Reconnaissance audiovisuelle de la parole par VMike

Fabian Brugger [1]*, Leila Zouari* [1]*, Hervé Bredin* [1]*, Asmaa Amehraye* [2]*,*
Gérard Chollet [1]*, Dominique Pastor* [2] *et Yang Ni* [3]

[1] GET-ENST/CNRS-LTCI - LastName@tsi.enst.fr
[2] GET-ENST Bretagne - FirstName.LastName@enst-bretagne.fr
[3] GET-INT - FirstName.LastName@int-evry.fr

ABSTRACT

This article presents a new Electronic Retina based Smart Microphone (VMike) and investigates the use of its novel parameters - lip profiles - in audiovisual speech recognition. In order to evaluate the parameterization, both an audio only and a video only speech recognition system are developed and tested. Then, two main fusion techniques are employed to test the usability of profiles in audiovisual systems : feature fusion and decision fusion. These results are compared to the performance of recognizers based on a state-of-the-art parameterization, and also to results obtained by applying perceptual filtering to the speech signal prior to recognition. When feature fusion is applied, and under noisy conditions, recognition using lip profiles improved by up to 13 percent with respect to audio-only recognition.

1. INTRODUCTION

Les systèmes actuels de reconnaissance de la parole sur petit vocabulaire et dans un environnement non bruité donnent des résultats satisfaisants (taux d'erreur inférieur à 1 %) [7]. Cependant, dans les applications réelles, les conditions d'enregistrement (voiture, avion, hélicoptère, etc.), d'acquisition et de transmission ne sont pas idéales : un bruit est nécessairement introduit dans la chaîne de traitement de la parole, entraînant une baisse des performances du système de reconnaissance.
Plusieurs méthodes ont été développées dans le but d'améliorer la reconnaissance de la parole en milieu bruité. On distingue trois principales catégories : le débruitage du signal, l'adaptation du système à l'environnement ou encore la reconnaissance audiovisuelle en considérant le mouvement des lèvres. Cette dernière méthode repose sur l'idée que la parole est un moyen audiovisuel de communication. En effet, le message parlé est plus intelligible quand il est accompagné de la vision du visage du locuteur, et particulièrement quand le milieu de transmission est bruité.
Un système de reconnaissance audiovisuelle de la parole résulte de la fusion de deux systèmes mono-modaux audio et vidéo. Classiquement, on distingue deux types de fusion : la fusion des paramètres et la fusion des scores.
Dans le cadre du projet VMike, un dispositif éponyme a été développé. Il s'agit d'un microphone augmenté d'une rétine électronique produisant un signal de parole audiovisuelle. Le but de cet article est de montrer que l'utilisation de cette nouvelle paramétrisation de la zone de la bouche -issue de VMike- peut être utilisée pour améliorer les taux de reconnaissance de la parole en milieu bruité.

Ce document est organisé comme suit. La section 2 présente rapidement le dispositif VMike ainsi que les paramètres visuels qu'il produit. Très peu de données ont été acquises avec le dispositif lui-même : nous les avons donc simulé à partir de la base de données audiovisuelles BANCA. Les paramètres utilisés sont décrits dans la section 3. Un tour d'horizon des techniques de fusion expérimentées constitue la section 4. Les expériences réalisées sont présentées en détails dans la section 5, où nous comparons notre système à un système audiovisuel état-de-l'art basé sur une transformation DCT de la zone des lèvres et à un système *audio seul* basé sur un filtrage perceptuel. Enfin, la section 6 conclut cet article et propose différentes perspectives et pistes d'amélioration.

2. VMIKE

2.1. Description du dispositif

VMike est un microphone augmenté d'une rétine électronique. Il a été développé dans le cadre du projet VMike[1] afin d'explorer l'apport des lèvres en reconnaissance de la parole. Les spécifications de ce dispositif (schématisé dans la figure 1) sont :

– un canal *voix* assuré par un microphone classique,
– un canal *vision* assuré par une rétine électronique.

