

Robust Audio-Visual Speech Synchrony Detection by Generalized Bimodal Linear Prediction

Kshitiz Kumar, Jiri Navratil, Etienne Marcheret, Vit Libal, Gerasimos Potamianos***

* Carnegie Mellon University, Pittsburgh, PA 15213, USA

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

** Institute of Informatics and Telecommunications, NCSR “Demokritos”, 15310 Athens, Greece

Contact Emails: kshitizk@ece.cmu.edu, jiri@us.ibm.com, gpotam@iit.demokritos.gr

Abstract

We study the problem of detecting audio-visual synchrony in video segments containing a speaker in frontal head pose. The problem holds a number of important applications, for example speech source localization, speech activity detection, speaker diarization, speech source separation, and biometric spoofing detection. In particular, we build on earlier work, extending our previously proposed time-evolution model of audio-visual features to include non-causal (future) feature information. This significantly improves robustness of the method to small time-alignment errors between the audio and visual streams, as demonstrated by our experiments. In addition, we compare the proposed model to two known literature approaches for audio-visual synchrony detection, namely mutual information and hypothesis testing, and we show that our method is superior to both.

Index Terms: Audio-Visual Synchronization, Mutual Information, Hypothesis Testing, Linear Prediction, Visual Features

1. Introduction

The problem of detecting audio-visual (AV) speech synchrony has recently attracted interest in the literature, due to its many potential applications [1]–[6]. Indeed, one can consider many scenarios where visual speech information extracted from video of the speaker lips does not correspond to the acoustic track in the particular video segment; for example, a camera may be focusing on a non-speaker, a subject may be translated (dubbed) to another language, or a static face may be provided together with a canned audio recording during a spoofing attack on a biometric access system. In addition, in general multi-speaker scenarios, AV synchrony detection can provide crucial information to help resolve spatio-temporal ambiguity of the “who spoke when and where” problems, namely in speech activity detection, speaker localization and diarization, and possibly assist in the difficult task of sound source separation.

Most approaches in the literature address the AV synchrony detection problem in terms of mutual information between audio and visual features [2]–[5]. Segments bearing high mutual information indicate that one of the audio or visual feature vectors provide some prediction about the other, hence the AV features can be considered in sync; else not. Additional criteria based on the correlation coefficient, parametric AV models and neural networks also appear in overview paper [5]. Recently, hypothesis testing methods have also been applied in AV synchrony detection, generalizing in a way the mutual information technique as discussed in Section 3. Both methods though treat neighboring AV feature frames as statistically independent. This contradicts the practical observation that such

AV features are strongly correlated, resulting from their generation/evolution in time.

In order to address this shortcoming, in earlier work [6], we proposed a time-evolution model for AV features, successfully capturing their correlation, and deriving an analytical way of quantifying AV synchrony based on the model parameters. In this paper, we extend that model to also include non-causal information – namely from future audio and/or visual observations. The proposed modifications result in significantly improved robustness to small time-alignment errors between the audio and visual streams, as demonstrated by our experiments on an appropriate bimodal corpus of both synchronous and asynchronous AV segments of a single speaker in frontal head pose. In addition, we compare our proposed approach to baseline AV speech synchrony detection systems that employ mutual information or hypothesis testing, with our results demonstrating our model to be superior to both.

The remainder of the paper is organized as follows: We review AV synchrony detection based on mutual information and hypothesis testing in Sections 2 and 3, respectively. We then present our proposed time-evolution based method in Section 4. Our experiments are described in Section 5, and the paper concludes with a summary in Section 6.

