Estimation de l'aire d'ouverture de la bouche à partir d'information acoustique du signal de parole en utilisant des techniques de déconvolution aveugle

Cong-Thanh DO, Abdeldjalil AÏSSA-EL-BEY, Dominique PASTOR et André GOALIC

Institut TELECOM; TELECOM Bretagne; UMR CNRS 3192 Lab-STICC
Université européenne de Bretagne, France
{thanh.do, abdeldjalil.aissaelbey, dominique.pastor, andre.goalic}@telecom-bretagne.eu

Résumé – Dans cet article, nous proposons une nouvelle méthode pour l'estimation de l'aire d'ouverture de la bouche à partir d'une séquence vidéo du mouvement des lèvres. Dans cette méthode, nous exploiterons les différents degrés de corrélation entre les enveloppes acoustiques et les mouvements visuels, reporté par Grant et Seitz 2000, pour établir un modèle mathématique d'un système SIMO (Single-Input Multiple-Output), où l'aire d'ouverture de la bouche est le "Single Input" inconnu que nous souhaitons estimer. Les énergies RMS (Root Mean Square) dans les sous-bandes du signal de parole sont les "Multiple Outputs" observables du modèle. Le signal d'entrée inconnu peut être directement estimé en utilisant des techniques existantes de déconvolution aveugle. Les séquences audiovisuelles, utilisées pour les tests d'estimation, ont été enregistrées par une "webcam" ordinaire. La moyenne des coefficients de corrélation qui sont calculés entre les meilleures estimations de l'aire d'ouverture de la bouche et celles mesurées, sur un ensemble de 16 phrases en français, est 0.73.

Abstract – In this paper, we propose a new method for estimation of area of mouth opening from a video sequence of the speaking person. In this method, we exploit the different degrees of correlation between acoustic envelopes and visible movements, reported by Grant and Seitz 2000, to establish a mathematical model of a Single-Input Multiple-Output (SIMO) system in which the area of mouth opening is the unknown Single Input that we need to estimate. The subband Root Mean Square (RMS) energies of the speech signal are the observable Multiple Outputs of the model. The unknown input signal can be directly estimated by using the existing blind deconvolution techniques. The audio-visual sequences used for the estimation tests have been recorded by an ordinary webcam. The average of the correlation coefficients, calculated between the best estimated area of mouth opening and the manually mesured one, on a set of 16 French sentences, is 0.73.

1 Introduction

Parmi les caractéristiques visuelles de la parole audiovisuelle, les caractéristiques géométriques des lèvres sont supposées contenir plus d'informations utiles pour la lecture labiale par homme et par machine. Cependant, leurs extractions requièrent des algorithmes robustes souvent difficiles, et une importante charge de calculs dans les scénarios réalistes. Compte tenu d'une séquence audiovisuelle de la parole, les caractéristiques géométriques des lèvres peuvent généralement être estimées en utilisant des méthodes basées sur les images [8] ou des méthodes de type "audio-to-visual mapping". Une courte étude sur les méthodes de type "audio-to-visual mapping" pour l'estimation des caractéristiques visuelles de la parole peut être trouvée dans [5], dans laquelle le mouvement facial est linéairement prédit avec une corrélation moyenne de 0.7 avec le mouvement mesuré. L'aire d'ouverture de la bouche est la zone contenue dans l'intérieur du contour des lèvres (voir Fig. 1), et elle est l'une des informations la plus utile pour la lecture labiale.

