Overcoming Asynchrony in Audio-Visual Speech Recognition

Virginia Estellers, Jean-Philippe Thiran

Signal Processing Laboratory 5 (LTS5)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
virginia.estellers@epfl.ch, jp.thiran@epfl.ch

***Abstract*—In this paper we propose two alternatives to overcome the natural asynchrony of modalities in Audio-Visual Speech Recognition. We first investigate the use of asynchronous statistical models based on Dynamic Bayesian Networks with different levels of asynchrony. We show that audio-visual models should consider asynchrony within word boundaries and not at phoneme level. The second approach to the problem includes an additional processing of the features before being used for recognition. The proposed technique aligns the temporal evolution of the audio and video streams in terms of a speech-recognition system and enables the use of simpler statistical models for classification. On both cases we report experiments with the CUAVE database, showing the improvements obtained with the proposed asynchronous model and feature processing technique compared to traditional systems.**

I. Introduction

Visual information can improve the performance of audio speech recognition systems, specially in presence of noise. The improvement is due to the complementary nature of the audio and visual modalities, as the visual information can help discern sounds easily confusable by ear but distinguishable by eye. The design of such an audio-visual system requires a fusion strategy for the different modalities, which constitutes a key topic on multimodal signal processing [1]. The fusion techniques should consider the asynchrony between the audio and visual modalities intrinsic to human speech where, for instance, the movement of the lips precedes or follows the actual production of sound at beginning or end of utterances. That asynchrony is a complex issue, as it is not a constant time delay between audio and video signals but changes with time, is context-dependent due to co-articulation effects and depends on the visibility and asynchrony of visible articulatory features as lips, teeth and tongue [2].

The statistical models commonly used in Audio-Visual Speech Recognition (AVSR) are multistream Hidden Markov Models, the natural extension of the Hidden Markov Models (HMM) used in audio speech recognition. However, several works [3], [4], [5] have proved the benefits of the more general Dynamic Bayesian Network (DBN) models allowing asynchrony between the audio and the visual streams. In that case, DBNs need to define some synchronization points for the models. These works reported that DBN word models imposing synchrony at word boundaries outperformed HMM-based ones. However, working with word models is restricted to small vocabulary tasks and the extension of the same asynchrononous DBNs to phoneme models did not obtain better results than phoneme multistream HMMs [6].

In this paper, we show that the use of DBNs is also beneficial for phoneme models when the synchonization constraints are correctly defined. We establish that the correct way to treat asynchrony in audio-visual speech recognition is within word boundaries and propose a new DBN phoneme model able to exploit its asynchrony without being limited to small vocabulary tasks. Our model thus enjoys the benefits of the DBNs documented in [3], [4], [5] overcoming the problems encountered by Graviet et al when extending their use to phonemes and large vocabulary tasks [6].

Analyzing recognition experiments with the standard CUAVE database [7], where our model outperforms the existing DBN ones, we develop an alternative strategy to overcome audio-visual asynchrony while working with a simple multistream HMM. The proposed technique introduces an extra processing step on the extracted audio and visual features in order to reduce the complexity of the subsequent statistical models. Such an approach is interesting as extensive work has been conducted on stream weighting for multistream HMMs, while it is not a so well-studied issue in DBNs.

The paper is organized in several sections as follows: the different (a)synchronous statistical models for audio-visual speech recognition are presented in section II and a new DBN phoneme model proposed. Based on those models, in section III we develop a processing technique designed to overcome the asynchrony of the feature streams while working with synchronous models. Section IV reports experiments comparing the proposed model and processing techniques to the state-of-the art and conclusions are drawn in section V.

II. Audio-Visual models in Speech Recognition

Audio-Visual Speech Recognition systems estimate the probability of a word or sentence by building statistical models of basic speech units given the observed audio and visual features o_A, o_V. On real-world tasks, those speech units are based on phonemes, as it is unfeasible to learn specific models for every word in the vocabulary. Instead, words are expressed in terms of phonemes, models for each phoneme are

MMSP'10, October 4-6, 2010, Saint-Malo, France.

learned and then whole-word models are created concatenating phoneme models. We focus, therefore, on the use of phonemes for speech recognition.

