Area of Mouth Opening Estimation From Speech Acoustics Using Blind Deconvolution Technique

Cong-Thanh Do, Abdeldjalil Aïssa-El-Bey, Dominique Pastor and André Goalic

Institut TELECOM; TELECOM Bretagne; UMR CNRS 3192 Lab-STICC
Technopôle Brest-Iroise, CS 83818, 29238 Brest Cedex 3
Université européenne de Bretagne, France
{thanh.do, abdeldjalil.aissaelbey, dominique.pastor, andre.goalic}@telecom-bretagne.eu

Abstract

We propose a new method for estimation of area of mouth opening from a video sequence of the speaking person. In a paper published in 2000, Grant and Seitz have reported the different degrees of correlation between acoustic envelopes and visible movements. In our method, we exploit these correlations 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 Squared (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. Our method necessitates only an audio sequence to estimate directly the area of mouth opening in the corresponding video sequence. Consequently, using this method permits us to avoid using complex images processing techniques of the conventional visual features extraction methods, or the training of the estimators in the audio-to-visual mapping methods. The audio-visual sequences used for the estimation tests have been recorded by an ordinary webcam. Estimation result is promising; the estimated area of mouth opening is sufficiently correlated with the manually measured one; the average of correlation coefficients obtained by the most effective configuration of the proposed method, on a set of 16 French sentences, is 0.73.

Index Terms: Lip geometric feature, area of mouth opening, speech temporal envelope processing, SIMO, blind deconvolution

1 Introduction

Amongst the visual features of the audio-visual speech, lip geometric features are assumed to contain most of useful information for speechreading by human and machine. However, their extraction requires robust algorithms which are often difficult and computationally intensive in realistic scenarios. Given an audio-visual sequence of speech, lip geometric features can generally be estimated by using image-based [1] or audio-to-visual mapping methods. In the image-based methods for lip visual features extraction, a region-of-interest (ROI) is needed to be determined for the visual feature extraction algorithm to proceed. Alternatively, specific image processing algorithms such as snakes [2], active shape and appearance models [3, 4] can be used to obtain lip contour estimates. The area of mouth opening is the area contained within the interior lip contour (see Fig. 1), and it is one of the most useful information for lipreading. To obtain the area of mouth opening, the interior lip contour must be firstly extracted and then, the area of mouth opening is calculated. Therefore, image-based area of mouth opening extraction is rather costly.

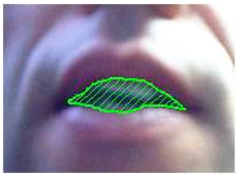

Figure 1: Area of mouth opening is defined as the area contained within the interior lip contour.

The area of mouth opening and other lip geometric features in the video sequence can also be estimated from speech acoustics of the corresponding audio sequence using audio-to-visual mapping methods. In the audio-to-visual mapping approach, we must have sufficiently audio-visual data to train the estimators. These estimators may be linear [5] or non-linear such as those based on hidden Markov models [6, 7] or on time delay neural networks [8]. The interrelation between visual speech features and acoustic ones can be generally approximated by linear models. However, the use of non-linear models is necessary to take into account the dynamics in the audio-visual speech [9]. The main inconvenience of the audio-to-visual mapping methods is the complexity of the realization, especially when we want to use the non-linear estimators. A short review about the audio-to-visual mapping methods for speech visual features estimation can be found in [5], in which facial motion is linearly predicted with a correlation average of 0.7 to the recorded motion.

In this paper, we propose a new method for the estimation of area of mouth opening from speech acoustics using blind deconvolution technique. The main originality of this method lies on its simplicity and its low cost of realization compared to other methods mentioned previously. Further, the method needs only an audio sequence to directly estimate the area of mouth opening in the corresponding video sequence. Consequently, images processing techniques and training are unnecessary. In [10], Grant and Seitz have found that the improvement of detectability of visible speech cues related to the degree of correlation between acoustic envelopes and visible movements. In our approach, we exploit these correlations 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 want to estimate. The

subband Root Mean Squared (RMS) energies of the speech signal are the observable Multiple Outputs of the model. The unknown input signal can be directly estimated by using blind deconvolution techniques in the literature [11]. Area of mouth opening estimation was performed on short audio-visual sequences recorded by an ordinary webcam.

