Débruitage de parole par fusion des informations auditives et visuelles : une étude des transitions vocaliques

Laurent Girin, Gang Feng et Jean-Luc Schwartz

Institut de la Communication Parlée, UPRESA 5009
INPG/ENSERG/Université Stendhal
B.P. 25, 38040 GRENOBLE CEDEX 09, FRANCE
girin@icp.grenet.fr

RÉSUMÉ

Dans cet article, nous présentons un système de débruitage de parole basé sur une fusion des informations auditives et visuelles. Dans un premier temps, la structure globale du système est présentée. Puis l'outil utilisé pour mélanger les deux sources d'information est décrit. Le système complet est implémenté dans le contexte de transitions vocaliques dégradées par un bruit blanc additif. Une évaluation complète est réalisée dans ce contexte : elle comporte des mesures de distance, des scores de classification gaussienne, et un test perceptif. Les résultats sont très prometteurs.

ABSTRACT

This paper deals with a noisy speech enhancement technique based on the fusion of auditory and visual information. We first present the global structure of the system, and then we focus on the tool we used to melt both sources of information. The whole noise reduction system is implemented in the context of vowel transitions corrupted with an additive white noise. A complete evaluation of the system in this context is presented, including distance measures, gaussian classification scores, and a perceptive test. The results are very promising.

1 Introduction

La parole est à la fois auditive et visuelle, et il existe une réelle complémentarité entre les deux modalités [2]. Ainsi, les informations visuelles peuvent partiellement compenser la déficience des informations auditives [2][3]. Cette propriété est déjà exploitée par les systèmes de reconnaissance bimodaux : leurs performances sont accrues par l'utilisation de la modalité visuelle, en particulier en environnement bruité [4].

Dans un article précédent [1], nous avons présenté un système tout à fait original, dédié aux télécommunications et au dialogue homme-machine. Sa fonction est de débruiter des signaux de parole en utilisant l'image du locuteur. Le principe est d'estimer, à partir de caractéristiques des lèvres du locuteur, un modèle du signal audio, et de filtrer le signal audio bruité à travers ce modèle. Les résultats obtenus sur des voyelles stationnaires ont été encourageants. Mais le point faible de cette structure réside dans la tentative d'estimer une information audio complète à partir du canal vidéo, alors que l'information labiale reste très partielle.

Dans cet article, nous proposons une nouvelle structure pour notre système : dans l'optique de mieux exploiter la complémentarité audiovisuelle, l'information auditive débruitée est maintenant estimée à partir des deux canaux, l'audio bruité et le vidéo. En premier lieu, nous présentons la structure globale du nouveau système. Puis nous décrivons le processus d'intégration bimodale. Enfin, nous présentons quelques résultats obtenus sur des transitions vocaliques dégradées par un bruit blanc additif.

2 Structure du système

Le nouveau système utilise abondamment le modèle de prédiction linéaire [5]. Considérons la figure 1. Tout d'abord, le signal audio bruité est analysé par LPC. Nous obtenons des paramètres spectraux, et l'excitation bruitée est extraite par filtrage à travers le filtre inverse $A_n(z)$. Puis, les paramètres spectraux bruités sont combinés avec les paramètres vidéo selon le processus décrit en section 3 pour donner des paramètres audio débruités. Enfin, les signaux de parole sont resynthétisés en filtrant l'excitation à travers le filtre LPC $1/A_e(z)$ dérivé des paramètres spectraux débruités. Le traitement de parole continue implique la réitération trame par trame du processus selon une dynamique adaptée à la parole et au traitement des images du locuteur.

