Audio-Visual Speech Synchronization Detection Using a Bimodal Linear Prediction Model

Kshitiz Kumar
Carnegie Mellon University, Pittsburgh, PA 15213, USA
kshitizk@ece.cmu.edu

Jiri Navratil, Etienne Marcheret, Vit Libal, Ganesh Ramaswamy
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
{jiri,etiennem}@us.ibm.com

Gerasimos Potamianos
Institute of Informatics and Telecommunications, NCSR "Demokritos", 15310 Athens, Greece
gpotam@iit.demokritos.gr

Abstract

In this work, we study the problem of detecting audio-visual (AV) synchronization in video segments containing a speaker in frontal head pose. The problem holds important applications in biometrics, for example spoofing detection, and it constitutes an important step in AV segmentation necessary for deriving AV fingerprints in multimodal speaker recognition. To attack the problem, we propose a time-evolution model for AV features and derive an analytical approach to capture the notion of synchronization between them. We report results on an appropriate AV database, using two types of visual features extracted from the speaker's facial area: geometric ones and features based on the discrete cosine image transform. Our results demonstrate that the proposed approach provides substantially better AV synchrony detection over a baseline method that employs mutual information, with the geometric visual features outperforming the image transform ones.

Index Terms– *Audio-Visual Synchronization, Mutual Information, Linear Prediction, Visual Features*

1. Introduction

Audio-visual (AV) synchronization detection bears importance in many biometrics related applications, for example in detecting spoofing and in automatically generating AV fingerprints of different individuals in AV data. Given pre-segmented AV data, AV synchronization can identify segments where audio and visual signals are in sync, implying that these segments most likely belong to the same individual. In typical AV data, we may observe static faces in images, or the camera focusing on a non-speaker, or a subject speaking in a foreign language with audio being translated to another language. In all of the above examples, there exists a mismatch between the audio and visual speech sources. AV synchrony indicates consistency between the audio and visual streams and thus the reliability for the segments to belong to the same individual. Such segments could then serve as building blocks for generating audio-visual fingerprints of the different individuals present in the AV data, which can be important for security, authentication, and biometric purposes. AV segmentation can also be important for speaker turn detection, as well as automatic indexing and retrieval of different occurrences of a speaker.

The problem of AV synchrony detection has already been considered in the literature. We refer to [2] for a comprehensive review on this topic. There, the authors present detailed discussion on different aspects of AV synchronization detection, including feature processing, dimensionality reduction, and correspondence measures. In that paper, AV synchrony detection is applied to the problem of identity verification, but the authors also mention additional applications in sound source localization, AV sequence indexing, film post-production and speech separation.

Most of the past work in AV synchronization detection has been based on mutual information (MI) between audio and visual features [4, 6, 2]. Segments bearing high MI indicate that observing one of audio or visual features provides some prediction about the other, hence the AV streams can be considered in sync; else not. Additional criteria based on the correlation coefficient, parametric AV models, and neural networks appear in [2]. One of the key assumptions in

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

the MI based approach is that AV feature frames are statistically independent. This approach thus extracts little information from the generation or evolution perspective of AV features, from which we empirically obtain that neighboring AV feature frames should be strongly correlated.

In our work, we specifically propose a time-evolution model for AV features and derive an analytical method to capture the notion of synchronicity between them. Throughout this work, we provide useful insights for the parameters in our model. We also extend the notion of correlation coefficient and relate our model with linear prediction coefficients (LPC) used in speech processing.

The rest of this paper is organized as follows: We review the mutual information based approach for AV synchrony detection in Section 2. We present our proposed time-evolution model in Section 3, accompanied by the use of canonical correlation analysis, as discussed in Section 4. We then detail our experimental setup and results in Sections 5 and 6, respectively. We present discussion and future work in Section 7, with Section 8 concluding the paper.

2. Mutual Information Based Audio-Visual Synchrony Detection

The problem of AV synchronization detection has primarily been approached using the mutual information (MI) criterion [4, 8, 2]. Under this criterion, mutual information is evaluated between sets of audio and visual features, and high MI values imply synchronicity. The criterion is mathematically defined as

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

where $I(A;V)$ denotes the MI between the audio (A) and visual (V) features, $p(a)$, $p(v)$, and $p(a,v)$ indicate the probability distribution functions (pdfs) of audio, visual and joint audio-visual feature vectors, respectively, and $\mathbb{E}$ denotes expectation.