2 Problem Formulation and Solution

2.1 Mathematical Modeling of Problem

A speech signal, $y(t)$, is decomposed into N subband signals, $y_i(t), i = 1 \dots N$, by using a N-channel filterbank:

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

The inherent correlation between the RMS energy, $x_i(t)$, of the i-th decomposed subband signal, $y_i(t)$, and the area of mouth opening, $s(t)$, can be modeled by a convolution $*$ between $s(t)$ and the finite impulse response (FIR) $h_i(t)$ of the i-th system channel. Below, the system channels are supposed to have finite impulse responses. We have

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

where $e_i(t)$ is the estimation error corresponding to the i-th system channel. This error represents the components of $x_i(t)$ that are uncorrelated with (or orthogonal to) $s(t)$. Hence, with N subbands, we have N equations:

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

The system (3) of equations represents the model of a Single-Input Multiple-Output (SIMO) system. In this model, the area of mouth opening, $s(t)$, is the unknown Single-Input and the subband RMS energies, $x_i(t), i = 1, \dots, N$, are the observable Multiple-Output.

2.2 Solution Using Blind Deconvolution Technique

Blind system identification is a fundamental signal processing technique aimed at retrieving a system unknown information from its outputs only [11]. In this case, the word "blind" means that we have neither information about the signal to estimate, $s(t)$, nor about the channel impulse responses, $h_i(t), i = 1, \dots, N$, of the system which is assumed to be linear and shift invariant. Our objective is to estimate the area of mouth opening $s(t)$ as the unknown input signal knowing the output signals $x_i(t), i = 1, \dots, N$. In the domain of blind system identification, the direct estimation of the SIMO system unknown input signal is an existent problem which has a number of solutions. The typical solutions such as input subspace (IS) method, mutually referenced equalizers (MRE) method, and linear prediction (LP) method can be found in [11].

The input subspace (IS) method, proposed in [12], is for identifying a Single-Input Multiple-Output finite impulse response system (SIMO-FIR), when only the outputs of the system are presented. Comparing to other methods, this method is computationally more efficient and it does not require any *a priori* knowledge of the input signal correlation. Further, this method gives good estimation results even for short signal frame [12]. Hence, these strengths suggest that input subspace method would be a good candidate for solving our estimation problem. In this paper, we introduce the results of the area of mouth opening estimation from speech acoustic using the input subspace method. In section 2.3, we will present briefly the mathematical solution of the problem using this method.

2.3 Input Subspace Method

It is more convenient to analyze (3) in its matrix form and to assume that the input and output are discrete signals having length M. Therefore, we write

$$\mathbf{x} = \mathbf{H}_N \mathbf{s} + \mathbf{e} \tag{4}$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1^T & \mathbf{x}_2^T & \cdots & \mathbf{x}_N^T \end{bmatrix}^T$$

$$\mathbf{e} = \begin{bmatrix} \mathbf{e}_1^T & \mathbf{e}_2^T & \cdots & \mathbf{e}_N^T \end{bmatrix}^T$$

and

$$\mathbf{x}_i = [x_i(0), \cdots, x_i(M-1)]^T$$

$$\mathbf{e}_i = [e_i(0), \cdots, e_i(M-1)]^T$$

The superscript ^{T} denotes the transpose and $\mathbf{s}$ is the input vector (area of mouth opening)

$$\mathbf{s} = [s(-L), s(-L+1), \cdots, s(M-1)]^T \tag{5}$$

where L is the model order or the length of the FIRs of the system channels. In (4), $\mathbf{H}_N$ is a generalized Sylvester matrix [12] of dimensions $NM \times (M+L)$

$$\mathbf{H}_N = \begin{bmatrix} \mathbf{H}_{(1)} \\ \mathbf{H}_{(2)} \\ \vdots \\ \mathbf{H}_{(N)} \end{bmatrix} \tag{6}$$

where $\mathbf{H}_{(i)}$ is the $M \times (M+L)$ Sylvester matrix of the i-th system channel response

$$\mathbf{H}_{(i)} = \begin{bmatrix} h_i(L) & \cdots & h_i(0) & \cdots & 0 \\ \vdots & \ddots & & \ddots & \vdots \\ 0 & \cdots & h_i(L) & \cdots & h_i(0) \end{bmatrix} \tag{7}$$

The previous formulation treats the system outputs as a large single vector whereas the input subspace method treats the system outputs as a sequence of small vectors by introducing a window parameter, W, to determine the length of each output vector. We rewrite (4) in the absence of noise, $\mathbf{e}$, as follows

$$[\mathbf{x}(0), \cdots, \mathbf{x}(k), \cdots, \mathbf{x}(M-W)] = \mathbf{H}_N \mathbf{S}_{W+L} \tag{8}$$

where $k = 0, 1, \dots, M-W$ and

$$\mathbf{x}(k) = \begin{bmatrix} \mathbf{x}_1^T(k), \cdots, \mathbf{x}_N^T(k) \end{bmatrix}^T$$

$$\mathbf{x}_i(k) = [x_i(k), \cdots, x_i(k+W-1)]^T$$

In (8), $\mathbf{H}_N$ is the generalized Sylvester matrix, as in (6), but with dimensions $NW \times (W+L)$ because the Sylvester matrices $\mathbf{H}_{(i)}, i = 1, \dots, N$ are now having dimensions $W \times (W+L)$.