Multistream HMMs are the traditional tools used to model those phonemes [8], but they are examples of a more general statistical model, Dynamic Bayesian Networks, recently also used for speech recognition [5], [9]. In the following section we present (a)synchronous models used in AVSR in the context of DBNs, compare them in terms of synchrony constraints and propose a new one.

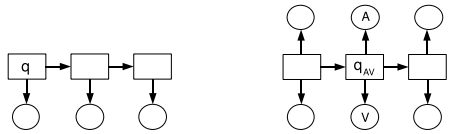
A. Dynamic Bayesian Networks

Bayesian Networks are directed acyclic graphs representing dependencies between variables in a probabilistic model. Variables are depicted as nodes in the network while arcs represent a conditional probability relation between the nodes they connect. A directed arc from node A to node B implies that B (called descendent or child node) is conditionally dependant on A (parent node). In a Bayesian Network each node is conditionally independent from its non-descendent given its parents. Dynamic Bayesian Networks are their natural extension when the variables are stochastic processes and we have Bayesian Networks in space and time. A HMM can thus be represented as a DBN, see figure 1a[1], where the hidden state is the parent of the observation variable and the state transition probabilities are encoded in the directed arcs between the state variables. DBNs, however, allow for a more general graph structure and flexible models than HMMs.

In AVSR the model parameters to estimate are the transition probabilities between audio and visual states $\alpha_{i,j} = p(q(t+1) = q_j | q(t) = q_i))$ and probability distributions of the audio and visual observations for each state $p(o|q_i)$. Taking into account the use of left-to-right models in speech, the number of parameters to estimate is then $O(N)$, where N corresponds to the number of states in our models. Maximum likelihood estimates of those parameters are computed with different implementations of the generalized Expectation-Maximization (EM) algorithm. In the case of HMMs, the Baum-Welch and Viterbi algorithms are used for parameter estimation and recognition [8], with respective time complexities $O(N^2T)$ and $O(NT)$, where T denotes the number of samples in the training or testing utterance. As parameter estimation is done once to build the system while recognition is constantly used, in the following we will only refer to the time complexity associated to recognition stage. For DBNs in general, implementations of the generalized EM are available for speech recognition, but their time complexity increases considerably [10], [9].

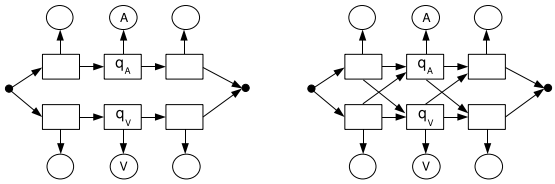
Our models assume stream independence for the combined observation likelihood $p(o_A, o_V|q_j) = p(o_A|q_j)p(o_V|q_j)$ and fit a Gaussian mixture to the audio and visual observations associated to each state. The complexity and estimation of those Gaussian mixtures is the same for all the models we will present, which will differ on the transition probability parameters associated to different synchrony constraints.

[1]On the DBN schemas, we will represent hidden state nodes *q* as rectangles, observation nodes as circles and smaller black circles as synchronization points of the models. The label *A,V,AV* on states and observations indicates their modality Audio, Video or Audio-Visual



(a) HMM (b) MSHMM

Fig. 1: DBN structure of HMMs



(a) IHMM (b) CHMM

Fig. 2: Asynchronous phoneme models

More complex stream fusion strategies can also be applied, but we do not adopt them because their effects on training and testing of AVSR systems have only been well-studied on MSHMM and not on more complex DBN models.

B. Synchronous Audio-Visual model

A multistream HMM (MSHMM), see figure 1b, assumes that the audio and video sequences are independent but state synchronous. The transition probabilities $\alpha_{i,j}^{AV}$ needs to be estimated for each of the N states of the P phonemes in our vocabulary, leading to a complexity $\mathcal{O}(NP)$ in terms of number of parameters and $\mathcal{O}(TNP)$ in recognition time for Viterbi.