To compute (1), we adopt a parametric approach, assume an underlying density function for the probability distributions in (1), and estimate the parameters of that density function. A convenient choice for such pdf is a single Gaussian distribution. Researchers in [4, 8] also make the same density assumption for the task of speaker localization. Under it, the MI score becomes

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

where Σ_A, Σ_V, and Σ_{AV} are the covariance matrices of audio, visual, and joint audio-visual feature vectors, respectively, and $|\bullet|$ denotes matrix determinant.

A potential problem in the above formulation is that it assumes the underlying AV feature frames to be independent. This contradicts empirical observations that AV features evolve as a time-series, and that AV feature frames at small time-lags are highly correlated. Capturing such correlations motivates our proposed approach, presented in the next Section, where we model the time-evolution of AV features. There, we specifically study linear dependence among the features evolving across time to analytically parametrize the notion of synchrony between AV features.

3. Time Evolution Based Bimodal Linear Prediction

In this Section, we develop our algorithm for AV synchrony detection, based on what we term *bimodal linear prediction coefficients* (BLPC). We specifically develop a time-evolution model for AV features that highlights the presence of correlation at small time-lags across feature frames. Our model caters to the two cases of synchronous vs. asynchronous AV features. It is of course difficult to analytically define synchronicity, but in our model we approximate it in terms of linear dependence among features. We assume that synchronous AV features should bear a linear dependence and be linearly explainable. From our model, we later derive an analytic way to parametrize synchronicity between AV features.

We thus propose a time-evolution model to capture linear dependence among AV features as

$$a[n] \approx \hat{a}[n] = \sum_{i=1}^{N_a}\alpha[i]a[n-i] + \sum_{j=0}^{N_v}\beta[j]v[n-j]\ . \tag{3}$$

In the above, $a[n]$ and $v[n]$ indicate audio and visual features, respectively, at discrete time-instant n. For now, we assume that AV features consist of a single audio and a single visual element, jointly referring to them as an AV feature pair. We later extend our model for multiple audio and visual features, in Section 4. In model (3), we assume that the current audio feature, $a[n]$, can be linearly explained or predicted using past N_a audio features, as well as the present visual feature $v[n]$ and past N_v visual features. The parameters involved in our model are coefficients α and β of lengths N_a and $N_v + 1$, respectively. Next, we note that parameters β will bear high values for AV features in sync. This holds because in our model we approximated synchronicity with correlation. Thus, synchronous visual features will be correlated with audio ones and can linearly explain some of them. Similarly, we claim that parameters β will ideally equal 0 for asynchronous AV features.

So far, in (3), we provided a time-evolution model of AV features, with parameters α and β encoding information useful for AV synchrony detection. Our next task is to estimate these parameters from observed AV features. For this purpose, we formulate a minimum square error estima-

tion problem and seek to minimize

$$E[a[n] - \hat{a}[n]]^2 \ . \tag{4}$$

To minimize (4), we differentiate it with respect to α and β, and we obtain the desired parameters by setting the differentials to zero. To proceed, we define

$$\begin{aligned}
\Phi_{aa} &= \begin{bmatrix} \phi_{aa}[0] & \dots & \phi_{aa}[N_a - 1] \\ \vdots & \ddots & \vdots \\ \phi_{aa}[N_a - 1] & \dots & \phi_{aa}[0] \end{bmatrix} \\
\Phi_{vv} &\overset{N_v \overset{v \leftarrow a}{+1 \Leftarrow} N_a}{\Longleftarrow} \Phi_{aa} \\
\Phi_{av} &= \begin{bmatrix} \phi_{av}[1] & \dots & \phi_{av}[1 - N_v] \\ \vdots & \ddots & \vdots \\ \phi_{av}[N_a] & \dots & \phi_{av}[N_a - N_v] \end{bmatrix} \\
P_{aa} &= [\phi_{aa}[1], \dots, \phi_{aa}[N_a]]^T \\
P_{av} &= [\phi_{av}[0], \dots, \phi_{av}[-N_v]^T \ ,
\end{aligned} \tag{5}$$

where Φ_{aa} is a Toeplitz 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 time-lags. Finally, vectors P_{aa} and P_{av} consist of autocorrelation coefficients of audio features and cross-correlation coefficients of AV features, respectively.