Dans cet article, nous proposons une nouvelle méthode pour l'estimation de l'aire d'ouverture de la bouche à partir d'information acoustique du signal de parole en utilisant des techniques de déconvolution aveugle. Dans [7], Grant et Seitz ont démontré que l'amélioration de la détectabilité de la parole visuelle est liée au degré de corrélation entre les enveloppes acoustiques et les mouvements visibles des lèvres. Dans notre approche, nous exploitons ces corrélations pour établir un modèle mathématique d'un système SIMO (Single-Input Multiple-Output) dans lequel l'aire d'ouverture de la bouche est le "Single Input" inconnu. Les énergies RMS (Root Mean Squared) dans les sous-bandes du signal de parole sont les "Multiple Outputs" observables du modèle. Le signal d'entrée inconnu peut être estimé directement en utilisant des techniques de déconvolution aveugle [1]. L'estimation de l'aire d'ouverture de

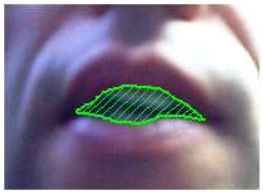

FIG. 1 – L'aire d'ouverture de la bouche est définie comme l'aire contenue dans l'intérieur du contour des lèvres.

la bouche est réalisée sur les séquences audiovisuelles courtes, enregistrées par une "webcam" ordinaire.

2 Formulation du problème et solution

2.1 Modélisation mathématique du problème

Un signal de parole, $y(t)$, est décomposé en N signaux de sous-bandes, $y_i(t), i = 1 \dots N$, en utilisant un banc de filtres de N-canaux :

$$y(t) \approx \sum_{i=1}^{N} y_i(t) \tag{1}$$

La corrélation intrinsèque entre l'énergie RMS, $x_i(t)$, de l'i-ième signal de sous-bande decomposé, $y_i(t)$, et l'aire d'ouverture de la bouche, $s(t)$, peut être modélisée par la convolution ($*$) entre $s(t)$ et la un filtre à réponse impulsionnelle finie (FIR) $h_i(t)$. Les canaux du système sont supposés avoir des réponses impulsionnelles finies. Par conséquent, nous avons

$$x_i(t) = h_i(t) * s(t) + e_i(t) \tag{2}$$

où $e_i(t)$ est l'erreur d'estimation correspondant à la i-ième sous-bande. Cette erreur représente les composantes de $x_i(t)$ qui sont non-corrélées avec (ou orthogonal à) $s(t)$. Donc, avec N sous-bandes, nous avons N équations :

$$\begin{cases} x_1(t) = h_1(t) * s(t) + e_1(t) \\ x_2(t) = h_2(t) * s(t) + e_2(t) \\ \quad \vdots \\ x_N(t) = h_N(t) * s(t) + e_N(t) \end{cases} \tag{3}$$

Le système d'équations (3) représente le modèle d'un système SIMO (Single-Input Multiple-Output). Dans ce modèle, l'aire d'ouverture de la bouche, $s(t)$, est le "Single-Input" inconnu et les énergies RMS dans les sous-bandes, $x_i(t), i = 1, \dots, N$, sont les "Multiple-Outputs" observables.

2.2 Solution du problème

L'estimation de l'aire d'ouverture de la bouche $s(t)$ dans le système d'équations (3) peut être trouvée en utilisant des techniques de déconvolution aveugle. L'estimation directe du signal inconnu $s(t)$ dans (3) est un problème connu comportant un certain nombre de solutions. Les solutions typiques telles que les méthode IS (Input Subspace), MRE (mutually referenced equalizers), and LP (Linear Prediction) peuvent être trouvées dans [1].

La méthode IS (input subspace), proposée dans [9], est pour identifier un système SIMO à réponse impulsionnelle finie (SIMO-FIR), lorsque seulement les sorties du système sont présentes. En comparant à d'autres méthodes, cette méthode est plus efficace et elle n'exige aucune connaissance *à priori* de la corrélation du signal entrée. De plus, cette méthode donne des bons résultats d'estimation même pour des trames courtes du signal [9]. Par conséquent, ces avantages suggèrent que l'IS serait une bonne candidate pour la solution de notre problème d'estimation. Dans cet article, nous appliquons la méthode IS pour estimer l'aire d'ouverture de la bouche $s(t)$ dans (3). La description mathématique de la méthode peut être trouvée dans [9, 1]. Cette méthode nécessite une représentation paramétrique minimale pour résoudre le système. Elle se base sur l'orthogonalité entre les sous-espaces signal et bruit, qui est exploitée pour construire un critère quadratique [9]. Sa minimisation donne l'estimation désirée à un facteur d'échelle prés.