C. Asynchronous Audio-Visual models

The assumption that the streams come from synchronous sources of information is valid when modelling information from the same modality, but that is not the case in AVSR systems. We can think of a HMM-like model representing each hidden state of the MSHMM as a pair of audio and visual states, allowing state asynchrony within the phoneme and forcing synchrony at model boundaries. Usual models of that kind include the independent HMM (IHMM), the product HMM, the coupled HMM (CHMM) and the factorial HMM. A good overview of those existing models and their complexity can be found in [5].

In this paper, we focus on IHMM and CHMM, which model independently the state observations of each stream but make different assumptions about their state evolution. IIHMMs assume independent state transitions of each stream while CHMMs allow the audio and visual states to interact with respect to their time evolution. Their DBN representation is presented in figure 2a and 2b, where we can see that both models allow more flexibility than the MSHMM in terms of synchrony. The IHMM, however, fails to model any correlation between the evolution of the audio and visual

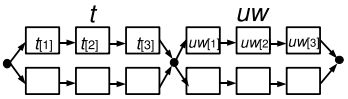


Fig. 3: Word 'two' built from concatenation phoneme IHMMs

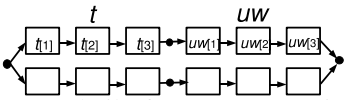


Fig. 4: Word 'two' built from concatenation of phoneme pHMMs

components while in the CHMM the coupling of states across streams accounts for this correlation. Actually those models can be build as non -eft-to-right MSHMMs with N^2 states $q_{i,j}^{AV} = (q_i^A, q_j^V)$ properly tying the Gaussian mixtures of the composed states

$$p(o_A, o_V)|q_{i,j}^{AV}) = p(o_A)|q_i^A)p(o_V)|q_j^V)$$

In IHMMs we have to factorize the transition probabilities as $\alpha^{AV}_{(i,j),(k,l)} = \alpha^A_{i,k}\alpha^V_{j,l}$ while in the CHMM we must consider a particular transition matrix structure in the N^2-state MSHMM [5]. That implementation leads respective parameter complexities of $\mathcal{O}(2NP)$ and $\mathcal{O}(N^2P)$ for the IHMM and CHMM and $\mathcal{O}(NPT)$ for the recognition time in both cases[5]. As usually $N = 3$ for phonemes is smaller than the number of words in the vocabulary, the complexity of a system based in IHMMs and CHMMs are similar between each other and to a MSHMM one.

D. Proposed asynchronous model

In the phoneme models previously presented, the synchronization points are the phoneme boundaries, but it is not necessarily the case for phoneme models based on DBNs. Indeed, we know that the delay between modalities can reach 120 ms [11], which surpasses the mean duration of phonemes. We thus propose to use two independent HMMs for each phoneme forcing synchronism at word boundaries once the constituting phonemes of the word are concatenated. We call that model piece-wise HMM (pHMM). Figure 3 and 4 show the DBN representation of words with IHMMs and pHMMs. It is important to note that those models are not word IHMMs [3], [2], [4], whose grouped states correspond to phonemes. Such a system would be defining word-depending phoneme models, which can only be used in reduced vocabulary tasks. We propose the use of independent phoneme models for the audio and visual domains, glued together for audio-visual recognition and with synchrony constraints imposed considering the words they form. The pHMM allows more flexibility in terms of asynchronism and is robust to co-articulation effects without building context-dependent phoneme or word models. When a certain phoneme presents different levels of asynchrony depending on the word or neighboring phonemes, our model does not impose any learnt asynchrony model for that word or context, but decides on testing between the possible words and asynchrony patterns. Those models can not be implemented as MSHMMs and their parameters must be estimated with the generalized EM-algorithm. The number of parameters to estimate is of order $\mathcal{O}(2NP)$ as they have independent transition probabilities for the audio and video streams and $\mathcal{O}(N^2VT)$ for the time complexity of the recognition stage, where V is the number of words in our vocabulary [12]. We see that the major flexibility of that model does not increase the number of parameters to train, but the complexity of recognition stage is considerably higher as the size of the vocabulary surpasses that of phonemes in real applications and large vocabulary tasks.