Further, $\mathbf{S}_{W+L}$ (8) is a Hankel matrix with dimensions $(W+L)\times(M-W+1)$ and the subspace defined by the rows of $\mathbf{S}_{W+L}$ is called the *input subspace*

$\mathbf{S}_{W+L} =$

$$\begin{bmatrix} s(-L) & s(-L+1) & \cdots & s(M-W-L) \\ s(-L+1) & s(-L+2) & \cdots & s(M-W-L+1) \\ \vdots & \vdots & & \vdots \\ s(W-1) & s(W) & \cdots & s(M-1) \end{bmatrix}$$

Let $\mathbf{V}_0$ be the null space of $\mathbf{S}_{W+L}$, i.e., $\mathbf{S}_{W+L}\mathbf{V}_0 = 0$. If $\mathbf{H}_N$ has full column rank, the data matrix, $\mathbf{x}$, has the same row span as $\mathbf{S}_{W+L}$ [11]. Therefore, the null space $\mathbf{V}_0$ of $\mathbf{S}_{W+L}$ can be calculated from the observable data matrix $\mathbf{x}$. Using the null space $\mathbf{V}_0$ of $\mathbf{S}_{W+L}$ and based on the property of Hankel matrices, we can repeatedly calculate the null space, $\mathbf{V}_r$, of $\mathbf{S}_r, r = W + L - k + 1, k = 1, \ldots, W + L$ by the following formula [11]

$$\mathbf{V}_r = \underbrace{\begin{bmatrix} \mathbf{V}_0 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{V}_0 \end{bmatrix}}_{k\ blocks} \tag{9}$$

where $\mathbf{0}$ is a $1 \times (M - 2W - L + 1)$ vector of zeros. The null space $\mathbf{V}_1$ of $\mathbf{S}_1 = [s(-L), \cdots, s(M-1)]$ is obtained when $k = W + L$. Having the null space $\mathbf{V}_1$, we can calculated the unknown input signal $\mathbf{S}_1$ by solving the following equation

$$\mathbf{S}_1\mathbf{V}_1 = 0 \tag{10}$$

In the presence of additive noise, the input signal $\mathbf{S}_1$ is the one that minimizes $\|\mathbf{S}_1\mathbf{V}_1\|^2$, where $\|.\|$ is the Euclidean norm. The IS method can be summarized as follows

- Calculate the null space $\mathbf{V}_0$ of $\mathbf{S}_{W+L}$ from the observable data matrix.
- Calculate $\mathbf{V}_1$ following (9).
- $\mathbf{S}_1 = \arg\min_{\mathbf{s}_1} \|\mathbf{S}_1\mathbf{V}_1\|^2$.

3 Speech Material and Filterbank Structure

3.1 Audio-Visual Data

We evaluate our estimation method on the audio-visual sequences recorded by an ordinary webcam. The purpose is to assess our method with non-high-quality audio-visual data. The 16 recorded sentences are in French and are selected from the French sentences of the Laval43 sequence of the ATR database [13]. These sentences were read consecutively by a male native French speaker (F. Berthommier at Gipsa-Lab, Grenoble), and were recorded into a long audio-visual sequence (about more than 2 minutes), by using a webcam. This long audio-visual sequence was then manually segmented into short sequences (from 3 to 5 seconds), each one corresponding to a single sentence. The video sampling frequency was 25 images/second whereas the audio sequence was sampled at 11025 Hz. The webcam was centered on the mouth region of the speaker to capture directly the ROI. The captured images are in BITMAP format and are with dimensions 204×148 pixels. Fig. 2 shows examples of such images.

Our method is based on only the speech acoustics information to perform the area of mouth opening estimation. Hence, the video images are not used in the estimation stage. They will be used for the evaluation of the estimation results only.

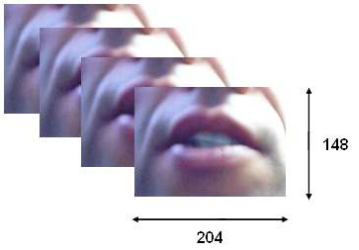

Figure 2: The images captured by the webcam are centered on the speaker's mouth region. They are of dimensions 204×148 pixels.