Using the definitions in (5), the final solution for the parameters can be compactly written as

$$\begin{bmatrix} \alpha_\rho \\ \beta_\rho \end{bmatrix} = \begin{bmatrix} \Phi_{aa} & \rho.\Phi_{av} \\ \rho.\Phi_{av}^T & \Phi_{vv} \end{bmatrix}^{-1} \cdot \begin{bmatrix} P_{aa} \\ \rho.P_{av} \end{bmatrix} \ , \tag{6}$$

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

$$\phi_{av}[n] = 0, \ \forall \, n \ , \tag{7}$$

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

$$\Phi_{av} = \mathbf{0} \quad \text{and} \quad P_{av} = \mathbf{0} \ . \tag{8}$$

This is equivalent to setting $\rho = 0$ in (6) and obtaining parameters $\{\alpha_0, \beta_0\}$ for asynchronous AV features. On the other hand, for synchronous features, no such assumption holds, hence the solution in (4) results in parameters $\{\alpha_1, \beta_1\}$ ($\rho = 1$).

Next, in (9), we define a measure of closeness for the two types of prediction coefficients, with and without the asynchrony assumption of (8), as

$$D = \left\| \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} - \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \right\| \ , \tag{9}$$

where $\| \bullet \|$ denotes the L2 norm. One can expect that D will remain small for asynchronous AV features and large for synchronous ones, thus providing a measure of AV synchronicity. Similarly to (9), we can derive two additional figures of merit, individually for each α and β, namely

$$D_\alpha = \|\alpha_0 - \alpha_1\| \quad \text{and} \quad D_\beta = \|\beta_0 - \beta_1\| \ . \tag{10}$$

Parameter D_α quantifies the distance between coefficients α_1 and α_0. One could view α_0 as corresponding to the well studied LPC-type coefficients [9] for audio, while α_1 corresponds to LPC coefficients when some of audio features can be linearly explained by β_1 coefficients in (3). Thus, D_α captures the change in LPC coefficients. We can also provide a similar understanding of D_β , as an indication of the change in parameters β , when visual features can explain audio ones in (3).

It is worth mentioning that, in parallel to (3), one could also try to predict visual features from past AV features as

$$v[n] \approx \hat{v}[n] = \sum_{j=1}^{N_v} \beta[j] v[n-j] + \sum_{i=0}^{N_a} \alpha[i] a[n-i] \ . \tag{11}$$

It is instructive to explore the similarities and dissimilarities in (3) and (11). Relevant experiments are reported in Section 6.

In summary, in this Section, we provided a time-evolution based approach to estimate bimodal linear prediction coefficients for an AV feature pair consisting of a single audio and a single visual element. In practice, one expects to obtain multi-dimensional feature vectors of different dimensionalities from the audio and visual streams. To extend our proposed method to such cases, we will need to design a feature transformation to establish a few only appropriate AV feature pairs to be modeled by (3). This is discussed next.

4. Audio-Visual Feature Pair Selection by Canonical Correlation Analysis

As already mentioned above, our model in (3) (as well as in (11)) considers scalar audio and visual features. To cover multi-dimensional such features, one would have to investigate all possible audio and visual feature pairs in (3), exponentially increasing the number of models. To avoid this, we seek an appropriate feature transformation and in particular a projection for feature dimensionality reduction, such that the resulting audio and visual features can be collected into distinct AV pairs, with their scalar components correlated within but uncorrelated across pairs. It will then suffice to only consider a small number of time-evolution models of the resulting distinct AV feature pairs.

For this purpose, we employ *canonical correlation analysis* (CCA). We briefly overview the formulation, referring

to [3] for details. The objective is to derive projection vectors $\{P, Q\}$ to transform features in a and v to respectively a' and v', namely

$$a' = P^T a \,, \;\; v' = Q^T v \,, \tag{12}$$

such as to maximize the correlation between $\{a', v'\}$. As a result of CCA, the following hold:

$$\mathbb{E}\, a' \, (a')^T = \mathrm{I} \,, \;\; \mathbb{E}\, v' \, (v')^T = \mathrm{I} \,, \;\; \mathbb{E}\, a' \, (v')^T = \mathfrak{D} \,, \tag{13}$$

where $\mathfrak{D}$ denotes a diagonal matrix. Thus, we note that the projected audio features are correlated with only one of the projected visual features. Further, projected audio and visual features are respectively uncorrelated with other projected audio and visual features.

