MEASURING AUDIO AND VISUAL SPEECH SYNCHRONY: METHODS AND APPLICATIONS

H. Bredin and G. Chollet

GET-ENST – CNRS-LTCI – Département TSI – 46 rue Barrault, 75013 Paris - FRANCE

Keywords: audio-visual speech, synchrony.

Abstract

Speech is a means of communication that is intrinsically bimodal: the audio signal originates from the dynamics of the articulators. This paper reviews recent works in the field of audiovisual speech and more specifically on techniques developed to measure the level of correspondence between audio and visual speech. It overviews the most common audio and visual speech front-end processing, transformations performed on audio, visual or joint audiovisual feature spaces and the actual measure of correspondence between audio and visual speech. Finally, applications of this specific task are described.

1 Introduction

Speech is a means of communication that is intrinsically bimodal: the audio signal originates from the dynamics of the articulators. Both audible and visible speech cues carry relevant information. Though the first automatic speech-based recognition systems were only relying on its auditory part (whether it is speech recognition or speaker verification), it is well known that its visual counterpart can be a great help, especially under adverse conditions [23]. In noisy environments for example, audiovisual speech recognizers perform better than audio-only systems. Using visual speech as a second source of information for speaker verification has also been experimented, even though resulting improvements are not always significant.

This review tries to complement existing surveys about audiovisual speech processing. It does not address the problem of audiovisual speech recognition nor speaker verification: these two issues are already covered in [6] and [8]. Moreover, because of length constraints, this paper does not tackle the question of the estimation of visual speech from its acoustic counterpart (or reciprocally): the reader might want to have a look at [33] and [1] showing that linear methods can lead to very good estimates.

This paper focuses on the measure of correspondence between acoustic and visual speech. How correlated the two signals are? Can we detect a lack of correspondence between them? Is it possible to decide (putting aside any biometric method), among a few people appearing in a video, who is talking?

Section 2 overviews the acoustic and visual front-ends processing. They are often very similar to the one used for speech recognition and speaker verification, though a tendency to simplify them as much as possible has been noticed. Moreover, linear transformations aiming at improving joint audiovisual modeling are often performed as a preliminary step before measuring the audio-visual correspondence: they will be discussed in section 3. The correspondence measures proposed in the literature are then presented in section 4. Finally, applications of these techniques in different technological areas are presented in section 5.

2 Front-end processing

This section reviews the speech front-end processing techniques used in the literature for audio-visual speech processing in the specific framework of audiovisual speech synchrony measures. They all share the common goal of reducing the raw data in order to achieve a good subsequent modelling.

2.1 Acoustic speech processing

Acoustic speech parameterization is classically performed on overlapping sliding window of the original audio signal.

Raw energy The raw amplitude of the audio signal can be used as is. In [18], the authors extract the average acoustic energy on the current window as their one-dimensional audio feature. Similar methods such as root mean square amplitude or log-energy were also proposed [1, 3].

Periodogram In [14], the periodogram was proposed for audio speech parameterization.

Mel-Frequency Cepstral Coefficients (MFCC) The use of MFCC parameterization is very frequent in the literature [26, 10, 22, 21, 7]. There is a practical reason for that: it is the state-of-the-art [25] parameterization for speech processing in general, including speech recognition and speaker verification.

Linear-Predictive Coding and Line Spectral Frequencies (LPC and LSF) Linear-Predictive Coding, and its derivation Line Spectral Frequencies [30], have also been widely investigated. The latter are often preferred because they were shown to be strongly related to the vocal tract geometry [33].

A comparison of these different acoustic speech features is performed in [26] in the framework of the *FaceSync* linear operator, which is presented below. To summarize, in their specific framework, the authors conclude that MFCC, LSF and LPC parameterizations lead to a stronger correlation with the visual speech than spectrogram and raw energy features. This result is coherent with the observation that these features are the ones known to give good results for speech recognition.

2.2 Visual speech processing

In this section, we will refer to the gray-level mouth area as the region of interest. It can be much larger than the sole lip area and can include jaw and cheeks. In the following, it is assumed that the detection of this region of interest has already been performed. Most of visual speech features proposed in the literature are shared by studies in audiovisual speech recognition. However, some much more simple visual features are also used for synchronization detection.

Raw pixels This is the visual equivalent of the audio raw energy. In [18] and [21], the intensity of gray-level pixels is used as is. In [3], their sum over the whole region of interest is computed, leading to a one-dimensional feature.