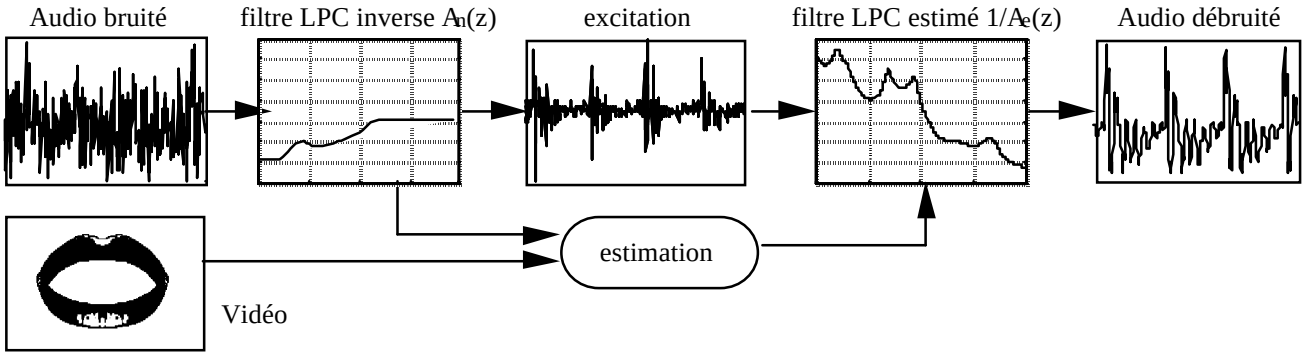

Figure 1 – Structure du système de débruitage

3 Estimation des paramètres audio débruités

Le problème central est donc d'estimer des paramètres audio débruités à partir des mêmes paramètres bruités et des paramètres vidéo. Nous avons utilisé une méthode de régression linéaire du fait de sa simplicité et de son efficacité pour traiter notre problème [1]. Le principe est le suivant. Considérons un vecteur audiovisuel comme étant la concaténation de paramètres audio et vidéo. La méthode implique une phase d'apprentissage où deux matrices sont construites. La première, notée M_{AV}, regroupe les vecteurs audiovisuels d'un corpus d'apprentissage en condition audio bruitée. La seconde matrice, notée M_A, regroupe les vecteurs audio correspondants, issus du même corpus mais en condition non bruitée. La matrice M réalisant la régression linéaire entre M_{AV} et M_A est alors calculée. A présent, pour chaque nouveau vecteur audiovisuel bruité V_{AV}, le produit entre V_{AV} et M fournit une estimation du vecteur audio débruité V_A.

4 Expérimentation

4.1 Les données audio et vidéo

Le poste « visage-parole » de l'ICP [6] permet l'extraction automatique de trois paramètres labiaux fondamentaux, l'étirement (A), la hauteur (B) et l'aire (S) intérolabiaux. Ces paramètres sont extraits toutes les 20 ms.

L'extraction et la transformation de l'information audio passe par la description des polynômes LPC mis en jeu. Les meilleures performances du système ont été obtenues avec une représentation sur 50 canaux spectraux qui sont les valeurs logarithmiques du module du filtre 1/A(z) pour 50 points également répartis sur la moitié supérieure du cercle unité. L'ordre de la LPC est de 20, les signaux sont échantillonnés à 16 KHz et les spectres LPC sont calculés sur 512 échantillons (32 ms, ce qui implique un recouvrement des fenêtres d'analyse de 12 ms pour se synchroniser avec la période vidéo de 20 ms).

4.2 Le corpus

Notre premier travail décrit en [1] a fourni des résultats prometteurs sur des voyelles stationnaires. Dans cette nouvelle implantation, nous nous intéressons au cas de transitions vocaliques $V_1V_2V_1$ en mode monolocuteur, V_1 et V_2 étant parmi [a, i, y, u]. Un échantillon des 16 stimuli possibles est utilisé à l'apprentissage, et un autre lors des tests. Avec une période d'acquisition de 20 ms, environ 350 vecteurs audiovisuels sont extraits pour la série de 16 stimuli (environ 24 trames par stimulus).

4.3 Protocole expérimental

Seul le cas d'un bruit blanc additif est abordé. Les résultats présentés ici ont été obtenus en utilisant deux matrices associatrices M : une est utilisée pour le débruitage des signaux avec un rapport signal à bruit (RSB) « fort ». Elle est calculée avec des trames de stimuli présentées à •, 18, 12, 6 et 0 dB de RSB. L'autre est utilisée pour le débruitage des signaux avec un RSB « faible ». Elle est calculée avec des trames de stimuli présentées à 6, 0, -6, -12, -18 dB de RSB. Nous avons vérifié qu'une analyse discriminante linéaire à deux classes permet, lors du traitement trame par trame, de classer les trames de stimuli de RSB inférieur à 0 dB ou supérieur à 6 dB avec moins de 1% d'erreur. Ceci assure le bon choix de l'associateur « fort ou faible », sachant que les deux donnent des réponses très similaires entre 6 et 0 dB du fait du recouvrement des RSB à l'apprentissage.