In our proposed approach, 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. We then employ model (3) (or (11)) to describe each AV feature pair, and we compute distance D. An overall distance is then obtained by summing up the distances over all pairs. Note that researchers in [10, 2] have applied CCA in AV synchrony detection as well.

5. Experimental Setup

We now proceed to our experiments on AV synchronization using the methods presented so far. In these, we employ two AV systems that differ in their visual feature extraction module, as explained later in this Section. We perform experiments on an appropriate AV speech database, part of the "CMU Audio-Visual Profile Frontal Corpus" [7]. The data have been recorded in an anechoic room with a "head and shoulder" camera view of the speaker and 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, and obtained asynchronous segments by randomly mixing different 4 sec chunks of the audio and video streams. The total duration of the resulting asynchronous database was about eight hours. The original 4 sec chunks constitute the synchronous part of our data. Note that none of the methods discussed in this work requires any specific training, with the exception of CCA. Projection vectors for the latter are estimated from a held-out synchronous data segment of 5 min in duration.

Next, we describe the features that we extracted from the audio and visual streams in our two AV systems. For audio, and for both systems, we used conventional 13-dimensional MFCC features [1]. However, the two systems differed in their visual features. The first, referred to as "System-1", employed 3-dimensional lip geometric (shape-based) features, consisting of the upper and lower lip heights and the lip width, as also depicted in Fig. 1. The second system, referred to as "System-2", employed appearance-based visual features, namely 40-dimensional discrete cosine transform (DCT) coefficients of the mouth region image pixels. Details of both feature extraction approaches are given in [7]. Note that the visual feature rate was 30 Hz, whereas the audio features were extracted at 100 Hz. We therefore upsampled the former to match the audio rate. Furthermore, we applied mean and variance normalization to all features.

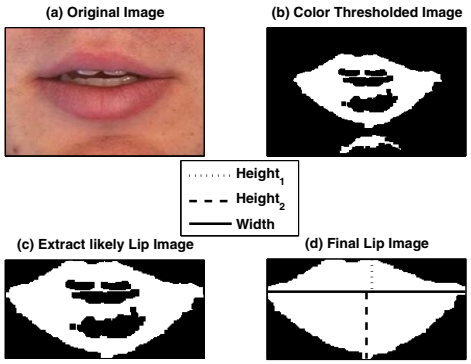

Figure 1. Visual feature extraction in "System-1" (see also [7]).

We applied CCA on the features and projected them to a lower-dimensional space. Since the maximum number of projections in CCA is limited to the minimum of the original feature vector dimensionalities, we projected audio and visual features each to 3-dimensional spaces in "System-1" and 13-dimensional ones in "System-2". As we noted in Section 4, the projected audio and visual features were collected into distinct AV feature pairs. We estimated distances in (9) or (10) for each pair and summed across all pairs to obtain an overall distance that became our metric for deciding AV feature synchrony. Note that for the MI criterion we used unprojected features, assuming multivariate Gaussians as the underlying audio and visual feature pdfs in (2).

6. Results

In this Section, we present AV synchrony detection results using the experimental setup of Section 5. In particular, we compare the two visual feature extraction systems considered ("System-1" and "System-2"), BLPC modeling approaches (3) and (11), figures of merit (9) and (10), as well as the baseline MI approach and our proposed BLPC. AV synchrony results are typically depicted in terms of *equal error rate* (EER), and in the final plot (Fig. 6) in terms of the *detection error tradeoff* (DET) curve. In particular, when reporting BLPC results, we assume $N_a = N_v$ in (3) or (11). It is also worth mentioning, that in our BLPC experiments, we observed that although P_{av} terms are small for asynchronous data-chunks, they are not iden-

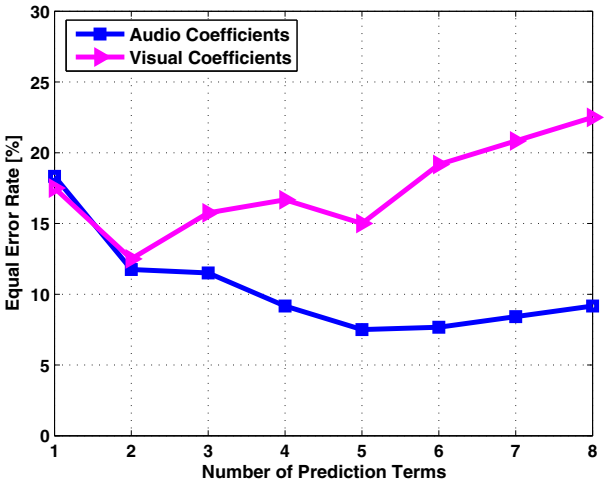

Figure 2. "System-1": Comparison of EER for models predicting audio features and figure of merit being based on norm of audio and visual prediction coefficients in (10).