Holistic methods Holistic methods consider and process the region of interest as a whole source of information. In [22], a two-dimensional discrete cosine transform is applied on the region of interest and the most energetic coefficients are kept as visual features: it is a well-known method in the field of image compression. Linear transformations taking into account the specific distribution of gray-level in the region of interest were also investigated. Thus, in [4], the authors perform a projection of the region of interest on vectors resulting from a principal component analysis: they call the principal components ‘eigenlips’ by analogy with the well-known ‘eigenfaces’ [32] principle used for face recognition.

Lip-shape methods Lip-shape methods consider and process lips as a deformable object from which geometrical features can be derived, such as height, width openness of the mouth, position of lip corners, etc. They are often based on fiducial points that need to be automatically located. In [1], videos available are recorded using two cameras (one frontal, one from side) and the automatic localization is made easier by the use of face make-up: both frontal and profile measures are then extracted and used as visual features. Mouth width, mouth height and lip protrusion are computed in [17], jointly with what the authors call the relative teeth count that can be considered as a measure of the visibility of teeth. In [13] [12], a deformable template composed of several polynomial curves follows the lip contours: it allows the computation of the mouth width, height and area. In [7], the lip shape is summarized with a one-dimensional feature: the ratio of lip height and lip width.

Dynamic features In [8] the authors underline that, though it is widely agreed that an important part of speech information is conveyed dynamically, dynamic features extraction is rarely performed: this observation is also verified for correspondence measures. However, some attempts to capture dynamic information within the extracted features do exist in the literature. Thus, the use of time derivatives is investigated in [16]. In [10], the authors compute the total temporal variation (between two subsequent frames) of pixel values in the region of interest, following the equation 1:

$$v_t = \sum_{x=1}^{W}\sum_{y=1}^{H} |I_t(x,y) - I_{t+1}(x,y)| \qquad (1)$$

where $I_t(x,y)$ is the grey-level pixel value of the region of interest at position (x,y) in frame t.

2.3 Frame rates

Audio and visual sample rates are classically very different. For speaker verification, for example, MFCCs are usually extracted every 10 ms ; whereas videos are often encoded at a frame rate of 25 images per second. Therefore, it is often required to down-sample audio features or up-sample visual features in order to equalize audio and visual sample rates. However, though the extraction of raw energy or periodogram can be performed directly on a larger window, down-sampling audio features is known to be very bad for speech recognition. Therefore, up-sampling visual features is often preferred (with linear interpolation, for example). One could also think of using a camera able to produce 100 images per second. Finally, some studies (like the one presented in section 4.3.2) directly work on audio and visual features with unbalanced sample rates.

3 Audiovisual subspaces

In this section, we overview transformations than can be applied on audio, visual and/or audiovisual spaces, with the aim of improving subsequent measure of correspondence between audio and visual clues.

3.1 Principal component analysis

Principal Component Analysis (PCA) is a well-known linear transformation that is optimal for keeping the subspace that has largest variance. The basis of the resulting subspace is a collection of principal components. The first principal component corresponds to the direction of greatest variance of a given dataset. The second principal component corresponds to the direction of second greatest variance, and so on. In [9], PCA is used in order to reduce the dimensionality of a joint audiovisual space (in which audio speech features and visual speech features are concatenated), while keeping the characteristics that contribute most to its variance.

3.2 Independent component analysis

Independent Component Analysis (ICA) was originally introduced to deal with the issue of source separation [19]. In [28], the authors use visual speech features to improve separation of speech sources. In [27], ICA is applied on an audiovisual recording of a piano session: the camera frames a close-up on the keyboard when the microphone is recording the music. ICA allows to clearly find a correspondence between the audio and visual note. However, to our knowledge, ICA has never been used as a transformation of the audiovisual speech feature space (as in [27] for the piano). A Matlab implementation of ICA is available on the Internet [20].

3.3 Canonical correlation analysis

Canonical Correlation Analysis (CANCOR) is a statistical analysis allowing to jointly transform the audio and visual feature spaces while maximizing the audiovisual cross-correlation. Given two synchronized random variables X and Y, the *FaceSync* algorithm presented in [26] uses CANCOR to find canonic correlation matrices A_X and A_Y that whiten X and Y under the constraint of making their cross-correlation diagonal and maximally compact. Let $\mathbf{X} = (X - \mu_X)^T A_X$, $\mathbf{Y} = (Y - \mu_Y)^T A_Y$ and $\Sigma_{\mathbf{XY}} = \mathbb{E}[\mathbf{XY}^T]$.
These constraints can be summarized as follows:

Whitening: $\mathbb{E}[\mathbf{XX}^T] = \mathbb{E}[\mathbf{YY}^T] = I$
Diagonal :
$\Sigma_{\mathbf{XY}} = \text{diag}\{\sigma_1, \ldots, \sigma_M\}$ with $1 \geq \sigma_1 \geq \ldots \geq \sigma_m > 0$ and $\sigma_{m+1} = \ldots = \sigma_M = 0$
Maximally compact:
For i from 1 to M, the correlation $\sigma_i = \text{corr}(\mathbf{X}_i, \mathbf{Y}_i)$ between $\mathbf{X}_i$ and $\mathbf{Y}_i$ is as large as possible.

The proof of the algorithm for computing A_X and A_Y is described in [26]. A Matlab implementation of this transformation is also available on the Internet [5].

3.4 Co-inertia analysis

Co-Inertia Analysis (CoIA) is quite similar to CANCOR. However, while CANCOR is based on the maximization of the correlation between audio and visual features, CoIA relies on the maximization of their covariance $\text{cov}(\mathbf{X}_i, \mathbf{Y}_i) = \text{corr}(\mathbf{X}_i, \mathbf{Y}_i) \times \text{var}(\mathbf{X}_i) \times \text{var}(\mathbf{Y}_i)$.
This statistical analysis was first introduced in biology and is relatively new in our domain. The proof of the algorithm for computing A_X and A_Y can be found in [11].

Remark Comparative studies between CANCOR and CoIA are proposed in [17, 13, 12]. The authors of [17] show that CoIA is more stable than CANCOR: the accuracy of the results is much less sensitive to the number of samples available. The *liveness* score (see section 5) proposed in [13, 12] is much more efficient with CoIA than CANCOR. The authors of [13] suggest that this difference is explained by the fact that CoIA is a compromise between CANCOR (where audiovisual correlation is maximized) and PCA (where audio and visual variances are maximized) and therefore benefits from the advantages of both transformations.

4 Correspondence measures

This section overviews the correspondence measures proposed in the literature to evaluate the synchrony between audio and visual features resulting from audiovisual front-end processing and transformations described in sections 2 and 3.

4.1 Pearson's product-moment coefficient

Let X and Y be two independent random variables which are normally distributed. Assuming a linear relationship between X and Y, the square of their Pearson's product-moment coefficient R(X,Y) (defined in equation 2) denotes the portion of total variance of X that can be explained by a linear transformation of Y (and reciprocally, since it is a symmetrical measure).

$$R(X,Y) = \frac{\text{cov}\,(X,Y)}{\sigma_X \sigma_Y} \qquad (2)$$

In [18], the authors compute the Pearson's product-moment coefficient between the average acoustic energy X and the value Y of the pixels of the video to determine which area of the video is more correlated with the audio. This allows to decide which of two people appearing in a video is talking.

4.2 Mutual information

In information theory, the mutual information MI(X,Y) of two random variables X and Y is a quantity that measures the mutual dependence of the two variables. In the case of X and Y are discrete random variables, it is defined as in equation 3.

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad (3)$$

It is non-negative (MI(X,Y) ≥ 0) and symmetrical (MI(X,Y)=MI(Y,X)). One can demonstrate that X and Y are independent if and only if MI(X,Y)=0. The mutual information can also be linked to the concept of entropy H in information theory as shown in equation 5:

$$\begin{aligned} MI(X,Y) &= H(X) - H(X|Y) \qquad (4) \\ MI(X,Y) &= H(X) + H(Y) - H(X,Y) \qquad (5) \end{aligned}$$

As shown in [18], in the special case where X and Y are normally distributed mono-dimensional random variables, the mutual information is related to R(X,Y) via the equation 6:

$$MI(X,Y) = -\frac{1}{2} \log \left(1 - R(X,Y)^2\right) \qquad (6)$$

In [18, 15, 22, 21], the mutual information is used to locate the pixels in the video that are most likely to correspond to the audio signal: the face of the person who is speaking clearly corresponds to these pixels. However, one can notice that the mouth area is not always the part of the face with the maximum mutual information with the audio signal: it is very dependent on the speaker.