4.4 Filtrage

Pour obtenir le filtre $1/A_e(z)$, une FFT inverse est appliquée aux coefficients spectraux débruités (plus précisément sur leur module carré en mode linéaire, de manière à obtenir l'estimation d'une densité spectrale de puissance). Les coefficients d'autocorrélation résultant sont ensuite soumis à l'algorithme de Levinson (toujours à l'ordre 20). Durant le traitement, une mémorisation des buffers des filtres ainsi qu'un fenêtrage trapézoïdal sont appliqués aux trames pour assurer la continuité du signal aux jonctions.

5 Résultats

Après une évaluation qualitative informelle du système, trois procédures d'évaluation quantitatives sont présentées : des mesures de distance, un test de classification gaussienne, et un test perceptif. Ces évaluations ont été réalisées pour les huit RSB standards (•, 18, 12, 6, 0, -6, -12, -18 dB).

5.1 Évaluation qualitative

Des tests d'écoute informels ont montré le bon comportement global de notre système. Pour des bruits faibles, le débruitage dégrade très peu la qualité des stimuli. Pour des bruits moyens, les effets du débruitage sont très bénéfiques : le message devient plus intelligible même si la qualité est parfois modifiée (le filtrage de l'excitation bruitée conduit à une voyelle en partie chuchotée). Pour des niveaux de bruit plus forts (perte d'intelligibilité), le système permet de retrouver l'intelligibilité de la plupart des stimuli (presque tous les [a] ou [i], du fait de leur distinction labiale, avec plus d'ambiguïté entre [u] et [y], qui restent toutefois départagés grâce à l'audio jusqu'à 0 dB).

5.2 Mesures de distance

La distance d'Itakura [5] a été utilisée pour mesurer la différence entre les spectres propres et débruités. La figure 2 montre cette distance moyennée sur le corpus de test complet (16 stimuli) pour trois conditions. AV indique l'utilisation en entrée des associateurs des informations auditives et visuelles, A indique la restriction à l'audio seul (vecteurs audio au lieu d'audio-vidéo), et V la restriction au vidéo seul (vecteurs vidéo au lieu d'audio-vidéo).

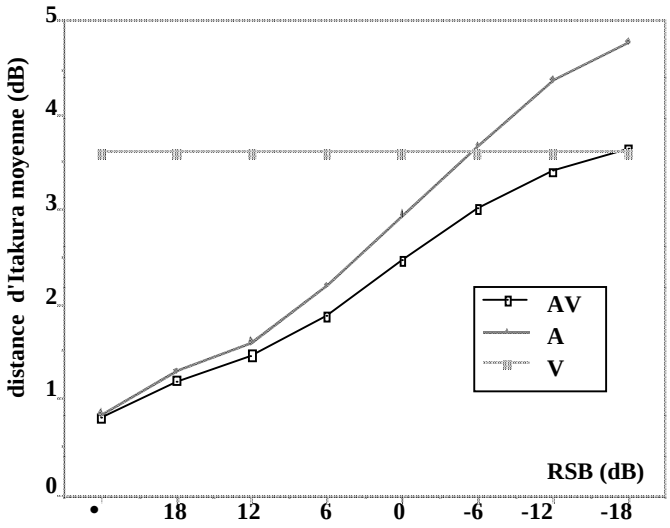

Figure 2 – Distance d'Itakura moyenne entre les spectres propres et débruités du corpus de test

Il y a effectivement débruitage, au sens où toutes les distances obtenues sont faibles par rapport aux distances entre les spectres bruités et les spectres propres. De plus, la condition AV est toujours meilleure que A, et presque toujours meilleure que V (jusqu'à près de -18 dB). Ceci confirme l'apport des informations visuelles, et la bonne complémentarité entre les deux modalités.

