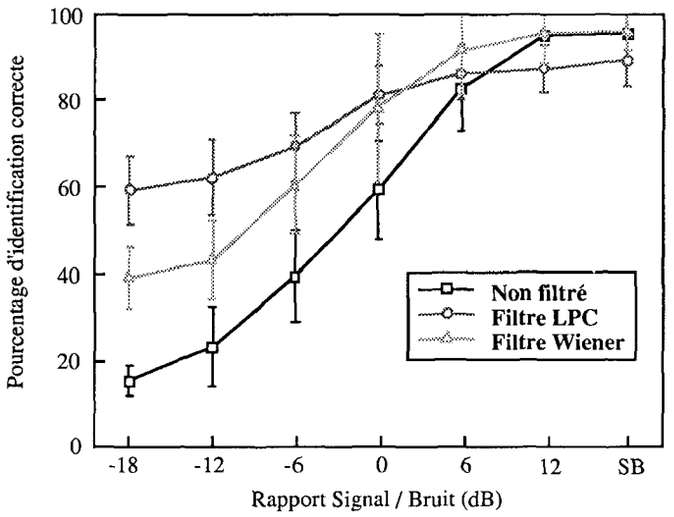


Figure 14. – Pourcentages d'identification correcte du test perceptif.

Les résultats sont très satisfaisants. En effet, 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. Pour ces mêmes RSB faibles les scores du filtrage de Wiener sont sensiblement en retrait par rapport au filtre LPC (38,5% et 43% pour des gains de 23,5% et 20%). En revanche, 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, vraisemblablement à cause des limites des associateurs lèvres-spectres (voir 4.2). On retrouve bien la répartition des performances des deux techniques le long de l'axe faible-fort bruit. Par ailleurs, on remarque qu'à -18 dB le pourcentage de reconnaissance globale sur signaux non filtrés (15%) est proche du pourcentage de choix aléatoire de 14,3%, le signal utile étant presque totalement noyé dans le bruit.

Les résultats ont été rassemblés pour chaque RSB dans des matrices de confusions qui regroupent le nombre d'identifications de la voyelle j (en colonne) sur présentation de la voyelle i (en ligne). Nous présentons dans le tableau 2 les matrices obtenues pour les conditions non-filtré et filtrage LPC pour les RSB ∞, 0, -6 et -18 dB. Ces matrices montrent que l'identification des signaux filtrés reste efficace au-delà de la reconnaissance exacte. On assiste en effet à une concentration des valeurs autour de la diagonale en trois blocs [a], [e i] et [ø y o u] qui reflètent 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 entre les deux classes [ø y] et [o u] que l'on cherche à départager grâce au sélecteur antérieures/périphériques, et à l'intérieur de chaque classe étant donnée la séparation visuelle limitée entre voyelles hautes et mi-hautes ([i] vs [e], [y] vs [ø], [u] vs [o]).

5. discussion

5.1. une synergie audiovisuelle confirmée

Le premier résultat positif qui ressort de cette étude est la confirmation d'une réelle synergie audiovisuelle qui découle de notre processus de filtrage. La récupération des informations auditives utilisées en entrée de nos deux filtres est effective puisque les courbes d'identification sont croissantes en fonction du RSB. Le signal d'origine bruité intervient ainsi pour une grande part dans le rehaussement, et ceci de façon de plus en plus prédominante au fur et à mesure de l'augmentation du RSB. Le processus de filtrage réalise donc intrinsèquement une « adaptation automatique » en fonction des informations disponibles. Ceci est particulièrement clair pour chaque domaine de prédilection de nos deux filtres. Dans le cas faiblement bruité, le filtre de Wiener favorise l'information acoustique en tendant vers le filtre plat. Cet effacement des informations visuelles n'est pas réalisé par le filtre LPC qui induit dans ce cas une faible dégradation de l'intelligibilité des signaux. Mais ce filtre se montre le plus efficace dans le cas fortement bruité où il s'agit de favoriser l'information visuelle. Le cas extrême est atteint lorsque le signal est complètement perdu dans le bruit et que l'on filtre un bruit à travers un filtre formantique (effet « vocodeur »). Mais ce cas précis ne peut réduire notre filtrage à un artefact fondé sur les seules informations visuelles. D'une part dans ce dernier cas extrême, les informations visuelles sont bien les seules disponibles, et d'autre part pour les niveaux de bruit plus faibles, le processus de filtrage permet une réelle complémentarité entre les informations auditives et visuelles. Nous nous rapprochons de ce point de vue de l'étude des phénomènes d'intégration audiovisuelle dans le sens où nous cherchons à vérifier la supériorité de l'audio plus vidéo sur l'audio seul et le vidéo seul (rappelons que seul le cas du filtrage LPC à faible niveau de bruit ne satisfait pas ce paradigme). Dans cette optique, il est intéressant de comparer nos scores avec ceux de Robert-Ribes [23] qui a mené un test perceptif audiovisuel sur le même corpus de voyelles pour mesurer le gain entre la condition audiovisuelle et la condition auditive seule. Les résultats de notre filtrage sont très proches de ceux de Robert-Ribes. Ceci 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).