Cette rétine permet la compression de l'information visuelle par projection sur les axes horizontal et vertical. L'interface stéréo entraîne une synchronisation parfaite et naturelle entre les deux signaux audio et vidéo. Les avantages d'une telle configuration sont :

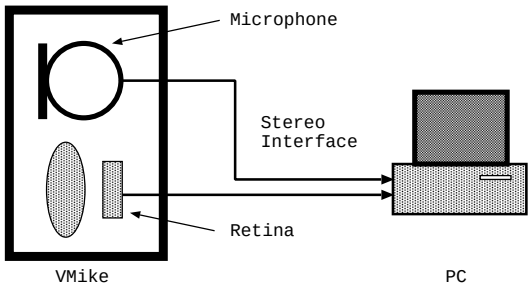

FIG. 1: le dispositif VMike

[1]Cette étude est soutenue par le groupe des Ecoles de Télécommunications (GET)

– sa simplicité,
– le nombre réduit de paramètres qu'elle produit,
– son importante vitesse de transmission.

2.2. *Sorties du VMike*

Deux signaux mono sont générés par le dispositif VMike :
– le signal audio de parole sur le canal gauche,
– les projections horizontales et verticales de la zone des lèvres sur le canal droit.

Alors que la parole est acquise de façon classique par le microphone, les images subissent un prétraitement avant leur transmission. En effet, pour chaque image, les projections sur les axes horizontal et vertical (voir figure 2) sont calculées (et concaténés), modulées puis transmises sur le canal droit.

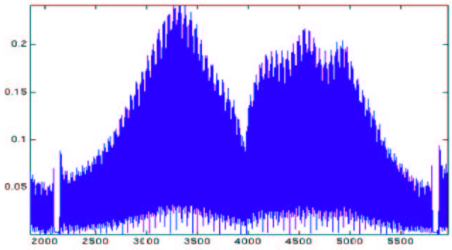

FIG. 2: Exemple d'un signal vidéo démodulé de VMike

3. PARAMÈTRES VISUELS

3.1. *Simulation du VMike*

Au moment de la réalisation des expériences dont les résultats sont présentés dans cet article, peu de données ont été enregistrées avec le dispositif VMike. Aussi, nous avons simulé des enregistrements à partir de la base de données audiovisuelles BANCA [3].
Rappelons que le VMike est utilisé comme un microphone classique, c'est-à-dire tenu à quelques centimètres de la bouche et orienté vers celle-ci. Aussi avons-nous implémenté un algorithme (présenté en détails dans [4]) permettant la localisation automatique de la zone de la bouche dans les séquences vidéo de la base de données BANCA. Un exemple de cette zone est présenté dans la figure 5.

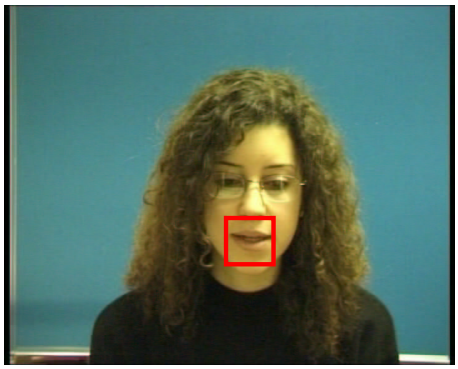

FIG. 3: Détection de la zone d'intérêt

3.2. *Projections XY*

Une fois la zone de la bouche localisée, elle est normalisée en taille (200 par 200) et les projections sur l'axe horizontal (respectivement vertical) sont calculées très simplement comme la somme des niveaux de gris sur chaque colonne (respectivement chaque ligne). La figure 4 illustre cette transformation.

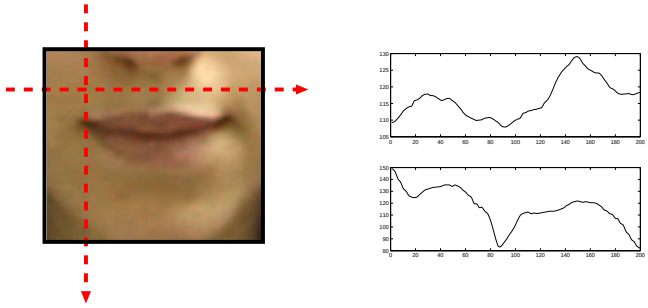

FIG. 4: Les projections de Vmike

3.3. *DCT*

Afin de comparer l'apport de cette nouvelle approche, une paramétrisation état-de-l'art [8] a aussi été implémentée. La zone de la bouche est normalisée en taille ($H = 64$ pixels par $W = 64$ pixels). Notant $I(i,j)$ l'intensité des pixels dans cette zone, une DCT (Discret Cosine Transform) est appliquée sur la zone d'intérêt :