2. The Mutual Information Based Approach

As already discussed, the problem of AV synchrony detection has been primarily approached using the *mutual information* (MI) criterion [2]–[5]. In this technique, the MI between sets of audio and visual features is evaluated, with high MI values implying AV synchrony. Mathematically, MI is defined as

$$I(A; V) = \mathbb{E} \log \frac{p(a, v)}{p(a)p(v)}, \quad (1)$$

where $p(a)$, $p(v)$, and $p(a, v)$ are the probability distribution functions (pdfs) of audio (A), visual (V), and joint AV feature vectors, respectively, and \mathbb{E} denotes expectation. As reviewed in [6], (1) can be easily computed assuming a single Gaussian density for all its three pdfs, thus resulting to

$$I(A; V) = \frac{1}{2} \log \frac{|\Sigma_A| |\Sigma_V|}{|\Sigma_{AV}|}, \quad (2)$$

where Σ_A , Σ_V , and Σ_{AV} are the covariance matrices of audio, visual, and AV feature vectors, respectively, and $|\bullet|$ denotes matrix determinant. As it is clear from the above, this approach treats AV feature frames as independent to each other.

3. The Hypothesis Testing Based Approach

Hypothesis testing (HT) is a statistical approach for deciding between two hypotheses. It assumes a finite number of underlying classes, and models the features corresponding to those classes in terms of a parametric pdf, for example a Gaussian mixture model (GMM). A classification decision is then made about a test feature on basis of its log-likelihood score against the models. In its application to AV synchrony detection, we are interested in two classes, namely

$$\begin{aligned} \mathcal{H}_1 &- \text{AV features in sync} \\ \mathcal{H}_0 &- \text{AV features not in sync} \end{aligned} \quad (3)$$

Following training of these two classes (e.g., using GMMs), at test time we evaluate a log-likelihood ratio (LLR) for the test data against the two class models. To proceed with our derivations, let us respectively denote audio, visual, and AV feature vectors at time instant n by a_n , v_n and z_n , where $z_n = [a_n^T, v_n^T]^T$. Let Z denote a N length sequence of z_n features across time. A normalized LLR can then be computed on sequence Z as in (4), and compared against an appropriate threshold λ to provide a hypothesis for the underlying class as \mathcal{H}_1 or \mathcal{H}_0 :

$$LLR = \frac{1}{N} \log \frac{p(Z; \mathcal{H}_1)}{p(Z; \mathcal{H}_0)} \geq \lambda. \quad (4)$$

Assuming that frames z_i are independent, LLR simplifies to

$$LLR = \frac{1}{N} \sum_i \log \frac{p(z_i; \mathcal{H}_1)}{p(z_i; \mathcal{H}_0)} \approx \mathbb{E} \log \frac{p(z; \mathcal{H}_1)}{p(z; \mathcal{H}_0)}. \quad (5)$$

It is interesting to note, that one could rewrite (5) as

$$LLR \approx \mathbb{E} \log \frac{p(z; \mathcal{H}_1)}{p(a)p(v)} - \mathbb{E} \log \frac{p(z; \mathcal{H}_0)}{p(a)p(v)}. \quad (6)$$

By comparing (6) and (1), it becomes clear that the normalized LLR can be interpreted as a ‘‘two-sided’’ MI, where an MI-like score is conditionally evaluated for the two classes, the difference of which becomes the HT score. In general of course, one expects that HT will outperform the MI approach, as presented in the previous Section, due to the training phase employed and the use of GMMs vs. single Gaussian densities (note that GMMs are mathematically intractable in (1) – see also [7]). However, both HT and MI treat AV features as statistically independent. This shortcoming is addressed by our proposed modeling approach presented in the next Section.

4. Generalized Bimodal Linear Prediction

In our earlier work [6], we specifically proposed the use of *bimodal linear prediction coefficients* (BLPC) as a model for AV feature evolution in time, successfully capturing auto- and cross-correlation information of the two feature streams. In particular the following *causal* model was proposed,

$$a_n \approx \hat{a}_n = \sum_{i=1}^{N_a} \alpha[i] a_{n-i} + \sum_{j=0}^{N_v} \beta[j] v_{n-j}, \quad (7)$$

for *scalar* audio (a) and visual (v) features. A quantitative measure was then proposed to quantify AV synchrony based on the difference of model parameter values of α and β (of lengths N_a and $N_v + 1$ respectively) under the presence or absence of an asynchrony assumption (as also further detailed in this Section). The approach was then extended for multi-dimensional AV feature vectors by employing canonical correlation analysis (also further discussed in this Section).