3 Données et structure de banc de filtres

3.1 Données audiovisuelles

Nous évaluons notre méthode sur les séquences audiovisuelles enregistrées par une "webcam" ordinaire. L'objectif est d'évaluer la méthode avec les données n'étant pas de haute qualité. Les 16 phrases enregistrées, en français, sont sélectionnées à partir de la séquence Laval43 de la base de données ATR [10]. Ces phrases sont lues consécutivement par un locuteur de langue maternelle française (F. Berthommier à Gipsa-Lab, Grenoble), et sont enregistrées dans une longue séquence audiovisuelle, en utilisant une "webcam". Puis, cette longue séquence est manuellement segmentée dans 16 séquences courtes (de 3 à 5 secondes), chacune correspondant à une phrase singulière. La fréquence d'échantillonnage vidéo est 25 images/seconde alors que celle du signal audio est 11025 Hz. La "webcam" est centrée sur la région de la bouche du locuteur pour capturer directement la région d'intérêt (Region Of Interest - ROI). Les images capturées sont de dimension 204×148 pixels. La Fig. 1 illustre un exemple de cette image.

3.2 Banc de filtres

Nous utilisons deux types de banc de filtres pour l'extraction des énergies RMS dans les sous-bandes, $x_i(t), i = 1, \dots, N$. Le premier comprend des filtres quasi-rectangulaires en échelle Bark et le deuxième comprend des filtres triangulaires en échelle Mel. D'après [7] et [3], les énergies RMS encodées dans 4 sous-bandes sont optimales pour l'encodage de la redondance audiovisuelle. Nous espérons que les résidus de la parole, encodés dans les énergies des 4 sous-bandes, contiennent les informations utiles pour estimer l'aire d'ouverture de la bouche. La Fig. 2 fait apparaître les 4 filtres quasi-rectangulaires en échelle Bark que nous utilisons pour l'extraction des énergies RMS du signal de parole [3].

Le banc de filtres, avec les filtres triangulaires, comprend 20 filtres en échelle Mel [6]. Les 10 premiers filtres ont leurs fréquences centrales linéairement distribuées de 0 à 1 kHz alors que les 10 derniers filtres ont leurs fréquences centrales régulièrement distribuées en échelle logarithmique, de 1 kHz à la moitié de la fréquence d'échantillonnage du signal de parole (11025 Hz). Nous utilisons deux formes de filtres pour savoir quels filtres donnent les meilleurs résultats d'estimation.

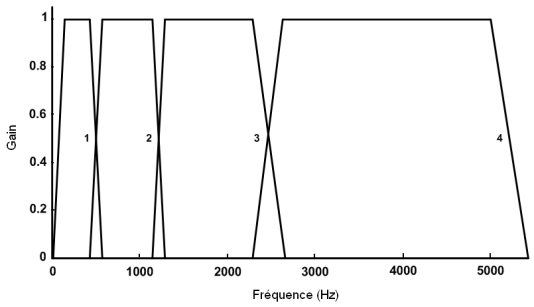

FIG. 2 – Banc de filtres comprenant 4 filtres quasi-rectangulaires, en échelle Bark [3]. Le signal de parole est échantillonné à 11025 Hz.

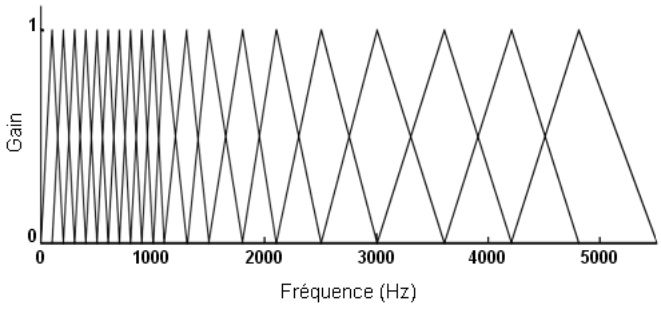

FIG. 3 – Banc de filtres comprenant 20 filtres triangulaires, en échelle Mel, conformément aux filtres utilisés pour le calcul des Mel Frequency Cepstral Coefficients (MFCCs) [6].