$$X(u,v) = C(u,v) \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \alpha(i,u,H)\, \alpha(j,v,W)\, I(i,j)$$

où $C(u,v)$ est un coefficient de normalisation et $\alpha(i,u,H) = \cos\left(\frac{2i+1}{2H} u\pi\right)$. Afin de réduire la dimension de ces paramètres (initialement $64 \times 64 = 4096$), les indices des coefficients les plus énergétiques sont obtenus sur un ensemble d'apprentissage. Seuls les coefficients correspondants sont conservés : dans notre cas, nous avons conservé les 100 coefficients les plus énergétiques. Ce principe est utilisé classiquement dans le cadre de la compression d'images : ils correspondent généralement aux fréquences u et v les plus basses.

3.4. *Traitement des paramètres bruts*

Plusieurs transformations sont alors appliquées à ces paramètres dans le but de réduire leur dimension, augmenter leur pouvoir discriminant et tenir compte de leur caractère dynamique.
Ainsi, une normalisation de la moyenne sur les paramètres bruts est suivie d'une transformation en analyse discriminante linéaire (LDA) pour réduire la dimension à 40 : ce sont les paramètres DCT et Pro (projections) dans la suite de cet article.
La dynamique de ces paramètres est modélisée en concaténant, à chaque instant d'échantillonnage, 15 échantillons consécutifs centrés sur l'échantillon courant. Une LDA est alors appliquée sur ces paramètres de dimension $40 \times 15 = 600$ pour obtenir des paramètres discriminants contenant l'information dynamique de dimension 40 : ce sont les paramètres DCT2 et Pro2.

4. TECHNIQUES DE FUSION

L'objectif d'un système de reconnaissance audiovisuelle est de combiner au mieux les performances de deux systèmes audio et vidéo afin d'améliorer les performances de reconnaissance de la parole, en particulier en présence de bruit. Classiquement, on distingue deux types de fusion : la fusion des paramètres et la fusion des scores.

4.1. Fusion des paramètres

Cette fusion est réalisée au moment de la paramétrisation des signaux audio et vidéo. Une fois les paramètres de chaque modalité sont extraits, les vecteurs audio $O_{a,t}$ et vidéo $O_{v,t}$, de dimension d_a et d_v respectivement, sont concaténés à chaque instant t pour ne former qu'un seul vecteur de paramètres audiovisuels $O_{av,t} = [O_{a,t}, O_{v,t}]$ de dimension $d_a + d_v$.
Dans les étapes suivantes de la chaîne de reconnaissance de la parole (estimation des paramètres, décodage, évaluation), aucune modification n'est nécessaire.

4.2. Fusion des scores

La fusion de scores ou de décision est possible lorsque l'on dispose de systèmes séparés (ici, audio et vidéo) et que leur fusion est réalisée au moment de la décision, par combinaison de leurs scores respectifs. Des poids différents peuvent être affectés à chaque système (ou parties de ces derniers) afin de privilégier l'une ou l'autre des deux modalités. Dans le cas de système de reconnaissance où les unités sub-lexicales (de type *phone*, par exemple) sont modélisées par des modèles de Markov cachés, cette fusion peut avoir lieu à différents niveaux qui sont l'état ou le phone ou le mot ou encore la phrase. Lorsque la fusion est éffectuée à chaque état, elle est dite synchrone, sinon elle est asynchrone.
Une fusion audiovisuelle synchrone est réalisée comme suit. Soient deux systèmes de reconnaissance de la parole audio et vidéo dont les modèles acoustiques ont la même topologie. Si $P(O_a; t, s)$ et $P(O_v; t, s)$ représentent les vraisemblances respectives d'une observation O émise à l'instant t par le même état s audio et vidéo respectivement, alors son score audiovisuel peut s'exprimer par :

$$P(O_{av}; t, s) = \lambda P(O_a; t, s) \times (1 - \lambda) P(O_v; t, s)$$

Le poids λ permet de donner plus d'importance à une modalité ou à l'autre. Pour chaque système, λ peut être choisi constant ou variable. Généralement, il dépend du rapport signal à bruit.