4.1. The Generalized BLPC Model

In this paper, we extend BLPC model (7) to include non-causal information – namely from future AV observations. The extension is motivated from empirical knowledge that audio features remain correlated with such future AV information. Model (7) is therefore generalized to the following two *non-causal* models:

$$\hat{a}_n = \sum_{i=1}^{N_a} \alpha[i] a_{n-i} + \sum_{j=-N_v}^{N_v} \beta[j] v_{n-j}, \quad (8)$$

$$\hat{a}_n = \sum_{\substack{i=-N_a \\ i \neq 0}}^{N_a} \alpha[i] a_{n-i} + \sum_{j=-N_v}^{N_v} \beta[j] v_{n-j}, \quad (9)$$

where (8) adds dependence on future visual observations, with model (9) further introducing dependence on future audio features. For simplicity, we will refer to models (7), (8), and (9) as BLPC/BLPC-1, BLPC-2, and BLPC-3, respectively. As in the original BLPC model, the generalized BLPC models can be extended to multi-dimensional AV features by employing canonical correlation analysis.

The proposed model generalizations over BLPC-1 not only allow a richer AV feature representation, but also prove especially useful for AV synchrony detection in the presence of small misalignments between the audio and visual streams. Such misalignment is often caused by intensive audio-visual data acquisition, and can be successfully captured by the non-causality introduced in the generalized BLPC models, as demonstrated in our experiments.

4.2. Model Parameter Estimation and Synchrony Measure

We now proceed to the problem of estimating the BLPC model parameters and measuring AV asynchrony based on their values. Our derivations constitute a generalization of the corresponding equations presented in our earlier work [6], and are provided for the most general model of the three considered, namely (9). Parameter estimation for other models can be subsequently deduced. The problem is formulated as a typical minimum square error estimation one, with the desired parameters obtained by differentiating $E[a_n - \hat{a}_n]^2$ with respect to α and β , and setting the differential to zero. To proceed, we define

$$\begin{aligned} \Phi_{aa} &= \begin{bmatrix} \phi_{aa}[0] & \dots & \phi_{aa}[2N_a] \\ \vdots & \ddots & \vdots \\ \phi_{aa}[2N_a] & \dots & \phi_{aa}[0] \end{bmatrix} \\ \Phi_{vv} &\stackrel{v-a}{\leftarrow} \begin{matrix} N_a \\ N_a \end{matrix} \Phi_{aa} \\ \Phi_{av} &= \begin{bmatrix} \phi_{av}[-N_a + N_v] & \dots & \phi_{av}[-N_a - N_v] \\ \vdots & \ddots & \vdots \\ \phi_{av}[N_a + N_v] & \dots & \phi_{av}[N_a - N_v] \end{bmatrix} \quad (10) \\ P_{aa} &= [\phi_{aa}[N_a], \dots, \phi_{aa}[0], \dots, \phi_{aa}[N_a]]^T \\ P_{av} &= [\phi_{av}[N_v], \dots, \phi_{av}[0], \dots, \phi_{av}[-N_v]]^T \\ J &= I_{(2N_a) \times (2N_a+1)}, \end{aligned}$$

where Φ_{aa} is a matrix consisting of autocorrelation values of audio features at different time lags, matrix Φ_{vv} is obtained in parallel to Φ_{aa} , but for visual features, and matrix Φ_{av} consists of cross-correlation coefficients between AV features at different lags. Vectors P_{aa} and P_{av} consist of autocorrelation coefficients of audio features and cross-correlation coefficients of AV

features, respectively. Finally, J is a row-eliminating matrix of size $(2N_a) \times (2N_a + 1)$, which eliminates the $(N_a + 1)^{th}$ row on its operand. Thus, operation $J\Phi_{aa}$ results in the removal of the middle row in Φ_{aa} , and $J\Phi_{aa}J^T$ eliminates both the middle row and middle column in Φ_{aa} .