Unlike the biphone models presented in [13], pHMMs have no restrictions on the degree of asynchrony within a word or are context dependent. This last approach increases dramatically the amount of necessary training data and is thus restricted to few audio-visual databases. Such is an important limitation for training of the system, as the most extensive audio-visual databases consists on recordings of natural meetings and their visual modality is usually too challenging for training due to mouth tracking issues, changing illuminations and speaker poses [14].

III. Proposed processing technique

In audio speech recognition it is well know that simpler statistical models can be used for recognition if the observed features are additionally processed, for instance, to include information of the subsequent recognition classes in tandem HMM approaches [15]. In that section we present such an alternative for audio-visual speech recognition, preprocessing the features that will be used as visual observations in subsequent MSHMM classifiers. The visual features are warped to the temporal evolution of the audio mimicking the behaviour of asynchronous DBN models compared to HMMs.

A. Preliminary idea based on model analysis

Previous works [3], [4], [5] and our own experiments, see section IV, show the benefits of allowing audio-visual asynchrony within word boundaries. Analyzing why those asynchronous DBN models work better than the traditional MSHMMs, we develop a processing technique to overcome asynchrony by an additional processing step on visual features when MSHMMs are used for recognition.

Figures 5 and 7 show an schema of word models based on MSHMMs and DBNs successfully applied to overcome asynchrony.[2] Assuming that there is a word utterance between the time instants t_0 and t_1, the DBN and the MSHMM model differ only on the synchronization points: the MSHMM forces the state variable of both streams to be the same at each time instant, $q_A(t) = q_V(t)\ \ t_0 \leq t \leq t_1$, while the DBN model just imposes synchrony on the initial and last frame of the words t_0 and t_1.

In speech recognition, the temporal evolution of the observed features $o_A(t)$ and $o_V(t)$ is described by the evolution

[2]Note that we can also work with phoneme MSHMMs and consider the word MSHMM obtained from the concatenation of their constituting phonemes.

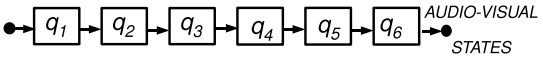


Fig. 5: MSHMM model for a word

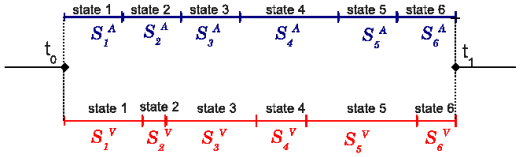


Fig. 6: Time partition of a word interval by their DBN modelling

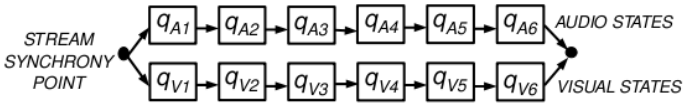


Fig. 7: DBN asynchronous word model

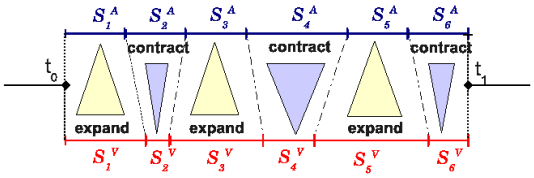


Fig. 8: Piece-wise definition of f

of the state variable in the corresponding statistical models. Those models define audio $\mathcal{S}^A = \{S_1^A, \ldots, S_n^A\}$ and video $\mathcal{S}^V = \{S_1^V, \ldots, S_n^V\}$ partitions of the time interval (t_0, t_1), where S_i^A corresponds to the time subinterval in which the audio state of models was q_i^A. Figure 6 shows a graphical representation of those time partitions of (t_0, t_1). For the MSHMM $\mathcal{S}^A$ and $\mathcal{S}^V$ coincide while the asynchronous DBN model allows the audio and video partition to differ within word boundaries.