4 Estimation de l'aire d'ouverture de la bouche

4.1 Extraction d'énergies RMS

Les énergies RMS dans les sous-bandes sont extraites de toutes les trames de longueur 40 ms, avec 50% de chevauchement entre deux trames contigües. Le fenêtrage Hanning est appliqué sur chaque trame. La fréquence d'échantillonnage d'énergies RMS est alors de 50 Hz. Supposons que $x^B = x_i^B(t), i = 1, \ldots, 4$ et $x^M = x_j^M(t), j = 1, \ldots, 20$ sont les énergies RMS extraites par les filtres quasi-rectangulaires et les filtres triangulaires, respectivement. Nous appliquons deux manipulations sur $x_j^M(t), j = 1, \ldots, 20$. Premièrement, on utilise un sous-ensemble d'énergies RMS $\widetilde{x}^M = x_j^M(t), j = 12, \ldots, 20$ au lieu d'utiliser toutes les énergies de 20 sous-bandes. Nous espérons que les énergies extraites dans les régions de haute fréquence de la parole donneront des meilleurs résultats d'estimation [7]. Deuxièmement, nous appliquons une analyse en composantes principales (ACP) pour extraire les C premières composantes principales, $x^P = x_j^P(t), j = 1, \ldots, C$, des énergies RMS, x^M. Puis, nous utilisons ces composantes principales dans l'algorithme de déconvolution aveugle. L'objectif de l'utilisation des composantes principales est de réduire la redondance dans l'ensemble d'énergies initiales [11, 4].

4.2 Méthode d'évaluation du résultat

Supposons que $\widehat{s}(t)$ et $s(t)$ sont respectivement les estimées et les vraies aires d'ouverture de la bouche. La vraie aire d'ouverture de la bouche, $s(t)$, est extraite à partir des images de la séquence vidéo, qui est synchronisée avec la séquence audio. Sur une image de la séquence vidéo, la largeur des lèvres, A, et la hauteur des lèvres, B, sont calculées en se basant sur les points manuellement marqués de 1 à 4 comme dans la Fig. 4. L'aire d'ouverture de la bouche, S, est approximativement calculée en utilisant la formule $\mathrm{S} = 0.75\mathrm{AB}$ [2].

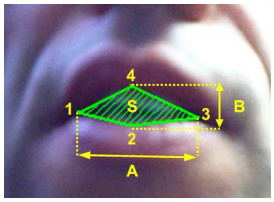

FIG. 4 – Aire d'ouverture de la bouche, S, dans une image, approximativement calculée en utilisant la formule $\mathrm{S} = 0.75\mathrm{AB}$ [2].

Le coefficient de corrélation de Pearson [1], $R\left(\widehat{s}(t), \widetilde{s}(t)\right)$, est utilisé pour évaluer le résultat d'estimation. Il est calculé entre la vraie aire d'ouverture de la bouche, $\widetilde{s}(t)$, et celle estimée, $\widehat{s}(t)$; $\widetilde{s}(t)$ est la vraie aire d'ouverture de la bouche après interpolation linéaire (longueur identique pour $\widehat{s}(t)$ et $\widetilde{s}(t)$).

4.3 L'algorithme d'estimation et résultat

L'algorithme complet pour l'estimation de l'aire d'ouverture de la bouche est présenté dans Fig. 5. Un filtrage temporel est appliqué pour lisser et éliminer les fréquences non-désirées dans le signal obtenu après la déconvolution aveugle. Le filtre pour le filtrage temporel est un filtre passe bas de Butterworth du quatrième ordre, qui a une fréquence de coupure très basse de 3.5 Hz, à la limite supérieure de la dynamique du mouvement oro-facial.