5.2. vers une application plus réaliste

Malgré des résultats très prometteurs, l'étude de faisabilité que nous venons de présenter est encore très restreinte étant donné la simplicité des stimuli (voyelles monolocuteur) et les contraintes

∞	a	e	i	ø	y	o	u
a	**99**	0	0	0	0	0	0
e	1	**100**	5	0	0	0	0
i	0	0	**95**	0	0	0	0
ø	0	0	0	**100**	3	0	0
y	0	0	0	0	**96**	0	0
o	1	0	0	0	0	**99**	22
u	0	0	0	0	1	1	**78**

∞	a	e	i	ø	y	o	u
a	**100**	0	0	0	0	0	0
e	0	**97**	22	0	0	0	0
i	0	0	**78**	2	0	0	0
ø	0	1	0	**69**	1	0	0
y	0	1	0	4	**98**	1	4
o	0	1	0	4	0	**86**	0
u	0	0	0	21	1	13	**96**

0	a	e	i	ø	y	o	u
a	**99**	0	0	0	0	0	0
e	1	**83**	2	63	4	20	11
i	0	1	**60**	0	28	0	13
ø	0	7	5	**21**	5	20	8
y	0	1	19	0	**45**	0	11
o	0	7	0	15	1	**60**	12
u	0	1	14	1	17	0	**45**

0	a	e	i	ø	y	o	u
a	**99**	0	0	0	0	0	0
e	0	**92**	56	0	0	0	0
i	0	8	**43**	0	0	0	0
ø	1	0	1	**71**	2	10	4
y	0	0	0	24	**98**	0	15
o	0	0	0	3	0	**85**	1
u	0	0	0	2	0	5	**80**

-6	a	e	i	ø	y	o	u
a	**97**	2	6	1	7	0	8
e	0	**32**	7	34	5	27	11
i	1	9	**39**	2	26	3	19
ø	0	24	11	**30**	12	31	11
y	0	11	19	3	**27**	2	24
o	2	13	4	25	0	**30**	7
u	0	9	14	5	23	7	**20**

-6	a	e	i	ø	y	o	u
a	**100**	0	0	0	0	0	0
e	0	**86**	64	0	0	0	0
i	0	14	**36**	0	1	0	0
ø	0	0	0	**64**	3	41	3
y	0	0	0	34	**74**	5	18
o	0	0	0	2	0	**50**	5
u	0	0	0	0	22	4	**74**

-18	a	e	i	ø	y	o	u
a	**25**	27	24	27	24	27	30
e	15	**10**	16	12	12	9	12
i	11	17	**18**	11	17	12	15
ø	24	19	20	**29**	21	26	22
y	7	11	11	10	**10**	12	8
o	6	7	3	5	8	**4**	4
u	12	9	8	6	8	10	**9**

-18	a	e	i	ø	y	o	u
a	**98**	1	0	0	0	0	0
e	0	**77**	67	0	0	0	1
i	0	22	**32**	0	0	0	0
ø	2	0	1	**52**	10	44	6
y	0	0	0	36	**53**	15	10
o	0	0	0	1	5	**25**	6
u	0	0	0	11	32	16	**77**