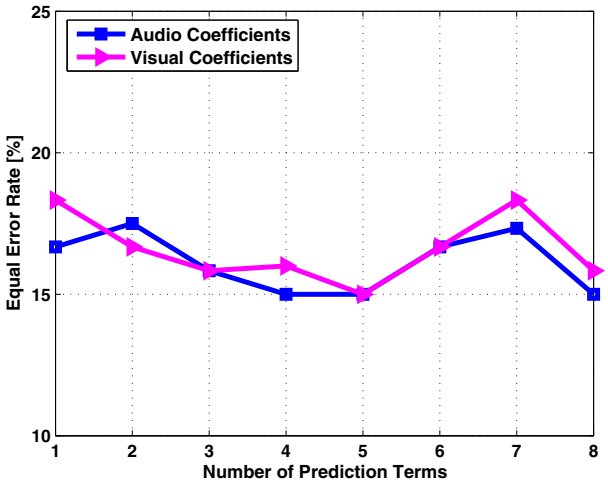

Figure 3. "System-1": Comparison of EER for models predicting visual features and figure of merit being based on norm of audio and visual prediction coefficients in (10).

tically zero. Consequently, we experimented with replacing ρP_{av} by P_{av} in (6). That resulted in better performance. We therefore consistently used this modification in all our BLPC experiments.

We first consider the time-evolution model in (3) for "System-1", and we present our results for the two individual figures of merit in (10). As depicted in Fig. 2, the figure of merit based on audio features, i.e. parameters α in (10), performs much better than the one based on visual features, i.e. parameters β. We therefore infer a fundamental insight that the audio features in themselves bear a very distinct signature in terms of their linear prediction, and that there is a significant change in their signature, when a part of audio feature can be explained by visual features. We note that the best performance is obtained when $N_a = N_v = 5 \sim 6$. We expect that there should be a middle range for the N_a parameter, when the algorithm should perform the best. Smaller values of N_a will not quite capture the autocorrelation existing in the audio features at different time-lags, whereas larger values of N_a far exceed the range of time-lags, where audio features bear reasonable autocorrelation to be useful to the BLPC model. In Fig. 2, we observe that the visual based coefficients do not perform as well, hence indicating that even though for synchronous AV features visual features explain some of audio features, the corresponding β parameters do not change a lot. We did an experiment where we combined the two individual figures of merit in a linearly weighted scale and obtained improvements over the figure of merit based on just the audio coefficients in (10). These improvements however were very small (less than 0.5% absolute), indicating that the audio based coefficients remain sufficient for our task.

Next, we present results using the time-evolution model (11) for "System-1". In model (11), β_0 indicates LPC coefficients of visual features. It is instructive to compare AV synchronization using the figure of merit employing change in audio LPC vs. change in visual LPC. Comparing Fig. 2 and Fig. 3, we note that audio coefficients in Fig. 2 perform much better than visual coefficients in Fig. 3. We infer that audio features in general bear a much better linear prediction representation than visual features. In Fig. 3 we also note that with respect to model (11), the audio based coefficients do better than visual based coefficients. This again justifies that audio features have a better signature with respect to linear prediction. In any case, one should emphasize that the interpretations developed in this Section about AV features are specific to the extracted features in Section 5.

In "System-2" we derived DCT features from images, whereas in "System-1", we worked on geometric features. We compare the performance of visual features in Fig. 2 and Fig. 4. Results in both figures are based on model (3). Comparing the results, we note that DCT features are not as good as geometric features for AV synchrony detection. It seems that 3-dimensional visual geometric features of lip-heights and lip-width carry more distinct information about synchrony to audio than the 40-dimensional DCT coefficients. Here too we observe that best performance is obtained when $N_a = N_v = 5 \sim 6$.