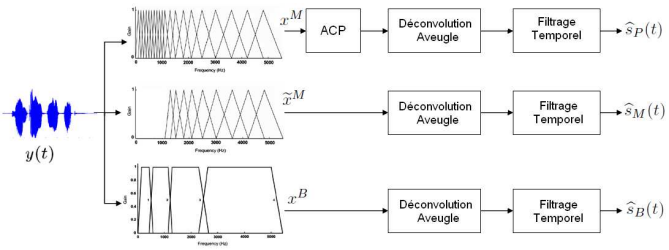

FIG. 5 – Algorithme d'estimation de l'aire d'ouverture de la bouche à partir d'information acoustique du signal de parole en utilisant la technique de déconvolution aveugle (la méthode IS - Input Subspace).

[1] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

L'estimation de l'aire d'ouverture de la bouche est effectuée sur 16 séquences audio courtes de longueur de 3 à 5 secondes, chacune correspondant à une simple phrase. Les coefficients de corrélation, $R\left(\widehat{s}_P(t), \widetilde{s}(t)\right)$, $R\left(\widehat{s}_M(t), \widetilde{s}(t)\right)$, et $R\left(\widehat{s}_B(t), \widetilde{s}(t)\right)$, entre les aires d'ouverture de la bouche estimées, $\widehat{s}_P(t)$, $\widehat{s}_M(t)$, et $\widehat{s}_B(t)$, respectivement, et la vraie aire d'ouverture de la bouche $\widetilde{s}(t)$, sont illustrés dans la Fig. 6. Les valeurs maximales des coefficients de corrélation, $\max\left(R\left(\widehat{s}_P(t), \widetilde{s}(t)\right), R\left(\widehat{s}_M(t), \widetilde{s}(t)\right), R\left(\widehat{s}_B(t), \widetilde{s}(t)\right)\right)$, atteintes par l'un des trois coefficients, pour chaque phrase, sont aussi représentées dans la Fig. 6.

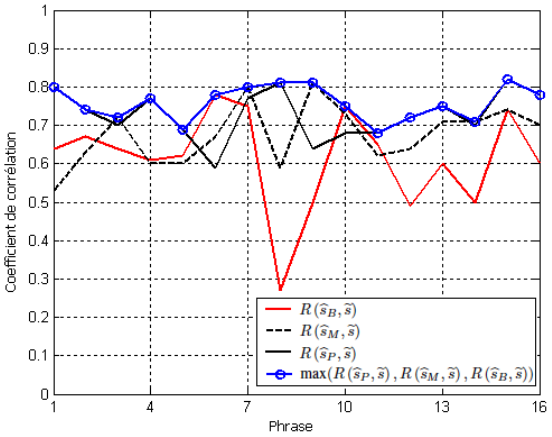

FIG. 6 – Les coefficients de corrélation de Pearson, calculés entre l'aire d'ouverture de la bouche estimée et celle mesurée, pour 16 phrases.

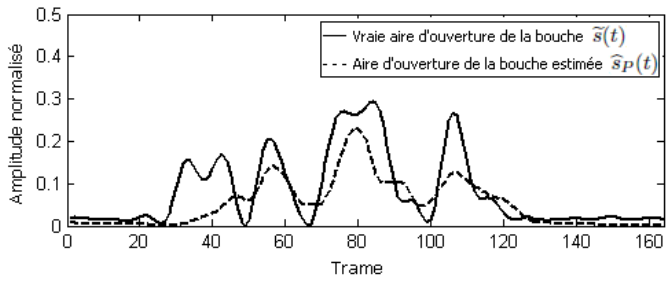

FIG. 7 – L'aire d'ouverture de la bouche estimée pour la phrase "*J'aimais obéir à mes parent*". Le coefficient de corrélation entre la vraie aire d'ouverture de la bouche et celle estimée est $R\left(\widehat{s}_P(t), \widetilde{s}(t)\right) = 0.82$.