Using the definitions in (10), the final solution for parameters α and β can be compactly written as:

$$\begin{bmatrix} \alpha_\rho \\ \beta_\rho \end{bmatrix} = \begin{bmatrix} J\Phi_{aa}J^T & \rho \cdot J\Phi_{av} \\ \rho \cdot \Phi_{av}^T J^T & \Phi_{vv} \end{bmatrix}^{-1} \cdot \begin{bmatrix} JP_{aa} \\ \rho \cdot P_{av} \end{bmatrix}, \quad (11)$$

where for later convenience, we parametrized the solution by a variable ρ . For asynchronous AV features, we can safely assume that

$$\phi_{av}[n] = 0, \forall n, \quad (12)$$

namely that the cross-correlation coefficients for AV features are identically 0 for all possible time-lags, and hence that

$$\Phi_{av} = \mathbf{0} \quad \text{and} \quad P_{av} = \mathbf{0}. \quad (13)$$

This is equivalent to setting $\rho = 0$ in (11) and obtaining parameters $\{\alpha_0, \beta_0\}$ for asynchronous AV features. On the other hand, for synchronous features, no such assumption holds, hence resulting in parameters $\{\alpha_1, \beta_1\}$ ($\rho = 1$).

Following parameter estimation, we can now define a measure of AV synchrony. Based on a number of experiments reported in [6], such measure is defined as

$$D_\alpha = \|\alpha_0 - \alpha_1\|, \quad (14)$$

namely as the distance between prediction coefficients α using the asynchrony assumption (13) and the coefficients without using that assumption. Large values of D_α therefore indicate AV synchrony. Note that in (14), $\|\bullet\|$ denotes the L2 norm. It can be easily shown that the parameter solutions for models (7) and (8) can be deduced from (11) via different instantiations of the row-eliminating matrix J .

4.3. Canonical Correlation Analysis for Vector Feature BLPC

The BLPC approach presented above has been considered for scalar audio and visual features. In practice, one expects to obtain multi-dimensional feature vectors from the audio and visual streams. To extend our proposed method to such cases, one would have to investigate all possible audio and visual feature pairs in (7)–(9), exponentially increasing the number of models. To avoid this, we employ *canonical correlation analysis* (CCA) [8], similarly to our earlier work [6] (note that CCA has also been considered for AV synchrony detection in [1, 5]).

In particular, we apply CCA on the audio and visual feature vectors, and we collect the correlated audio and visual features resulting from the projection into distinct feature pairs. As a result of CCA, the scalar components of these pairs are correlated within but uncorrelated across pairs. We then employ the BLPC model of our choice to describe each AV feature pair, computing individual distances D_a . An overall distance is then obtained by summing up the distances over all pairs. Note that in addition to BLPC, we also employed CCA in conjunction with the HT baseline approach, due to the resulting computational savings.

5. Experiments

We now proceed to discuss our AV synchrony detection experiments. We first provide information on the bimodal database and experimental setup, followed by our results.

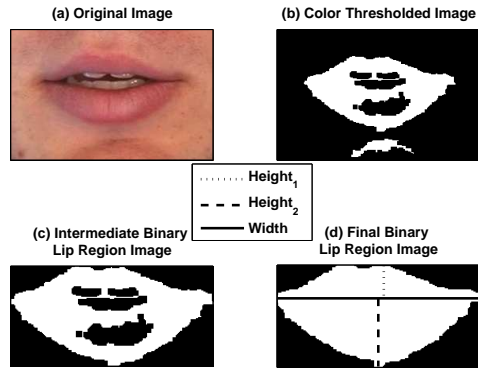


Figure 1: Visual feature extraction in our system [9]. 3 estimated visual features are shown in (d).