Remark In [4], the mutual information between audio X and time-shifted visual Y_t features is plotted, as a function of their temporal offset t. It shows that the mutual information reaches its maximum for a visual delay of between 0 and 120 ms. This observation led the authors of [13, 12] to propose a liveness score L(X,Y) based on the maximum value R_{ref} of the Pearson's coefficient for short time offset between audio and visual features.

$$R_{\text{ref}} = \max_{-2 \leq t \leq 0} [R(X, Y_t)] \tag{7}$$

4.3 Joint audiovisual models

Though the Pearson's coefficient and the mutual information are good at measuring correspondence between two random variables even if they are not linearly correlated (which is what they were primarily defined for), some other methods do not rely on this linear assumption.

4.3.1 Parametric models

Gaussian Mixture Models Let consider two discrete random variables $X = \{x_t, t \in N\}$ and $Y = \{y_t, t \in N\}$ of dimension d_X and d_Y respectively. Typically, X would be acoustic speech features and Y visual speech features [29, 7]. One can define the discrete random variable $Z = \{z_t , t \in N\}$ of dimension d_Z where z_t is the concatenation of the two samples x_t and y_t, such as $z_t = [x_t , y_t]$ and $d_Z = d_X + d_Y$.
Given a sample z, the Gaussian Mixture Model λ defines its probability distribution function as follows:

$$p(z|\lambda) = \sum_{i=1}^{N} w_i \mathcal{N}(z; \mu_i, \Gamma_i) \tag{8}$$

where $\mathcal{N}(\bullet; \mu, \Gamma)$ is the normal distribution of mean μ and covariance matrix Γ. $\lambda = \{w_i,\mu_i,\Gamma_i\}$ $i\in[1,N]$ are parameters describing the joint distribution of X and Y. Using a training set of synchronized samples x_t and y_t concatenated into joint samples z_t, the Expectation-Maximization algorithm (EM) allows the estimation of λ.
Given two sequences of test $X = \{x_t,t \in [1, T]\}$ and $Y = \{y_t,t \in [1, T]\}$, a measure of their correspondence $C_\lambda(X,Y)$ can be computed as in (9).

$$C_\lambda(X, Y) = \frac{1}{T} \sum_{t=1}^{T} p([x_t, y_t]|\lambda) \tag{9}$$

Then the application of a threshold θ decides on whether the acoustic speech X and the visual speech Y correspond to each other (if $C_\lambda(X,Y) > \theta$) or not (if $C_\lambda(X,Y) \leq \theta$).

Remark λ is well known to be speaker-dependent: GMM-based systems are the state-of-the-art for speaker identification. However, there is often not enough training samples from a speaker S to correctly estimate the model λ_S using the EM algorithm. Therefore, one can adapt a world model λ_Ω (estimated on a large set of training samples from a population as large as possible) using the few samples available from speaker S into a model λ_S. This is not the purpose of this paper to review adaptation techniques: the reader can refer to [25] for more information.

Hidden Markov Models Like the Pearson's coefficient and the mutual information, time offset between acoustic and visual speech features is not modeled using GMMs. Therefore, the authors of [22] propose to model audio-visual speech with Hidden Markov Models (HMMs). Two speech recognizers are trained: one classical audio only recognizer [24], and an audiovisual speech recognizer as described in [23]. Given a sequence of audiovisual samples ($[x_t,y_t]$, $t \in [1, T]$), the audio only system gives a word hypothesis W. Then, using the HMM of the audiovisual system, what the authors call a measure of plausibility P(X,Y) is computed as follows:

$$P(X, Y) = p([x_1, y_1]...[x_T, y_T]|W) \tag{10}$$

An asynchronous hidden Markov model (AHMM) for audio-visual speech recognition is proposed in [2]. It assumes that there is always an audio observation x_t and sometimes a visual observation y_s at time t. It intrinsically models the difference of sample rates between audio and visual speech, by introducing the probability that the system emits the next visual observation y_s at time t. AHMM appears to outperform HMM in the task of audio-visual speech recognition [2] while naturally resolving the problem of different audio and visual sample rates.

4.3.2 Non-parametric models

The use of neural networks (NN) is investigated in [10]. Given a training set of both synchronized and not synchronized audio and visual speech features, a neural network with one hidden layer is trained to output 1 when the audiovisual input features are synchronized and 0 when they are not. Moreover, the authors propose to use an input layer at time t consisting of $[X_{t-N_X}, \ldots, X_t, \ldots, X_{t+N_X}]$ and $[Y_{t-N_Y}, \ldots, Y_t, \ldots, Y_{t+N_Y}]$ (instead of X_t and Y_t), choosing N_X and N_Y such as about 200 ms of temporal context is given as an input. This proposition is a way of solving the well-known problem of coarticulation and the already mentioned lag between audio and visual speech. It also removes the need for down-sampling audio features (or up-sampling visual features).