The graphical representation of the problem suggests how to obtain a system performing equally to the DBN, but working with the simpler MSHMM and processed visual features $\tilde{o}_V$. The processed feature streams should have the same state evolution, i.e. $q_A(t) = \tilde{q}_V(t)$ but still be as close as possible to the original stream o_V. We could modify the temporal variable of the video stream $\tilde{o}_V(t) = o_V(f(t))$, with f defined piece-wisely expanding and contracting each S_i^V to fit S_i^A, so that $q_A(t) = \tilde{q}_V(t)$. See figure 8. That reasoning assumes that both audio and video HMMs have decided on a word with the same number of state, which is not always the case. Indeed, with single-stream HMMs we can just have the state partition $\mathcal{S}^A$ and $\mathcal{S}^V$ of each modality individually, where the start and end of the detected words and the number of states of the word recognized by each stream might be different. In those cases we could not expand and contract the partitioned intervals in a one-to-one basis. We can, however, take the audio modality as model for the time evolution of the features within word boundaries and adapt the visual features to it. Note that the audio modality is taken as model because it is the most reliable stream for recognition of speech, specially for silences. In the following, we denote $\mathcal{S}^A$ the time partition obtained with an audio-only system applied to the audio stream.

The time clustering of the observed features $o_A(t)$ and $o_V(t)$ explains the definition of the time partitions $\mathcal{S}^A$ and $\mathcal{S}^V$. Within any S_i^A the audio is in the same model state q_i^A, whose observations in our HMMs and DBNs are modeled with the same Gaussian mixture. On the contrary, samples indexed by S_{i-1}^A and S_i^A are modelled with different Gaussian mixtures and belong to different clusters on the feature space. Indeed, when performing recognition, a sample $o(t)$ will be assigned to state q_i or q_{i+1} depending mainly[3] on the likelihoods $p(o(t)|q_{i+1})$ and $p(o(t)|q_{i+1})$.

B. Proposed pre-processing of features

For each interval (t_0, t_1) where a word is recognized, our goal is to adapt the time variations of o_V so that its state partition coincides with $\mathcal{S}^A$ while keeping the feature values close to the originals, that is, we construct a new visual feature vector $\tilde{o}_V$ from o_V samples taking into account the clustering in time of the audio stream. For each S_i^A, we define the $\tilde{o}_V$ video features associated to it taking the necessary samples of the original o_V corresponding to time instants from the central region of S_i^A and indexing them as if they came from a uniform time sampling on S_i^A. We force in this way the time clustering of the new stream, and consequently its state partition, to be closer to the audio one. That is explained by the fact that we do not have a real analogue signal $o_V(t)$, but periodical samples from some visual feature from which we can construct a continuous version of $o_V(t)$. When we say that we take samples associated to time instants central to S_i^A, as the video rate is kept, in fact we are interpolating $o_V(t)$ in S_i^A in order to increase the number of samples associated to the center of the subinterval and neglect the rest. As the interpolation is associated to an up-sampling and low-pass filtering process, when the obtained samples from o_V are treated as coming from a uniform sampling of $\tilde{o}_V$ in S_i^A, the resulting feature stream presents a cluster in the central part of each subinterval S_i^A.