5.1. Experimental Framework

We conducted our experiments on an appropriate AV speech database, part of the “CMU Audio-Visual Profile Frontal Corpus” [9]. The data have been recorded in an anechoic room with a “head and shoulder” camera view of the speaker, and they contain isolated word utterances. In particular, we used the frontal only part of this corpus, consisting of just over one hour of data. To facilitate our experiments, we further split the data into chunks of four seconds each. These individual chunks constitute the synchronous segments of our corpus. We then randomly mixed different 4-sec chunks of the audio and video streams thus obtaining asynchronous segments of a total duration of approximately eight hours. Since HT requires a training stage, we reserved 60% of the data for training and the remaining for testing. In addition, for estimating CCA vectors, we reserved held-out synchronous data of 5 min in duration.

For audio-visual feature extraction, we employed our system reported in [9]. In particular, we used conventional 13-dimensional MFCC audio features [10] and extracted 3-dimensional “geometric” visual features from the frontal mouth images of the subject, that consisted of the upper and lower lip heights and the lip width [9] (see also Fig. 1). Due to their lower extraction rate, visual features were upsampled from 30 to 100 Hz, to match the audio feature rate. Furthermore, all features were mean and variance normalized. In addition, for the HT and BLPC methods, we applied CCA on the extracted features. This yielded 3-dimensional projected features that were collected into distinct AV feature pairs. For the MI criterion though, we used the unprojected features.

Some additional implementation details in our experiments are as follows: For the MI based approach, we assumed a multivariate underlying Gaussian density and evaluated MI as in (2). For the HT based system, we trained 512-mixture Gaussian models using the EM algorithm [10] for both classes of interest in (3). Finally, for BLPC, we used $N_a = N_v = 6$ in (7)–(9).

5.2. Results

In our first set of experiments, we compared MI, HT, and BLPC-1 for AV synchrony detection. Results are reported in Fig. 2, in terms of a *detection error tradeoff* (DET) curve. It is clear from the figure that HT performs significantly better than the MI criterion, and that in almost all DET curve regions the BLPC approach is by far the best. In particular, BLPC provides respec-

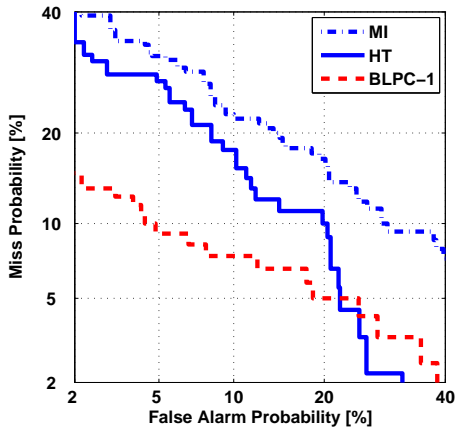


Figure 2: AV synchrony DET curve for the MI and HT baselines, as well as the causal BLPC approach.

tively 50% and 33% relative reduction in *equal error rate* (ERR) over MI and HT methods.

Next, we consider the two non-causal models introduced in (8) and (9). We are particularly interested in the robustness of these models to small misalignments of otherwise synchronous audio and visual streams, due possibly to capturing delays because of intensive data rates. We thus introduced an artificial relative delay in the AV stream and used such data during testing. Results are reported in Fig. 3 in terms of EER. Note that we considered both positive and negative delays, and averaged the results. We clearly observe that BLPC not only provides better EER, but it also remains more robust to AV stream misalignments than the MI and HT methods. Interestingly, HT performs better than MI in the absence of misalignment, but its performance degrades rapidly. Furthermore, among the two generalized BLPC models, BLPC-2 proved superior, indicating that including future audio information does not contribute to robustness on top of the already incorporated future visual information. Finally, with respect to computational efficiency, the BLPC method and its generalizations are quite efficient and can be implemented on devices with smaller computational power.