3.2 Filterbank for RMS Energy Extraction

For the extraction of the speech subband RMS energies, $x_i(t), i = 1, \ldots, N$, we use two types of filterbank. The first one consists of Bark-scaled quasi-rectangular filters and the second one consists of Mel-scaled triangular filters. Following [10] and [14], the 4-subband envelope energy features are found to be optimal for encoding the audio-visual redundancy. We expect that the residual speech cues, encoded in the 4-subband temporal envelopes, contain useful information to estimate the area of mouth opening. Fig. 3 shows the four Bark-scaled and quasi-rectangular filters that we use for subband RMS energies extraction from speech signal, the same as in [14].

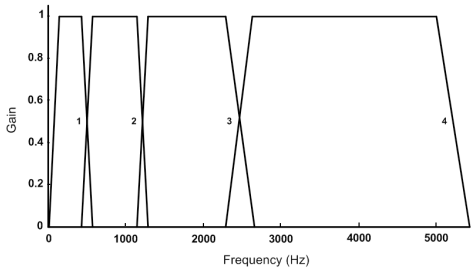

Figure 3: Bark-scaled filterbank for subband RMS energy extraction, the same as in [14]. Four quasi-rectangular filters having high frequency cutoff frequencies at: (1) 515 Hz, (2) 1175 Hz, (3) 2440 Hz, (4) 5250 Hz, respectively. The speech signal sampling frequency is 11025 Hz.

As mentioned above, we use also the Mel-scaled triangular filters, which are used in the calculation of the Mel frequency cepstral coefficient (MFCC) [15], to extract the subband RMS energies. The filterbank consists of 20 Mel-scaled triangular filters (N) as in the original version [15]. The Mel-scaled warping of the frequency axis creates a scaling that is linear below 1 kHz and logarithmic above this limit. The first 10 triangular filters have their central frequencies linearly distributed from 0 to 1 kHz whereas the last 10 filters have their central frequencies equally distributed on a logarithmic scale from 1 kHz to a half of the speech signal sampling frequency (11025 Hz). Our motivation for using Mel-scaled triangular filters beside the Bark-scaled quasi-rectangular

filters is: (1) to compare between the two filter configurations, Bark-scaled quasi-rectangular and Mel-scaled triangular, which one gives better useful information for the area of mouth opening estimation, (2) to investigate the possibility of recovering visual information (area of mouth opening) from the conventional automatic speech recognition acoustic feature (MFCC) by a novel manner. The filterbank consisting of 20 Mel-scaled triangular filters, used in this study, is shown in Fig. 4.

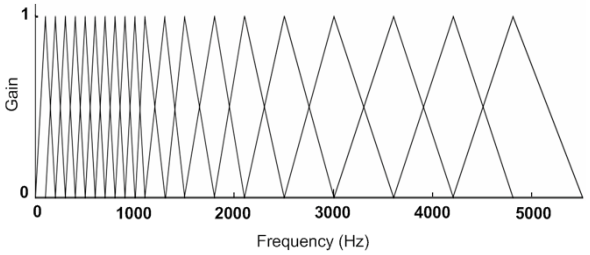

Figure 4: Filterbank consisting of 20 Mel-scaled triangular filters for subband RMS energies extraction, in accordance with the filters used for the MFCC calculation [15]. The first 10 triangular filters have their central frequencies linearly distributed from 0 to 1 kHz whereas the last 10 filters have their central frequencies equally distributed on a logarithmic scale from 1 kHz to a half of the speech signal sampling frequency (11025 Hz).

4 Area of Mouth Opening Estimation

4.1 RMS Energy Extraction

The subband RMS energies are extracted from every 40 ms Hanning windowed subband speech signal with 50% overlap between two adjacent windows. The frame rate of the subband RMS energies is thus 50 Hz. This frame rate guarantees a subband RMS energies bandwidth equals 50/2 = 25 Hz, which is greater than the upper bound of the vocal tract motion (6.25 Hz) [14]. The trade-off between the signal bandwidth and the blind deconvolution time is also satisfied at this frame rate since the longer the signals are, the slower the blind deconvolution is. Let $\mathbf{x}_B = \left[\mathbf{x}_{B1}^T\, \mathbf{x}_{B2}^T\, \mathbf{x}_{B3}^T\, \mathbf{x}_{B4}^T\right]^T$ and $\mathbf{x}_M = \left[\mathbf{x}_{M1}^T\, \mathbf{x}_{M2}^T \ldots \mathbf{x}_{M20}^T\right]^T$ are the subband RMS energies vectors extracted by the Bark-scaled quasi-rectangular and the Mel-scaled triangular filters, respectively. The subvectors of $\mathbf{x}_B$ are $\mathbf{x}_{Bi} = [x_{Bi}(0), \ldots, x_{Bi}(M-1)]^T, i = 1, \ldots, 4$ where M is the signal length. Similarly, $\mathbf{x}_{Mj} = [x_{Mj}(0), \ldots, x_{Mj}(M-1)]^T, j = 1, \ldots, 20$ are the subvectors of $\mathbf{x}_M$.