Tableau 2. – Matrices de confusion obtenues pour le test perceptif en *condition* non filtré (à gauche) et filtré LPC (à droite) et pour les RSB ∞, 0, -6 et -18 dB (case supérieure gauche).

d'enregistrement vidéo (lèvres maquillées pour des estimations très propres des paramètres labiaux). Dans le cadre d'une application plus réaliste, il faudrait lever chacune de ces restrictions. Passons rapidement sur celles qui concernent des problèmes généraux partagés avec le domaine de la reconnaissance audio-visuelle de parole : le passage au multilocuteur et à des images moins voire pas du tout contraintes. Il est clair que les outils de caractérisation de visages parlants sont nombreux et en plein développement [18]. Les progrès dans ce domaine peuvent être directement récupérables dans notre perspective, en remplaçant visage maquillé et paramètres $[A\,B\,S]$ par d'autres situations et d'autres descripteurs.

Il reste le problème-clé très complexe du passage des voyelles à la parole continue. Nous envisageons dans un premier temps le passage à des séquences voyelle-consonne-voyelle qui nous permettra de tenter de franchir la barrière du dynamique dans des conditions bien contrôlées. Cette perspective, qui représente notre préoccupation majeure dans cette voie originale que nous tentons d'ouvrir, pose deux problèmes principaux. Le premier concerne la visibilité des gestes consonantiques de la parole et le deuxième la caractérisation de leur conséquence acoustique pour la synthèse de filtres débruitants. En conséquence, deux voies simplifiées s'ouvrent à nous suivant la présence ou non d'une information visuelle significative :

– D'une part, nous avons l'intention d'exploiter un certain nombre de gestes visuellement pertinents, plus ou moins facilement repérables, et dont nous connaissons quelques propriétés du signal acoustique associé. Ainsi les bilabiales [p], [b], [m] sont les gestes consonantiques les plus visibles et doivent s'associer pendant la fermeture à un signal de silence ou de voisement. Les plosives dentales [t], [d], [n], sont plus discrètes visuellement mais nous

pouvons envisager une piste dans le repérage des dents et du bout de la langue qui impose une nouvelle procédure d'extraction visuelle.

– D'autre part, de nombreux gestes consonantiques sont invisibles et par conséquent visuellement ambigus soit entre eux soit par rapport aux combinaisons uniquement vocaliques. C'est le cas de [aga] vs [aea], ou pire pour un geste absolument non repérable tel que [ygy] vs [y#y]. Malgré cette difficulté majeure, des travaux de Öhman [19] à la théorie du « timing intrinsèque » de Fowler [7], une idée forte nous donne des raisons d'être relativement optimistes. L'idée est que la parole est produite comme une séquence de gestes vocaliques (gestes de voyelle à voyelle) lents, mettant en mouvement l'ensemble du conduit vocal, ligne de base sur laquelle sont superposés des gestes consonantiques rapides, ne mettant en jeu que certaines parties spécifiques et localisées du conduit vocal (la pointe de la langue, les lèvres, le dos de la langue, par exemple). Dans ce cadre, nous pouvons donc espérer que le filtrage de certaines séquences consonantiques sur la base des seuls gestes vocaliques fournisse des performances intéressantes. Aussi la démonstration que nous avons faite ici de l'efficacité de la méthode pour débruiter les voyelles nous semble particulièrement encourageante.

6. conclusion

Nous avons présenté dans cet article une étude de faisabilité concernant une nouvelle méthode de débruitage de parole par un filtrage utilisant l'image du locuteur. Les acquis de notre étude portent sur trois points essentiels :

1) Deux techniques de filtrage ont été étudiées et validées. Le filtre de Wiener permet de s'adapter au RSB mais reste complexe, dépendant de l'estimation de ce RSB et finalement perceptivement peu rentable. A l'inverse, le filtre LPC est plus simple mais plus efficace pour notre application et sa non-adaptation au RSB ne s'avère pas trop pénalisante dans le cas de fort RSB. Enfin, l'élaboration d'un filtre LPC hybride pondéré avec le filtre plat unitaire en fonction du RSB reste une piste intéressante pour l'avenir.