Fig. 5 plots the results for "System-2", where we used the time evolution model for visual features in (11). As in "System-1", audio feature based coefficients perform better than visual based coefficients. Visual coefficients perform better in model (11) than in (3), but their error rates are

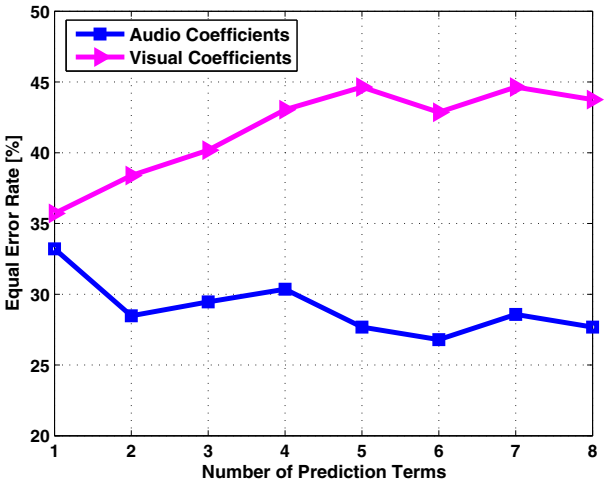

Figure 4. "System-2": Comparison of EER for models predicting audio features and figure of merit being based on norm of audio and visual prediction coefficients in (10).

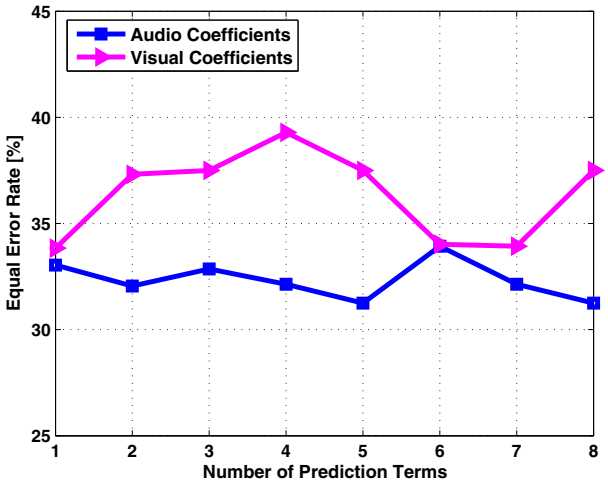

Figure 5. "System-2": Comparison of EER for models predicting visual features and figure of merit being based on norm of audio and visual prediction coefficients in (10).

still high. Other interpretations that we developed for AV features in "System-1" carry over to "System-2" as well.

Next, we plot the DET curve in Fig. 6 for the two broad methods we discussed in our work, namely MI and BLPC. For BLPC, we use "System-1" with time evolution model (3), as this system performed the best in our task. For MI we used a single Gaussian density assumption for the features. We see that BLPC performs significantly better than the MI method and improves EER absolutely by 10%. Omitting details, we compared BLPC to our best implementation of another state-of-the-art method, namely *hypothesis testing* (HT) [2]. There, BLPC provided 27% relative reduction in EER over HT. The key difference between our method and MI is that MI extracts little information from the time-evolution perspective in the features, which suggests that features in a small time-window should bear high correlation among themselves. In MI, the features are assumed to be independent, but in our approach we specifically capture the correlation across the features for different time-lags in parameters α and β. For MI results, we relaxed the single Gaussian density assumption to Gaussian mixture models (GMM) using the approximations in [5], but single density still provided better detection results. This suggests that the approximation methods for evaluating MI scores for GMMs are still inadequate for our task, and hence better approximations are needed.

7. Discussion

In this Section, we discuss and relate the two methods presented in our paper. From [4], we know that the MI criterion for AV features under the single Gaussian density assumption is equivalent to cross-correlation at 0^{th} lag between audio and visual features. Next, for the BLPC model in (3), we can derive that with $N_a = 0$ and $N_v = 0$, β becomes the cross-correlation at 0^{th} lag between the AV features. Thus, our measure in (9) will be equivalent to the measure in MI in (2). Hence, our work essentially provides a way to extend the MI based approach. MI approaches gather autocorrelation statistics and cross-correlation statistics at just 0^{th} lag between the features but our approach gathers those statistics at additional time-lags in an attempt to explain the time evolution model in (3) and capture the notion of synchrony between AV features.