Before performing the blind deconvolution, we apply two manipulations to the subband RMS energies vector, $\mathbf{x}_M$, extracted by using the Mel-scaled triangular filters. First, motivated by the stronger correlation between the area of mouth opening and the acoustic energy modulations in the *F2* (800–2200 Hz) and *F3* (2200–6500 Hz) regions [10], we use only a subvector, $\widetilde{\mathbf{x}}_M = \left[\mathbf{x}_{M12}^T\, \mathbf{x}_{M13}^T \ldots \mathbf{x}_{M20}^T\right]^T$, of nine subband RMS energies from the original vector, $\mathbf{x}_M$, for blind deconvolution. This subvector, $\widetilde{\mathbf{x}}_M$, contains the subband RMS energies extracted from the high frequency region of the speech signal, is expected to carry the most appropriate information regarding the area of mouth opening. The subband RMS energies contained in $\widetilde{\mathbf{x}}_M$, are extracted by nine Mel-scaled triangular filters, which cover the speech frequency region above 1.1 kHz, and have the central frequencies equally distributed on a logarithmic scale. The second manipulation is to extract the principal components of the subband RMS energies, $\mathbf{x}_M$, by performing a principal component analysis (PCA). These first C principal components, $\widehat{\mathbf{x}}_P = \left[\widehat{\mathbf{x}}_{P1}^T\, \widehat{\mathbf{x}}_{P2}^T \ldots \widehat{\mathbf{x}}_{PC}^T\right]^T$, are then used in the blind deconvolution algorithm. The principal components $\widehat{\mathbf{x}}_{Pi} = [\widehat{x}_{Pi}(0), \ldots, \widehat{x}_{Pi}(M-1)]^T, i = 1, \ldots, C$, have the same length, M, as of the subvectors of $\mathbf{x}_M$. The approach of using the principal components of a dataset instead of using the original dataset was first introduced by Turk and Pentland in face recognition [16], and then by Bregler and Konig in automatic audio-visual speech recognition [17]. The principal components, that is the eigenvectors of the covariance matrix of the subband RMS energies vector, $\mathbf{x}_M$, can be thought of as a set of features that together characterize the variation between the subband RMS energies, $\mathbf{x}_{M1}^T, \mathbf{x}_{M2}^T, \ldots, \mathbf{x}_{M20}^T$. We expected that the detrimental effect of the subband RMS energies redundancy on the blind deconvolution process, will be eliminated by using the subband RMS energies principal components, $\widehat{\mathbf{x}}_P$, instead of using the subband RMS energies, $\mathbf{x}_M$, itself.

4.2 Evaluation Method

Assuming that $\widehat{\mathbf{s}} = [\widehat{s}(0), \widehat{s}(1), \ldots, \widehat{s}(M-1)]^T$, and $\mathbf{s} = [s(0), s(1), \ldots, s(K-1)]^T$, are the estimated and true areas of mouth opening, respectively. The true area of mouth opening, $\mathbf{s}$, is extracted from the images of the video sequence, which is synchronous with the audio signal. On the i-th image, $i = 1, \ldots, K$ of the video sequence, the lip width, A_i, and the lip height, B_i, are calculated based on the manually marked points from 1 to 4 as in Fig. 5. The areas of mouth opening, $s(i), i = 1, \ldots, K$, are approximately calculated by using the formula $s(i) = 0.75\mathrm{A}_i\mathrm{B}_i, i = 1, \ldots, K$ [18].

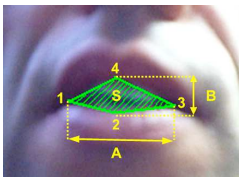

Figure 5: The area of mouth opening, S, in an image is approximately calculated by using the formula S = 0.75AB [18]. The lip width, A, and the lip height, B, are calculated based on the manually marked points from 1 to 4.