5 Applications

Measuring the synchrony between audio and visual speech features can be a great help in many applications dealing with audiovisual sequences.

Sound source localization Sound source localization is the most cited application of audio and visual speech correspondence measure. In [10], a sliding window performs a scan of the video, looking for the most probable mouth area corresponding to the audio track (using a Time-Delayed Neural Network). In [22], the principle of mutual information allows to choose which of the four faces appearing in the video is the source of the audio track: the authors announce a 82% accuracy (averaged on 1016 video tests). One can think of an intelligent video-conferencing system making extensive use of such results: the camera could zoom in on the person who is currently speaking.

Liveness test The main weakness of a biometric system based on talking-faces is that it might be fooled by an impostor showing a picture of the face of his/her target while playing a recording of his/her voice. Two relatively simple solutions can address this problem. First, one can ask the user to pronounce a random sentence, thus preventing the use of a pre-recorded sample of the voice of the target. The other solution is to check the correspondence between audio and visual speech: a few papers [13, 12, 7, 3] investigate this solution. Thus, replay attacks scenarios are tackled in [7] (with GMMs for joint audiovisual modeling) and [3] (with the Pearson's coefficient).

Indexation of audiovisual sequences Another field of interest is the indexation of audiovisual sequences. In [21], the authors combine scores from three systems (face detection, speech detection and a measure of correspondence based on the mutual information between the soundtrack and the value of pixels) to improve their algorithm for detection of monologue. Experiments performed in the framework of the TREC 2002 Video Retrieval Track [31] show a 50% relative improvement on the average precision.

Film post-production During the post-production of a film, dialogues are often re-recorded in a studio. An audio-visual speech correspondence measure can be of great help when synchronizing the new audio recording with the original video. Such measures can also be a way of evaluating the quality of a dubbed film into a foreign language: does the translation fit well with the original actor facial motions?

Other applications In [29], audio-visual speech correspondence is used as a way of improving an algorithm for speech separation. The authors of [15] design filters for noise reduction, with the help of audio-visual speech correspondence.

6 Conclusion and perspectives

This paper has reviewed techniques proposed in the literature to measure the degree of correspondence between audio and visual speech. However, it is very difficult to compare these methods since no common framework is shared among the laboratories working in this area. There was a monologue detection task (where using audiovisual speech correspondence showed to improve performance in [21]) in TRECVid 2002 but unfortunately it disappeared in the following sessions (2003 to 2006). Moreover, tests are often performed on very small datasets, sometimes only made of a couple of videos and difficult to reproduce. Therefore, drawing any conclusions about performance is not an easy task: the area covered in this review clearly lacks a common evaluation framework.

Nevertheless, experimental protocols and databases do exist for research in biometric authentication based on talking-faces: one can quote the XM2VTS and BANCA databases and corresponding protocols, for example. One could think of augmenting the existing biometric authentication tasks with audiovisual speech synchrony detection tasks.

A simple idea would be to artificially create desynchronized sequences from the original sequences. Thus, in [3], two replay attacks scenarios are introduced and experimented on the BANCA database. The first scenario, called "Paparazzi", simulates an attack where the impostor owns a picture of the face (which he/she puts in front of the camera) and an audio recording of the voice (which he/she plays in the microphone) of his/her target. In the second scenario, called "Big Brother", the impostor gains access to a video of the face and an audio recording of the voice of his/her target, but from two different utterances and therefore desynchronized. A simple algorithm based on the correlation between the audio energy and the value of grey-level pixels in the mouth area gives 0% Equal Error Rate (EER) on the first scenario and 35% EER on the second one in the replay attack detection task.

Acknowledgment

The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space.

References

[1] J. P. Barker and F. Berthommier. Evidence of Correlation between Acoustic and Visual Features of Speech. International Congress of Phonetic Sciences, 1999.

[2] S. Bengio. An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition. Advances in Neural Information Processing Systems, 2003.

[3] H. Bredin, A. Miguel, I. H. Witten, and G. Chollet. Detecting Replay Attacks in Audiovisual Identity Verification. In IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2006.

[4] C. Bregler and Y. Konig. "Eigenlips" for Robust Speech Recognition. International Conference on Acoustics, Speech, and Signal Processing, 2:19–22, 1994.

[5] CANCOR. http://people.imt.liu.se/~magnus/.