The procedure that we propose is the following: we use an audio-only HMM system to obtain the state partition of the time intervals $\mathcal{S}^A$, we up-sample o_V and form a new $\tilde{o}_V$ from the necessary samples associated to time instants close to the center of each S_i^A. Assuming that for the subinterval S_i^A we have N_i periodic samples of o_V, we can obtain the corresponding analog signal $o_V(t)$ by any smoothing interpolation

[3]The state transition probability of the HMM and the surrounding frame likelihoods also affect the decisions taken on the Viterbi decoder because of the forward path limitation of the HMMs, but as those transition probabilities are more or less similar for all the states and the surrounding frames at certain time instant are the same, the decision of assigning a sample to certain state or the following depends mainly on the observation likelihoods of the states

method we prefer. We then sample that signal at the sorted time instants drawn from a normal distribution with mean the center of S_i^A and variance with associated 90% confidence interval of the gaussian in relation $\frac{1}{n}$ to the lenght of S_i^A. In our implementation, we used cubic spline interpolation, values $n = 1, 2, 4, 8$ and restricted the N_i sampling instants to lie within the interval S_i^A. The resulting samples of $o_V(t)$ are treated as if they came from a periodic sampling of the cotinuous signal $\tilde{o}_V(t)$, which in fact does not need to be constructed.

The choice of the parameter σ in relation to the length of each time subinterval is done based on the results obtained with an evaluation set.

IV. Experiments and Results

We perform continuous speechreading experiments on the CUAVE database. We use the static portion of the 'isolated digits' section of the database, consisting of 36 speakers repeating the digits five times. Our experiments are speaker independent, using 6-fold cross validation with 30 speakers for training, 3 for evaluation and 3 for testing. The results are given in terms of word accuracy.

The audio features used are normalized mel-frequency cepstral coefficients with their first and second temporal derivatives. We train any model parameters on clean audio data and artificially add white noise on testing with Signal to Noise Ratios (SNR) ranging from clean to 0 dB. The visual features are selected DCT coefficients on a region of interest defined around the mouth, which consists of a 128x128 image of the speaker's mouth, normalized for size, centred and rotated. The DCT coefficients are the 15 most important ones taken in a zig-zag order, as in the MPEG/JPEG standard, together with first and second temporal derivatives and their means removed. No noise is added to the visual features.

A. Experiments on modeling

For all the experiments, the phoneme models are made of 3 hidden states with independent audio and visual observations described by Gaussian mixtures with diagonal covariance matrices. As the Expectation Maximization is an optimization algorithm finding local minima, which makes the choice of the initial parameters a critical issue, we obtained initial estimates of the parameters by separate training of audio and video-only HMMs before jointly retraining the audio-visual models. We used the GMTK toolkit [10] to build and train audio-visual models for the MSHMMs, IHMMs, CHMMs and pHMMs already described.

The results obtained are presented in table I, where the video-only word accuracy is 62.22%. They show that the asynchrony goes beyond phonemes and that allowing asynchronous phoneme models does not outperform the traditional MSHMM properly trained under the different SNR conditions. The CHMM only obtains better recognition performance than the MSHMM in low-noise conditions, whereas the proposed pHMM outperforms it through all the SNRs. In relation to the CHMM and IHMM, we see that the state evolution of the streams are not independent, as the coupled system exploits the temporal correlation of the audio and visual streams to obtain better performance.

TABLE I: Word accuracy of different asynchronous models

SNR	audio HMM	MSHMM	IHMM	CHMM	pHMM
clean	97.4	97.9	97.9	98.0	97.9
25 dB	97.3	97.8	97.8	98.0	97.8
20 dB	96.8	97.6	97.4	97.8	97.7
15 dB	94.2	95.6	95.4	95.4	96.5
10 dB	87.9	91.7	91.2	91.9	93.0
5 dB	74.2	82.2	80.3	81.9	86.1
0 dB	48.9	62.2	54.7	58.6	71.4

In order to analyze if the improvement obtained with the proposed model is significant and coherent through the different train-test sets and SNRs, we performed a Wilcoxon signed rank test comparing the results of each asynchronous model with the synchronous MSHMM. The null hypothesis being that the corresponding asynchronous model outperforms the MSHMM, we obtained p-values of 0.03, 0.40 and 0.97 for IHMM, CHMM and pHMM, respectively. We we can thus only state that introducing asynchrony on the models is beneficial for the pHMM through all SNRs levels, which can not be stated with the other DBN phoneme models. These results prove that audio-visual asynchrony in speech goes beyond phoneme level.