We use the Pearson product-moment correlation coefficient between the estimated area of mouth opening, $\widehat{\mathbf{s}}$, and the true area of mouth opening, $\mathbf{s}$, to evaluate the correctness of the estimation results. In general, the length, M, of $\widehat{\mathbf{s}}$ is greater than the length, K, of $\mathbf{s}$, because the frame rate of the subband RMS energies is greater than that of the video sequence. The true area of mouth opening is therefore linearly interpolated to have the same length, M, as the estimated area of mouth opening. The Pearson product-moment correlation coefficient, $R\left(\widehat{\mathbf{s}}, \widetilde{\mathbf{s}}\right)$, is then calculated between the estimated area of mouth opening, $\widehat{\mathbf{s}}$, and the true area of mouth opening after linear interpolation,

$\widetilde{\mathbf{s}} = [\widetilde{s}(0), \ldots, \widetilde{s}(M-1)]^T$, by the formula

$$R\left(\widehat{\mathbf{s}}, \widetilde{\mathbf{s}}\right) = \frac{1}{M} \sum_{i=0}^{M-1} \left(\frac{\widehat{s}(i) - \mu_{\widehat{\mathbf{s}}}}{\sigma_{\widehat{\mathbf{s}}}} \right) \left(\frac{\widetilde{s}(i) - \mu_{\widetilde{\mathbf{s}}}}{\sigma_{\widetilde{\mathbf{s}}}} \right) \qquad (11)$$

where $\mu_{\widehat{\mathbf{s}}}$ and $\mu_{\widetilde{\mathbf{s}}}$ are the sample means, whereas $\sigma_{\widehat{\mathbf{s}}}$ and $\sigma_{\widetilde{\mathbf{s}}}$ are the standard deviations of $\widehat{\mathbf{s}}$ and $\widetilde{\mathbf{s}}$, respectively.

4.3 Area of Mouth Opening Estimation Algorithm

The complete algorithm for the area of mouth opening estimation is shown in Fig. 6. As mentioned previously, the extracted subband RMS energies, $\mathbf{x}_B, \widetilde{\mathbf{x}}_M$, and $\widehat{\mathbf{x}}_P$, are used as the inputs of the blind deconvolution algorithm. After performing the blind deconvolution on the subband RMS energies, a temporal filtering is applied to smooth the blind deconvolved signals, $\widetilde{\mathbf{s}}_P, \widetilde{\mathbf{s}}_M$, and $\widetilde{\mathbf{s}}_B$. This filtering eliminates also the frequencies existing in the subband RMS energies but these frequencies are the undesired components in the signals, $\widetilde{\mathbf{s}}_P, \widetilde{\mathbf{s}}_M$, and $\widetilde{\mathbf{s}}_B$, after deconvolution. The filter used for the temporal smoothing is a fourth-order Butterworth lowpass filter, which has a very low cutoff frequency of 3.5 Hz, in the upper bound of the orofacial motion range [14]. The signals after temporal filtering, $\widehat{\mathbf{s}}_P, \widehat{\mathbf{s}}_M$, and $\widehat{\mathbf{s}}_B$, are the estimated areas of mouth opening.

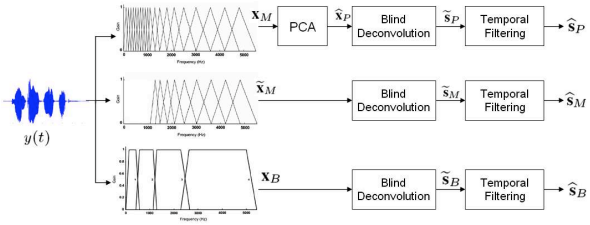

Figure 6: Algorithm for the estimation of the area of mouth opening from speech acoutics using blind deconvolution techniques. The subband RMS energies, $\widehat{\mathbf{x}}_P, \widetilde{\mathbf{x}}_M$, and $\mathbf{x}_B$, are extracted from the speech signal by using different types of filterbanks. The estimated areas of mouth opening, $\widehat{\mathbf{s}}_P, \widehat{\mathbf{s}}_M$, and $\widehat{\mathbf{s}}_B$, are obtained after a temporal filtering of the blind deconvolved signals, $\widetilde{\mathbf{s}}_P, \widetilde{\mathbf{s}}_M$, and $\widetilde{\mathbf{s}}_B$, respectively.

The algorithm used for the blind deconvolution is the input subspace method as mentioned previously. This method needs two *a priori* parameters, the model order, L, and the window parameter, W (see section 2.3). In this current work, we vary systematically the values of L and W and choose the values of L and W which maximize the correlations between the estimated signals, $\widehat{\mathbf{s}}_P, \widehat{\mathbf{s}}_M$, and $\widehat{\mathbf{s}}_B$ and the true area of mouth opening, $\widetilde{\mathbf{s}}$. The automatic calculation of the parameters L and W is beyond the scope of this paper. In addition, the number of principal components, C , for $\widehat{\mathbf{x}}_P$ calculation is also manually adjusted to attain the maximum correlation between the estimated, $\widehat{\mathbf{s}}_P$, and the true, $\widetilde{\mathbf{s}}$, areas of mouth opening.