La Fig. 6 montre que $\widehat{s}_P(t)$ est le meilleur résultat d'estimation en terme de moyenne des coefficients de corrélation (0.73 par rapport à 0.68 et 0.61 de $\widehat{s}_M(t)$ et $\widehat{s}_B(t)$, respectivement). Il est aussi le plus stable en terme de l'écart-type (0.06 par rapport à 0.08 et 0.13 de $\widehat{s}_M(t)$ et $\widehat{s}_B(t)$, respectivement). La Fig. 7 illustre un exemple de l'aire d'ouverture de la bouche estimée de la phrase *"J'aimais obéir à mes parents."*, 15-ième phrase dans l'ensemble de 16 phrases (voir Fig. 6). Dans cet exemple, les entrées de l'algorithme de déconvolution aveugle sont les 10 premières composantes principales de x^M.

5 Conclusion

Cet article propose une nouvelle méthode d'estimation de l'aire d'ouverture de la bouche à partir d'information acoustique du signal de parole en utilisant des techniques de déconvolution aveugle. Les principaux avantages de cette méthode sont sa simplicité et son faible coût de calcul. Les estimations effectuées sur les séquences audiovisuelles, enregistrées par une "webcam", sont prometteuses. Actuellement, l'utilisation de la méthode IS n'est pas encore complètement automatique car elle exige des paramètres *à priori*, notamment l'ordre du modèle SIMO-FIR. Les travaux futurs se concentreront sur le calcul automatique de l'ordre du modèle SIMO-FIR et sur l'amélioration des performances de la méthode.

Remerciement

Nous tenons à remercier F. Berthommier (Gipsa-Lab, Grenoble) pour l'autorisation d'utilisation de ses données audiovisuelles pour les expérimentations dans ces travaux.

Références

[1] K. Abed-Meraim, W. Qiu, et Y. Hua, "Blind system identification", *Proc. IEEE*, vol. 85, no. 8, pp. 1310-1322, Aug. 1997.

[2] C. Abry, et L.-J. Boë, "Laws for lips", *Speech Communication*, vol. 5, no. 1, pp. 97-104, Mar. 1986.

[3] F. Berthommier, "A phonetically neutral model of the low-level audio-visual interaction", *Speech Communication*, vol. 44, no. 1-4, pp. 31-41, Oct. 2004.

[4] C. Bregler, et Y. Konig, "Eigenlips for robust speech recognition", in *Proc. IEEE ICASSP*, vol. 2, pp. 669-672, 1994.

[5] M.S. Craig, P. Van Lieshout, et W. Wong, "A linear model of acoustic-to-facial mapping : Model parameters, data set size, and generalization across speakers", *J. Acoust. Soc. Am*, vol. 124, no. 5, pp. 3183-3190, Nov. 2008.

[6] S.B. Davis, et P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. ASSP*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

[7] K.W. Grant, et P.-F. Seitz, "The use of visible speech cues for improving auditory detection of spoken language", *J. Acoust. Soc. Am*, vol. 108, no. 3, pp. 1197-1208, Sep. 2000.

[8] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox et R. Harvey, "Extraction of visual features for lipreading", *IEEE Trans. PAMI*, vol. 24, no. 2, pp. 198-213, Feb. 2002.

[9] E. Moulines, P. Duhamel, J.-F. Cardoso, et S. Mayrargue, "Subspace method for the blind identification of multichannel FIR filters", *IEEE Trans. Sig. Process.*, vol. 43, no. 2, pp. 516-525, Feb. 1995.

[10] K.G. Munhall, E. Vatikiotis-Bateson et Y. Tohkura, "X-ray film database for speech research", *J. Acoust. Soc. Am*, vol. 98, no. 2, pp. 1222-1224, Aug. 1995.

[11] M. Turk, et A. Pentland, "Eigenfaces for recognition", *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, Dec. 1990.