2) Le choix d'un associateur linéaire simple pour l'estimation des paramètres des filtres à partir des lèvres semble pertinent : cet outil s'est révélé remarquablement efficace. Certes l'utilisation d'outils plus puissants et plus complexes (réseaux de neurones par exemple) peut être envisagée pour améliorer les performances du système. Mais nous restons prudent sur ce point car ces outils ne peuvent résoudre directement le problème crucial de notre processus : la limite des informations disponibles sur les lèvres. De plus, à l'image du sélecteur antérieures/périphériques fondé sur le même type de méthode linéaire, notre idée est moins l'implantation en série de modules ultra-sophistiqués de type reconnaissance de parole, que l'élaboration d'un système de débruitage complet simple et fiable.

3) Les résultats qualitatifs et quantitatifs fournis par l'évaluation de ce système dans le cadre d'un corpus de voyelles stationnaires ont permis de valider notre démarche. Ils offrent une perspective très intéressante pour la poursuite de notre travail dans le cadre plus complexe de transitions voyelle-consonne-voyelle avec pour ambition à plus long terme le débruitage de parole continue.

Remerciements

Nous tenons à remercier particulièrement Jordi Robert-Ribes pour l'apport de ses travaux à notre étude et pour la gentillesse avec laquelle il a mis son corpus à notre disposition, ainsi que Tahar Lallouache, qui a rendu possibles les développements en « labiométrie » à l'ICP grâce à son poste visage-parole. Nous remercions par ailleurs les trois reviewers pour leurs critiques constructives.

BIBLIOGRAPHIE

[1] D. Baudois, C. Servière, & A. Silvent, « Algorithmes adaptatifs et soustraction de bruit », *Traitement du Signal*, vol. 6, No. 5, 1989, pp. 391–497.

[2] C. Benoît, T. Mohamadi, & S. Kandel, « Effects of phonetic context on audio - visual intelligibility of French », *J. 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] P. Ducknowski, U. Meier, & A. Waibel, « See me, hear me : integrating automatic speech recognition and lip reading », *Int. Conf. on Spoken Language Processing*, Yokohama, Japan, 1994, pp. 547–550.

[5] N.P. Erber, « Interaction of audition and vision in the recognition of oral speech stimuli », *J. of Speech and Hearing Research*, vol. 12, 1969, pp. 423–425.

[6] K.E. Finn, *An investigation of visible lip information to be used in automated speech recognition*, Doctoral dissertation, Georgetown University, Washington DC, 1986.

[7] C. Fowler, « Coarticulation and theories of extrinsic timing », *J. of Phonetics*, vol. 8, 1980, pp. 113–133.

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

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

[10] J.S. Lim, *Speech enhancement*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[11] H. McGurk, & J. MacDonald, « Hearing lips and seeing voices », *Nature*, vol. 264, 1976, pp. 746–748.

[12] A. McLeod, & Q. Summerfield, « Quantifying the contribution of vision to speech perception in noise », *British Journal of Audiology*, vol. 21, 1987, pp. 131–141.

[13] M.W. Mak, & W.G. Allen, « Lip-motion analysis for speech segmentation in noise », *Speech Communication*, vol. 14, 1994, pp. 279–296.

[14] J. Makhoul, « Linear prediction : a tutorial review », *Proc. IEEE*, vol. 63, No. 4, 1975, pp. 561–580.

[15] J.D. Markel, & A.H.Jr. Gray, *Linear prediction of speech*, Springer-Verlag, New York, 1976.

[16] K. Mase, & A. Pentland, « Automatic lipreading by optical-flow analysis », *Systems and Computers in Japan*, vol. 22, No. 6 , 1991, pp. 67–76.

[17] D.W. Massaro, « Testing between trace model and the fuzzy logical model of speech perception », *Cognitive Psychology*, vol. 21, 1989, pp. 398–421.

[18] NATO ASI Workshop, *Speechreading by man and machine : models, systems and applications*, D.G. Stork (Ed.), à paraître.

[19] S. Öhman, « Coarticulation in VCV utterance : spectrographic measurements », *J. Acoust. Soc. Am.*, vol. 39, 1966, pp. 151–168.

[20] A. Papoulis, *Signal analysis*, McGraw–Hill, New York, 1977.