4.4 Estimation Results

The area of mouth opening estimation are performed on 16 short audio sequences of length ranging from 3 to 5 seconds, each one corresponding to a single sentence (see section 3.1). The correlation coefficients, $R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right)$, $R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right)$, and $R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right)$, between the estimated areas of mouth opening, $\widehat{\mathbf{s}}_P$, $\widehat{\mathbf{s}}_M$ and $\widehat{\mathbf{s}}_B$, respectively, and the true area of mouth opening after linear interpolation, $\widetilde{\mathbf{s}}$, are shown in Fig. 7. In addition, the correlation coefficients maximum values, $\max(R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right), R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right), R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right))$, attained from one of three correlation coefficients for each sentence are also represented in Fig. 7. The variation of the correlation coefficients in Fig. 7 shows that, the goodness of the estimation results depends not only on the type of subband RMS energies that have been used, but also on each particular sentence. This dependence is comprehensible since the method is based on the inherent correlation between the acoustic envelopes and the lip visible movements. Meanwhile, this inherent correlation is sentence-dependent as reported by Grant and Seitz [10].

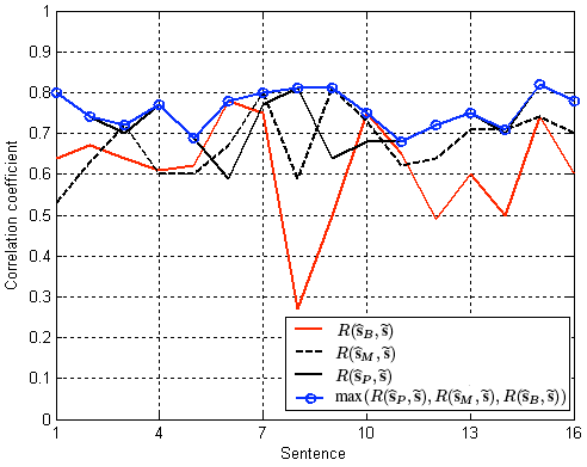

Figure 7: Pearson product-moment correlation coefficients, $R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right)$, $R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right)$, and $R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right)$, between the estimated areas of mouth opening, $\widehat{\mathbf{s}}_P$, $\widehat{\mathbf{s}}_M$ and $\widehat{\mathbf{s}}_B$, respectively, and the true area of mouth opening after linear interpolation, $\widetilde{\mathbf{s}}$. $\max(R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right), R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right), R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right))$ is the maximum correlation coefficient attained for each sentence.

Table 1: Means (μ) and standard deviations (σ) of the correlation coefficients, $R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right)$, $R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right)$, $R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right)$, and $\max(R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right), R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right), R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right))$.

	$R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right)$	$R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right)$	$R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right)$	Max
μ	0.73	0.68	0.61	0.76
σ	0.06	0.08	0.13	0.04

Table 1 shows the empirical means and the standard deviations of the correlation coefficients. A one-way ANOVA reveals that $R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right)$ is significantly greater than $R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right)$ $[F(1,30) = 4.30, p < 0.05]$, and $R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right)$ $[F(1,30) = 10.19, p < 0.005]$. However, the difference between $R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right)$ and $R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right)$ is not significant $[F(1,30) = 2.72, p > 0.1]$. In addition, no significant difference is revealed between $R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right)$ and the $\max(R\left(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}\right), R\left(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}\right), R\left(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}\right))$ $[F(1,30) = 2.45, p > 0.1]$. Therefore, the estimation results obtained with $\widehat{\mathbf{x}}_P$ is the highest ($\mu_{R(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}})} = 0.73$) and the most stable ($\sigma_{R(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}})} = 0.06$).

Fig. 8 shows an example of the estimated areas of mouth opening for the French sentence *"J'aimais obéir à mes parents."*,