5.3 Test de classification gaussienne

Afin d'évaluer le système dans une tâche de reconnaissance, un test de classification gaussienne a été réalisé sur les 4 voyelles du corpus. L'échantillon utilisé dans ce test est constitué des deux trames les plus proches des noyaux vocaliques de chaque stimulus. Cette sélection assure l'absence d'effets de coarticulation (stabilité relative des signaux audio et vidéo dans ces zones) et un étiquettage sûr des voyelles. Pour chaque niveau de bruit, 96 échantillons sont donc disponibles (2 trames, 3 voyelles par stimulus, 16 stimuli), soit 24 par voyelle. Ce nombre étant petit par rapport au nombre de paramètres audio, ce dernier est réduit de 50 à 5 par une analyse en composantes principales (ACP). Pour les résultats donnés en figure 3, les paramètres de l'ACP et du classifieur gaussien sont déterminés en présentant les données à 3 niveaux de bruit (•, 18, et 12 dB). La figure 3 montre les scores de classification correcte obtenus pour 3 conditions : A signifie un apprentissage du classifieur avec les paramètres audio bruités seulement (5 paramètres). Dans cette condition, nous avons la comparaison entre $A_{bruité}$, qui indique l'utilisation du corpus de test audio bruité, et $A_{débruité}$ qui indique l'utilisation du même corpus après débruitage par notre système. En comparaison, les scores notés AV correspondent à un classifieur audiovisuel appliqué sur des vecteurs combinant les 5 paramètres audio et les 3 paramètres vidéo. tous les scores sont normalisés entre 0 (choix aléatoire) et 100% (reconnaissance parfaite).

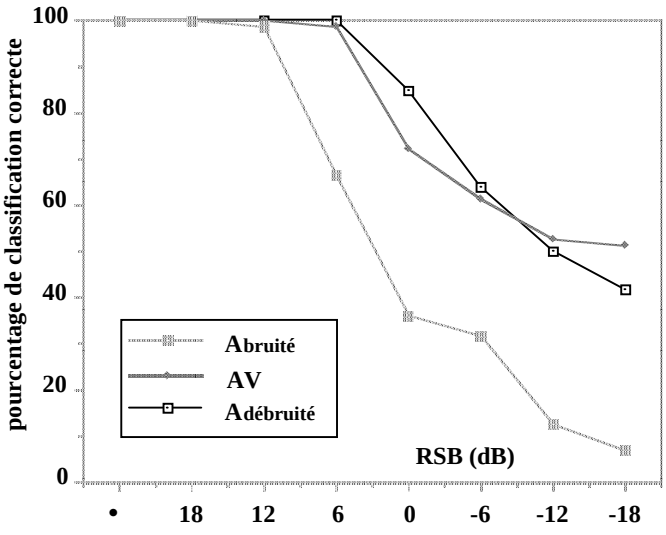

Figure 3 – Scores du test de classification gaussienne

La différence entre les scores des conditions bruité et débruité confirme l'efficacité du système. De plus, la condition audio débruité est compétitive par rapport à la classification audiovisuelle jusqu'à un RSB de l'ordre de -12 dB.

5.4 Test perceptif

L'évaluation finale du système passe par la mise en œuvre d'un test perceptif. 17 sujets devaient identifier les stimuli présentés aléatoirement en condition bruitée et débruitée pour les 8 RSB standards. Les transitions $V_1V_2V_1$ ont été segmentées à la main en V_1V_2 et V_2V_1 de manière à ne présenter V_1 qu'une seule fois par échantillon. Ainsi, chaque point des courbes d'identification de la figure 4, c'est-à-dire chaque niveau de bruit et chaque condition (bruité et débruité), correspond à 1088 réponses (16 stimuli à 2 voyelles, 2 segments V_1V_2 et V_2V_1, 17 sujets).
La figure 4 montre que le débruitage est efficace dès 0 dB de RSB. Les gains obtenus (différence entre les conditions débruité et bruité) sont d'environ 6% à 6 dB, 17,5% à 0 dB, 18,5% à -6 dB, 30% à -12 dB, et atteignent 42,5% à -18 dB.