With respect to model (3), we can further improve our AV synchrony detection results by exploiting correlations due to past as well as future feature frames. In this case, for the current audio feature, the index for visual features will range from $-N_v$ to N_v, and similarly for audio features. This generalization could also account for possible misalignment between the two streams (e.g. due to the acquisition hardware). Note also that the time-evolution model in (3) currently considers speech information in a context-independent fashion. One should expect that autocorrelation and cross-correlations among AV features differ among various phonetic classes and contexts, and such variation could be useful in the BLPC model. We intend to address these issues in our next work.

8. Conclusions

In conclusion, we note that we worked on the problem of reliable audio-visual synchrony detection, where the task was to decide if a particular AV segment belonged to the same person. We motivated different applications for this work, specifically in automatically generating AV finger-

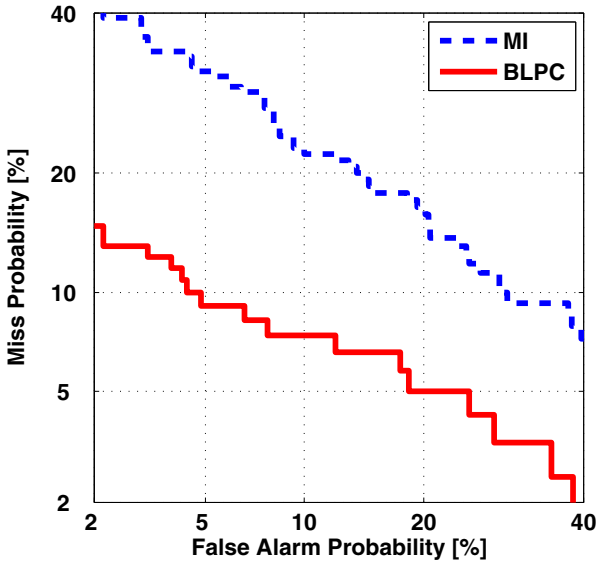

Figure 6. DET plot for reliable AV segmentation using mutual information and bimodal linear prediction model.

prints for individuals in AV data. We initially presented the mutual information criterion for this problem, as a baseline, and we highlighted that the approach ignored information from audio-visual feature correlation at small time-lags. In our approach, we specifically proposed a time-evolution model for audio-visual features in the form of linear prediction. We highlighted that the model parameters capture the notion of audio-visual synchrony, and we derived a measure for synchronicity detection based on these parameters. We also justified the use of canonical correlation analysis in our work as a means to extend our approach to multi-dimensional audio-visual features. Throughout our work, we provided useful analysis and discussion on our model and the parameters involved, also relating our approach to the mutual information criterion, and indicating that our method extends its. We applied our proposed method on an appropriate audio-visual database considering two visual feature extraction approaches, and obtained significant improvements over the mutual information based baseline.

References

[1] Sphinx open source speech recognition engines. http://cmusphinx.sourceforge.net/html/cmusphinx.php. 4

[2] H. Bredin and G. Chollet. Audiovisual speech synchrony measure: Application to biometrics. *EURASIP Journal on Advances in Signal Processing*, 2007. 1, 2, 4, 6

[3] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. *Canonical Correlation Analysis - An Overview with Application to Learning Methods*. Royal Holloway, University of London, 2003. CSD-TR-03-02. 4

[4] J. Hershey and J. Movellan. Using audio-visual synchrony to locate sounds. In *Advances in Neural Information Processing Systems 12*, pages 813–819. MIT Press, 1999. 1, 2, 6

[5] J. Hershey and P. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 317–320, 2007. 6

[6] G. Iyengar, H. Nock, and C. Neti. Audio-visual synchrony for detection of monologues in video archives. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 329–332, 2003. 1

[7] K. Kumar, T. Chen, and R. Stern. Profile view lip reading. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 429–432, 2007. 4

[8] H. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony: An empirical study. In *Proceedings of the 10th ACM International Conference on Multimedia*, pages 488–499, 2003. 2

[9] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall Inc., 1993. 3

[10] M. Slaney and M. Covell. Facesync: a linear operator for measuring synchronization of video facial images and audio tracks. In *Advances in Neural Information Processing Systems 13*, pages 814–820. MIT Press, 2000. 4