[21] E.D. Petajan, *Automatic lipreading to enhance speech recognition*, Doctoral thesis, University of Illinois, 1984.
[22] L.R. Rabiner, & R.W. Schafer, *Digital processing of speech signals*, Prentice - Hall (Signal Processing Series), 1978.
[23] J. Robert-Ribes, *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles*, Thèse doctorale, INPG, Grenoble, 1995.
[24] M.R. Schroeder, B.S. Atal, & J.L. Hall, « Objective measure of certain speech signal degradations based on masking properties of human auditory perception », *Frontiers of Speech Communication Research*, B. Lindblom & S.Ohman (Eds.) Academic Press, London, 1979, pp. 217–229.
[25] D.G. Stork, G. Wolff, & E. Levine, « Neural network lipreading system for improved speech recognition », *Int. Joint Conf. on Neural Networks*, Baltimore, 1992, pp. 285–295.
[26] W.H. Sumby, & I. Pollack, « Visual contribution to speech intelligibility in noise », *J.Acoust. Soc. Am.*, vol. 26, 1954, pp. 212–215.
[27] Q. Summerfield, « Some preliminaries to a comprehensive account of audio visual speech perception », *Hearing by eye : the psychology of lipreading*, B.Dodd & R.Campbell (Eds.), Lawrence Erlbaum Associates, London, 1987, pp. 3–51.
[28] Q. Summerfield, A. McLeod, M. McGrath, & M. Brooke, « Lips, teeth and the benefits of lipreading », *Handbook of research on face processing*, A.W. Young & H.D. Ellis (Eds.), Elsevier Science Publishers B.V., North-Holland, 1989, pp. 223–233.
[29] P. Vary, « Noise suppression by spectral magnitude estimation-mechanism and theoretical limits », *Signal Processing*, vol. 8, No. 4, 1985, pp. 387–400.
[30] D.L. Wang, & J.S. Lim, « The unimportance of phase in speech enhancement », *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, No. 4, 1982, pp. 679–681.
[31] B.P. Yuhas, M.H. Goldstein, T.J. Sejnowski, & R.E. Jenkins, « Neural network models of sensory integration for improved vowel recognition », *Proc. IEEE*, vol. 78, No. 10, 1990, pp. 1658–1668.

Manuscrit reçu le 6 Juillet 1995.

LES AUTEURS

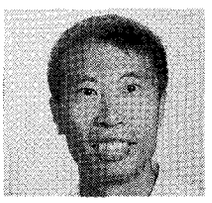
Laurent GIRIN

Laurent Girin est diplômé de l'ENS d'Ingénieurs Electriciens de Grenoble et de l'ENS d'Electronique et de Radioélectricité de Grenoble. Titulaire du DEA Signal - Image - Parole de l'INP de Grenoble en 1994, il poursuit depuis une thèse à L'Institut de la Communication Parlée sous la direction de Gang Feng et Jean-Luc Schwartz. Ses travaux portent sur l'utilisation des informations visuelles pour le rehaussement de la parole bruitée.

Jean-Luc SCHWARTZ

Ancien élève de l'Ecole Normale Supérieure de Paris, Jean- Luc Schwartz a obtenu un doctorat de 3ème cycle à l'INP de Grenoble en 1981 et un doctorat d'Etat en 1987. Depuis 1983 il est chargé de recherches au CNRS à l'Institut de la Communication Parlée, et dirige depuis 1988 l'équipe « Perception auditive et visuelle de la parole ». Ses domaines d'intérêt recouvrent la modélisation de l'audition, la psychoacoustique, la perception audiovisuelle de la parole, les interfaces pour la reconnaissance automatique de la parole et les relations entre perception, motricité, cognition et langage.

Gang FENG



Docteur de l'INP de Grenoble en 1986, Gang Feng est maître de conférence à l'ENS d'Electronique et de Radioélectricité de Grenoble. Il est responsable de l'équipe « Traitement et Codage de la parole » à l'Institut de la Communication Parlée. Ses principaux domaines d'intérêt sont la compression, le codage, la modélisation, la caractérisation, l'inversion et le débruitage de la parole.