[6] T. Chen. Audiovisual Speech Processing: Lip Reading and Lip Synchronization. IEEE Signal Processing Magazine, pages 9–21, 2001.

[7] G. Chetty and M. Wagner. "Liveness" Verification in Audio-Video Authentication. Australian International Conference on Speech Science and Technology, pages 358–363, 2004.

[8] C. C. Chibelushi, F. Deravi, and J. S. Mason. A Review of Speech-Based Bimodal Recognition. IEEE Transactions on Multimedia, 4(1):23–37, 2002.

[9] C. C. Chibelushi, J. S. Mason, and F. Deravi. Integrated Person Identification Using Voice and Facial Features. IEE Colloquium on Image Processing for Security Applications, (4):1–5, 1997.

[10] R. Cutler and L. Davis. Look Who's Talking: Speaker Detection using Video and Audio Correlation. International Conference on Multimedia and Expo, pages 1589–1592, 2000.

[11] S. Dolédec and D. Chessel. Co-Inertia Analysis: an Alternative Method for Studying Species-Environment Relationships. Freshwater Biology, 31:277–294, 1994.

[12] N. Eveno and L. Besacier. A Speaker Independent Liveness Test for Audio-Video Biometrics. 9th European Conference on Speech Communication and Technology, 2005.

[13] N. Eveno and L. Besacier. Co-Inertia Analysis for ”Liveness” Test in Audio-Visual Biometrics. International Symposium on Image and Signal Processing Analysis, pages 257–261, 2005.

[14] J. W. Fisher and T. Darell. Speaker Association With Signal-Level Audiovisual Fusion. IEEE Transactions on Multimedia, 6(3):406–413, 2004.

[15] J. W. Fisher, T. Darrell, W. T. Freeman, and P. Viola. Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. Advances in Neural Information Processing Systems, 2001.

[16] N. Fox and R. B. Reilly. Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features.International Conference on Audio- and Video-Based Biometric Person Authentication, pages 743–751, 2003.

[17] R. Goecke and B. Millar. Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English. International Conference on Audio-Visual Speech Processing, 2003.

[18] J. Hershey and J. Movellan. Audio-Vision: Using Audio-Visual Synchrony to Locate Sounds. Neural Information Processing Systems, 1999.

[19] A. Hyvarinen. Survey on Independent Component Analysis. Neural Computing Surveys, 2:94–128, 1999.

[20] ICA. http://www.cis.hut.fi/projects/ica/fastica/.

[21] G. Iyengar, H. Nock, and C. Neti. Audio-Visual Synchrony for Detection of Monologues in Video Archives. International Conference on Acoustics, Speech, and Signal Processing, pages 329–332, 2003.

[22] H. Nock, G. Iyengar, and C. Neti. Assessing Face and Speech Consistency for Monologue Detection in Video. Multimedia'02, pages 303–306, 2002.

[23] G. Potamianos, C. Neti, J. Luettin and I. Matthews. Audio-Visual Automatic Speech Recognition: An Overview. Chapter to appear in: Issues in Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds., MIT Press, 2004.

[24] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In IEEE, volume 77, pages 257–286, 1989.

[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. Digital Signal Processing, 10:19 – 41, 2000.

[26] M. Slaney and M. Covell. FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. Neural Information Processing Society, 13, 2000.

[27] P. Smaragdis and M. Casey. Audio/Visual Independent Components. International Symposium on Independent Component Analysis and Blind Signal Separation, pages 709–714, 2003.

[28] D. Sodoyer, C. Jutten, and J.-L. Schwartz. Speech Extraction based on ICA and Audio-Visual Coherence. International Symposium on Signal Processing and Its Applications, 2:65–68, 2003.

[29] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten. Separation of Audio-Visual Speech Sources: A New Approach Exploiting the Audio-Visual Coherence of Speech Stimuli. EURASIP Journal on Applied Signal Processing, 11:1165–1173, 2002.

[30] N. Sugamura and F. Itakura. Speech Analysis and Synthesis Methods developed at ECL in NTT–From LPC to LSP. Speech Communications, 5(2):199–215, June 1986.

[31] Text Retrieval Conference Video Track. http://trec.nist.gov/.

[32] M. Turk and A. Pentland. Eigenfaces for Recognition. Journal of Cognitive Neuroscience, 3(1):71 – 86, 1991.

[33] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative Association of Vocal-Tract and Facial Behavior. Speech Communication, (28):23–43, 1998.

[34] G. Saporta, Probabilités, Analyse des Données et Statistique. Editions Technip, 1990.