5. EXPÉRIENCES ET RÉSULTATS

Deux types d'expérience ont été réalisées : reconnaissance audiovisuelle de la parole et débruitage par filtrage perceptuel en amont de la reconnaissance de la parole. Dans le premier cas, les paramètres vidéo de type DCT et projections sont testés. Le même protocole expérimental est adopté dans les deux cas.

5.1. Protocole expérimental

Les expériences de reconnaissance de chiffres décrites ci-dessous ont été réalisées sur les données en condition *studio* de la base de données audiovisuelle BANCA (partie *controlled*). 52 locuteurs (26 femmes et 26 hommes) ont enregistré 4 sessions (S1 à S4) de 2 phrases durant lesquelles il/elle prononce une suite aléatoire de 12 chiffres (parmi les chiffres 1 à 9). Les 6 phrases des trois premières sessions sont utilisées lors de l'apprentissage et les 2 phrases de la quatrième session pour effectuer les tests. Par conséquent, $52 \times 6 = 312$ phrases de 12 chiffres (environ 38 minutes) consituent l'ensemble d'apprentissage et $52 \times 2 = 104$ tests ont été réalisés.
Des modèles de Markov cachés (3 états chacun et 16 gaussiennes par état) indépendants du contexte sont construits et estimés par l'algorithme de Baum-Welch. Le signal de parole est paramétrisé par des vecteurs de 12 coefficients MFCC et de leurs première et deuxième dérivées. Pour la vidéo, selon les expériences, les coefficients DCT ou les projections sont utilisés. Le bruit testé est de type *babble*. Il est extrait de la base de données NoiseX [1]. Enfin, le décodage est réalisé par le décodeur HTK [9], en utilisant un algorithme de Viterbi.

5.2. Reconnaissance audiovisuelle

Comme mentionné auparavant, l'objectif de ce travail est d'expérimenter l'apport des projections des lèvres en reconnaissance de la parole bruitée. Pour ce faire, la base de données BANCA est bruitée par un bruit de type *babble*. Puis deux systèmes audio et vidéo sont construits et leurs résultats sont évalués. Suivant le type de paramètres, on obtient une précision de 37.02% pour le système vidéo utilisant les projections et 46.69% avec les DCT. Enfin, une technique de fusion (des paramètres ou des scores) permet de générer le système audiovisuel. Les résultats de reconnaissance audiovisuelle sans bruit et à -5dB sont reportés sur le tableau 1 :

TAB. 1: Résultats des systèmes de reconnaissance audio, vidéo et et audiovisuels

	Monomodal		Fusion param.		Fusion scores	
	studio	-5dB	studio	-5dB	studio	-5dB
Audio	97.55	45.59	-	-	-	-
Pro	35.49	35.49	96.71	49.93	97.55	46.01
Pro2	37.02	37.02	95.31	51.68	97.55	46.71
DCT	46.69	46.69	93.98	52.52	97.55	50.91
DCT2	44.40	44.40	95.45	54.22	97.55	51.15

On peut remarquer que :

- Dans les deux cas de fusion, la reconnaissance audiovisuelle apporte une amélioration par rapport au système audio, et ce, pour tous les types de paramétrisation (Pro, Pro2, DCT, et DCT2).
- Les résulats de la fusion des paramètres sont légèrement supérieurs à ceux de la fusion synchrone des scores.
- Les meilleurs résultats de fusion des paramètres sont obtenus avec les paramètres dynamiques (Pro2 et DCT2) qui correspondent à une augmentation relative de la précision de 13% et de 19%.