6. Summary

In conclusion, we investigated the problem of audio-visual synchrony detection in video segments containing a speaker in frontal head pose but possibly speech that does not belong to this subject. We presented two baseline approaches to the problem, namely based on the mutual information criterion and hypothesis testing, highlighting that both ignored information from audio-visual feature correlation at small time-lags. In our approach, we specifically addressed this shortcoming, by discussing a time-evolution model for audio-visual features in the form of linear prediction that we recently introduced. We then generalized the model to a non-causal one by including future visual or audio-visual features, and we derived parameter estimation formulas and an AV synchrony measure for it. We demonstrated that the introduced models outperform the baseline, and in particular that one of the models remains robust to small misalignments of the audio and visual streams.

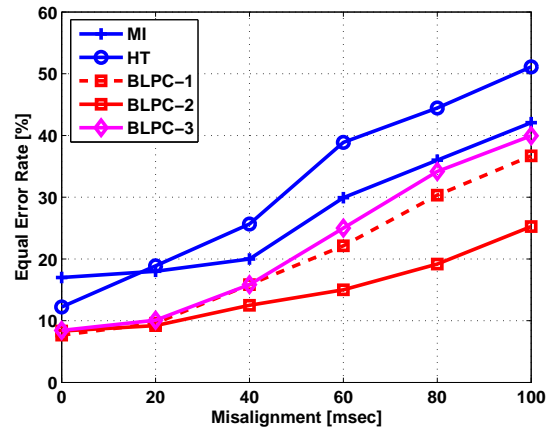


Figure 3: Robustness to misalignment in AV streams of the various AV synchrony detection algorithms considered in this paper.

7. References

- [1] M. Slaney and M. Covell, “Facesync: a linear operator for measuring synchronization of video facial images and audio tracks,” in *Advances Neural Information Process. Sys.*, vol. 13, pp. 814–820, MIT Press, 2000.
- [2] J. Hershey and J. Movellan, “Using audio-visual synchrony to locate sounds,” in *Advances Neural Information Process. Sys.*, vol. 12, pp. 813–819, MIT Press, 1999.
- [3] G. Iyengar, H. Nock, and C. Neti, “Audio-visual synchrony for detection of monologues in video archives,” *Proc. Int. Conf. Multimedia Expo*, vol. 1, pp. 329–332, 2003.
- [4] H. Nock, G. Iyengar, and C. Neti, “Speaker localisation using audio-visual synchrony: An empirical study,” *Proc. ACM Int. Conf. Multimedia*, pp. 488–499, 2003.
- [5] H. Bredin and G. Chollet, “Audiovisual speech synchrony measure: Application to biometrics,” *EURASIP J. Advances Signal Process.*, 2007.
- [6] K. Kumar, J. Navratil, E. Marcheret, V. Libal, G. Ramaswamy, and G. Potamianos, “Audio-visual speech synchronization detection using a bimodal linear prediction model,” (To appear:) *Proc. CVPR Biometrics Works.*, 2009.
- [7] J. Hershey and P. Olsen, “Approximating the Kullback Leibler divergence between Gaussian mixture models,” *Proc. Int. Conf. Acoustics Speech Signal Process.*, vol. 4, pp. 317–320, 2007.
- [8] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, *Canonical Correlation Analysis - An Overview with Application to Learning Methods*, Royal Holloway, University of London, CSD-TR-03-02, 2003.
- [9] K. Kumar, T. Chen, and R. Stern, “Profile view lip reading,” *Proc. Int. Conf. Acoustics Speech Signal Process.*, vol. 4, pp. 429–432, 2007.
- [10] The Sphinx Open Source Speech Recognition Engines, [Online] Available at: <http://cmusphinx.sourceforge.net/html/cmusphinx.php>