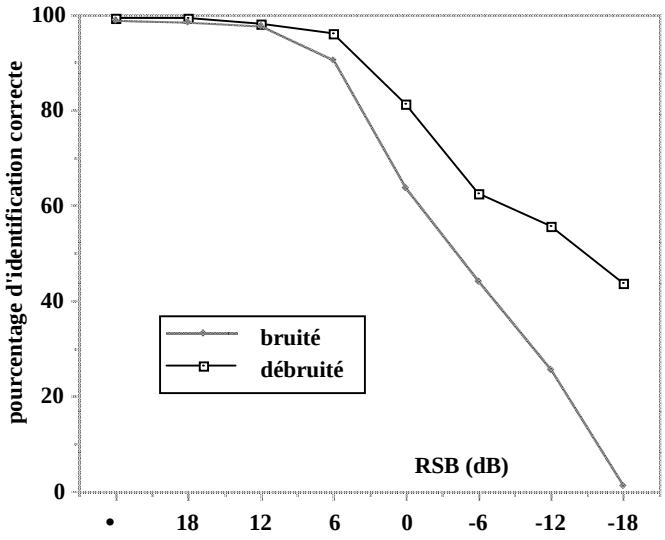

Figure 4 – Scores d'identification du test perceptif

Les matrices de confusion pour les 5 plus faibles RSB sont présentées dans le tableau 1 (les matrices pour les 3 RSB forts sont presque diagonales et présentent peu d'intérêt). Elles permettent de mettre à jour les principales performances de notre système :

1) La désambiguisation du contrast [i, y], qui est fortement dégradé par le bruit avant débruitage. Ce cas est un bon exemple de la complémentarité audiovisuelle de la parole (distinction visuelle robuste pour une robustesse auditive faible dans le bruit).

2) La relative désambiguisation de la confusion [a, i]. Celle-ci apparaît seulement pour les bruits forts, du fait de la robustesse du son [a] dans le bruit, et de la moins bonne distinction labiale des deux voyelles en dynamique en comparaison avec les mêmes voyelles stationnaires [1]. Cette effet est plus important de [a] vers [i] que de [i] vers [a].

3) Le renforcement du trait d'arrondissement ([y] et [u] sont très contrastés avec [a] et [i]). Malheureusement, cet effet entraîne la confusion de [u] avec [y] pour les forts niveaux de bruit. Dans ce cas, la faible récupération d'informations auditives reste un point faible du processus de fusion.

		signaux bruités				signaux débruités			
RSB	V	a	i	y	u	a	i	y	u
6 dB	a	**271**	1	0	0	**272**	0	0	1
	i	1	**223**	6	2	0	**270**	2	7
	y	0	37	**260**	16	0	2	**260**	11
	u	0	11	6	**254**	0	0	10	**253**
0 dB	a	**271**	2	1	17	**272**	3	0	0
	i	0	**131**	35	17	0	**253**	7	15
	y	1	104	**187**	45	0	14	**222**	70
	u	0	35	49	**203**	0	2	43	**187**
-6 dB	a	**272**	12	17	15	**268**	7	0	1
	i	0	**96**	60	60	4	**243**	28	22
	y	0	128	**155**	89	0	21	**222**	201
	u	0	36	40	**108**	0	1	22	**48**
-12 dB	a	**234**	61	45	42	**247**	6	1	3
	i	27	**108**	94	107	19	**228**	18	25
	y	8	69	**89**	73	6	36	**221**	213
	u	3	34	44	**50**	0	2	32	**31**
-18 dB	a	**128**	129	121	105	**168**	26	4	2
	i	109	**95**	108	113	77	**211**	23	15
	y	24	37	**29**	44	20	33	**218**	225
	u	11	11	14	**10**	7	2	27	**30**

Tableau 1 – Matrices de confusion pour le test perceptif.

6 Conclusion

Nous avons présenté dans cet article une méthode originale de débruitage de parole utilisant à la fois l'information auditive bruitée et des paramètres décrivant la forme des lèvres du locuteur. Son implantation dans le cadre de transitions vocaliques dégradées par un bruit blanc additif a montré qu'un débruitage efficace peut être obtenu grâce à la complémentarité entre les modalités auditive et visuelle. Ces résultats sont très prometteurs pour la future étape de nos travaux qui concernera le traitement de transitions voyelle-consonne-voyelle.

Références

[1] Girin L., Feng G., & Schwartz J.-L., Débruitage de parole par un filtrage utilisant l'image du locuteur : une étude de faisabilité, *Traitement du Signal*, vol. 13, n°4, 1996, pp. 319-334.

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

[3] Sumby W.H., & Pollack I., Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Am.*, 26, 1954, pp. 212-215.

[4] Stork D., & Hennecke M., (Eds.), *Speechreading by humans and machines*, Springer-Verlag, Berlin, 1996.

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

[6] Lallouache M.T., Un poste « visage-parole », *18th JEPs*, Montréal (Québec), Canada, 1990, pp. 282-286.