5.3. Débruitage de la parole

Nous avons étudié un filtrage perceptuel [5] [2] conçu de manière à respecter le phénomène de masquage simultané. Une modélisation de ce dernier permet de calculer pour chaque trame du signal de parole une courbe de masquage représentant les points de pression acoustiques nécessaires

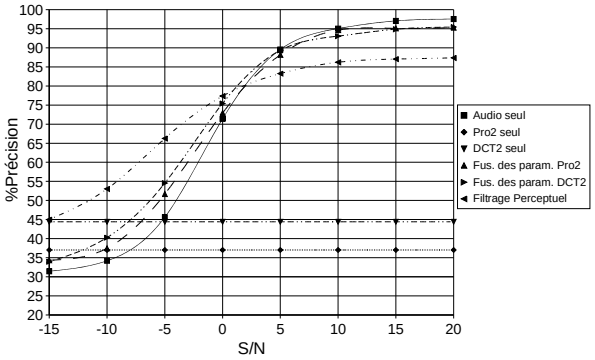

FIG. 5: Résultats de la fusion des paramètres

pour qu'un son devienne audible en présence d'un masquant [6]. Le but est de masquer les composantes audibles du bruit résiduel et de diminuer les distorsions du signal. Nous avons alors comparé les taux de reconnaissance de la méthode audiovisuelle à ceux obtenus par le filtrage perceptuel [5] dans les conditions idéales où la courbe de masquage est calculée à partir de la version non bruitée du signal de parole et la densité spectrale de puissance du bruit est estimée à partir d'une référence de bruit seul. Les résultats sont reportés sur la figure 5.

On peut voir que le filtre perceptuel améliore les performances du système bruité. Ces performances dépassent celle du système de reconnaissance audiovisuelle pour des rapports signal sur bruit inférieurs à 0 dB.

6. CONCLUSION ET PERSPECTIVES

Une nouvelle rétine électronique augmentée d'un microphone a été développée. Elle permet d'acquérir un signal audiovisuel de parole où audio et vidéo sont naturellement synchronisés. Le signal vidéo est constitué des projections horizontales et verticales de l'image de la zone des lèvres de l'utilisateur. L'apport de cette paramétrisation originale a été expériementé dans le cadre d'un système de reconnaissance audiovisuelle de la parole en milieu bruité sur la base de données audiovisuelles BANCA. Deux techniques de fusion audiovisuelle ont été envisagées (au niveau des paramètres et des scores) et comparées à un système audiovisuel état-de-l'art et un système audio avec débruitage par filtre perceptuel.
Bien que moins performantes que l'état-de-l'art basé sur la transformation DCT de la zone des lèvres, les projections apportent une amélioration des performances en milieu bruité comparé à un système audio seul (+13% de précision relative). La fusion niveau des paramètres donne de meilleurs résultats que celle au niveau des scores. Cependant, cette dernière pourrait être améliorée en appliquant une fusion asynchrone.
Aussi, prévoyons-nous d'appliquer des transformations de type PCA ou LDA dans l'espace des paramètres audiovisuels. En outre, il est prévu de combiner les deux approches (débruitage et fusion audiovisuelle) afin d'obtenir un système profitant de ces deux sources d'amélioration. Enfin, à plus long terme, il serait intéressant de tester l'utilisation du VMike en situation réelle (les expériences ayant ici été menées sur des données simulées).

RÉFÉRENCES

[1] The NoiseX database. http ://spib.rice.edu/spib.

[2] A. Amehraye, D. Pastor, and S. Ben Jebara. On the application of recent results in statistical decision and estimation theory to perceptual filtering of noisy speech signals. In *ISCCSP 06*, 2006.

[3] Enrique Bailly-Baillière, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad Popovici, Fabienne Porée, Belen Ruiz, and Jean-Philippe Thiran. The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.

[4] Hervé Bredin, Guido Aversano, Chafic Mokbel, and Gérard Chollet. The Biosecure Talking-Face Reference System. Accepté à MMUA 2006, May 2006.

[5] Y. Hu and P. Loizou. Incorporating a psychoacoustic model in frequency domain speech enhancement. In *IEEE Signal Processing Letters*, volume 6, pages 270–273, 2004.

[6] J. D Johnston. Transform coding of audio signals using perceptual noise criteria. In *IEEE Jour. Selected Areas Commun*, volume 6, pages 9956–9963, 1998.

[7] John Makhoul and Richard Schwartz. State of the art in continuous speech recognition. In *Natl. Acad. Sci.*, pages 9956–9963, 1995.

[8] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior. Recent Advances in the Automatic Recognition of Audiovisual Speech. In *IEEE*, volume 91, pages 1306 – 1326, September 2003.

[9] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book (for HTK Version 3.2)*. Cambridge University Engineering Department, December 2002.