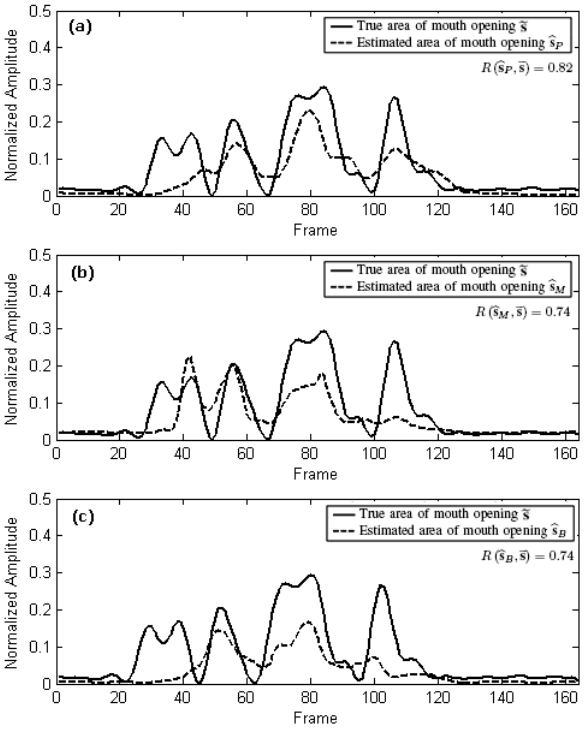

Figure 8: Estimated areas of mouth opening for the French sentence *"J'aimais obéir à mes parents."*. The true area of mouth opening, $\widetilde{\mathbf{s}}$, is represented by solid line whereas the estimated areas of mouth opening, $\widehat{\mathbf{s}}_P$, $\widehat{\mathbf{s}}_M$, and $\widehat{\mathbf{s}}_B$ are represented by dashed lines in the panel (a), (b), and (c), respectively. The correlation coefficient in each case is figured in each panel.

which is the 15th sentence in the 16-sentence set for estimation (see Fig. 7). In each subfigure, the true area of mouth opening is represented by a solid line whereas the estimated area of mouth opening is represented by a dashed line. The correlation coefficients obtained for this sentence, $R(\widehat{\mathbf{s}}_P, \widetilde{\mathbf{s}}) = 0.82$, $R(\widehat{\mathbf{s}}_M, \widetilde{\mathbf{s}}) = 0.74$, and $R(\widehat{\mathbf{s}}_B, \widetilde{\mathbf{s}}) = 0.74$ are sufficiently good. In this example, $\widehat{\mathbf{x}}_P$ consists of the first ten principal components.

5 Conclusion

This paper proposes a new method for the estimation of the area of mouth opening from only speech acoustics using blind deconvolution technique. The main advantage of this method lies on its simplicity and the low cost of realization. On the basis of a given audio sequence only, we can estimate directly the area of mouth opening in the corresponding video sequence, without manipulating the images of the video sequence or training audio-to-visual mapping estimators. Estimation result performed on webcam-recorded audio-visual sequences is promising; the estimated area of mouth opening is sufficiently correlated with manually measured one. This method suggests that the area of mouth opening estimation from speech acoustics might be analytically done. Actually, automatic calculation of the *a priori* parameters of the input subspace method, such as the model order L and the window length W, is still an open issue. Further study is needed to be carried out to make the method completely automatic.

Acknowledgement

We would like to thank F. Berthommier (Gipsa-Lab, Grenoble) for permission to use his webcam-recorded audio-visual sequences for the experiments in this paper.

References

[1] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, 2002.

[2] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *Intl. J. Comp. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models – Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," in *Proc. Eur. Conf. Comp. Vis.*, 1998, pp. 484–498.

[5] M. S. Craig, P. Van Lieshout, and 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, 2008.

[6] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip Movement Synthesis from Speech based on Hidden Markov Models," *Speech Commun.*, vol. 26, no. 1-2, pp. 105–115, 1998.

[7] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. K. Kakumanu, and O. N. Garcia, "Audio/Visual Mapping with Cross-Modal Hidden Markov Models," *IEEE Trans. on Multimedia*, vol. 7, no. 2, pp. 243–252, 2005.

[8] S. Morishima and H. Harashima, "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 4, pp. 594–600, 1991.

[9] J. P. Barker and F. Berthommier, "Estimation of Speech Acoustics from Visual Speech Features: A Comparison of Linear and Non-Linear Models," in *Proc. Intl. Conf. AVSP, Santa Cruz, CA, USA*, 1999, pp. 112–117.

[10] K. W. Grant and 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, 2000.

[11] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind System Identification," *Proc. IEEE*, vol. 85, no. 8, pp. 1310–1322, 1997.

[12] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace Method for the Blind Identification of Multichannel FIR Filters," *IEEE Trans. on Sig. Process.*, vol. 43, no. 2, pp. 516–525, 1995.

[13] K. G. Munhall, E. Vatikiotis-Bateson, and Y. Tohkura, "X-ray Film Database for Speech Research," *J. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 1222–1224, 1995.

[14] F. Berthommier, "A Phonetically Neutral Model of the Low-level Audio-visual Interaction," *Speech Commun.*, vol. 44, no. 1-4, pp. 31–41, 2004.

[15] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on ASSP.*, vol. 28, no. 4, pp. 357–366, 1980.

[16] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1990.

[17] C. Bregler and Y. Konig, "Eigenlips for Robust Speech Recognition," in *Proc. IEEE ICASSP*, vol. 2, 1994, pp. 669–672.

[18] C. Abry and L.-J. Boë, "Laws for Lips," *Speech Commun.*, vol. 5, no. 1, pp. 97–104, 1986.