B. Experiments on feature processing

To test the alternative proposed processing technique, we have applied it to the same AVSR framework, using the audio stream and MSHMM speech recognizers already explained but with two different video streams. For the first stream we used directly the extracted visual features and for the second stream we applied the proposed processing technique with different values of the parameter σ. In experiments with the evaluation set the best performance was obtained adjusting σ to fit the 90% confidence interval of the Gaussian to half of the length of each S_i^A, that same value was retained for testing.

We present results with both a weighted and non-weighted MSHMMs. The first ones allow us to compare the proposed modeling and feature processing approaches to overcome asynchrony, while a weighted MSHMM is the current state-of-the-art in AVSR. The weighted and non-weithed results, however, can not be directly compared, as they were obtained with different software toolkits, GMTK for the non-weighted systems and HTK [16]for the weighted one, as the former does not allow for the use of different audio and video weights. In the weighted MSHMM, the fusion strategy is based on weighing the likelihoods of the audio and visual observations for each state $p(o_A, o_V|M) = p(o_A|M)^{\lambda_A} p(o_V|M)^{\lambda_V}$. Separately training the audio and visual models with clean audio data, we assume independency of the audio and visual observations and set both weights to one. However, testing the system with different SNR audio data, we choose the best weights (in terms of the performance on the evaluation set) for each SNR from the possible combinations satisfying

TABLE II: Word accuracy for the original and processed feature streams

Accuracy	non-weighted			weighted		
SNR	A	AV	A$\tilde{V}$	A	AV	A$\tilde{V}$
clean	97.4	97.9	97.9	98.5	98.4	99.1
25 dB	97.3	97.8	97.9	97.0	97.0	98.3
20 dB	96.8	97.6	97.7	96.2	94.4	98.1
15 dB	94.2	95.6	96.0	94.2	90.3	95.3
10 dB	87.9	91.7	92.2	88.5	82.5	92.1
5 dB	74.2	82.2	83.0	73.9	74.7	83.2
0 dB	48.9	62.2	63.7	49.97	65.6	69.8

$\lambda_A + \lambda_V = 1$ and ranging from 0 to 1 at 0.05 steps.

The results are shown in table II compared to an audio-only HMM system. We observe that the proposed technique, denoted $A\tilde{V}$, outperform both the traditional audio-visual system AV and the audio-only one A. Compared to the modelling approach, the processing technique does not perform as well as the proposed asynchronous model, even though it obtains better results than the asynchronous IHMM and CHMM phoneme models. For the weighted strategy implemented in HTK, the AV system performs worse than the audio-only system for high SNR conditions. This degradation is due to the state synchrony assumption of the MSHMM, which is too constraining, but also to the mismatch between the weights at training and testing conditions. Such a mismatch is unavoidable if we do not want to train a system for each possible SNR condition. Training the MSHMM those weights affects the association of the training example to one state or another of the MSHMM depending on the clustering of the audio and visual features. Setting $\lambda_A > \lambda_V$ favours the definition of audio-visual states according to the clustering of the audio stream and viceversa for $\lambda_V > \lambda_A$. It is natural, then, that the mismatch on the weights does not affect so much the modified stream, where audio and video clusters are similar. In the experiments, the A$\tilde{V}$ system manages to profit from the visual modality through all SNR levels, specially in noisy conditions. In that case we also performed a Wilcoxon signed rank hypothesis test for the A$\tilde{V}$ outperforming the AV system and obtained a p-value of 0.89, showing that the results are also statistically significant.

V. Conclusions

Our work proves that audio-visual asynchronism in speech recognition goes beyond the phoneme level and that word boundaries constitute a good choice for stream synchronization models. The proposed pHMM model exploits the DBN possibilities defining phoneme models with word synchrony and offers a good trade-off between vocabulary size and the amount of training data needed, not provided by asynchronous word or context-dependent models.

We also show that it is possible to reduce the audio-visual asynchrony effects by an additional processing of the feature streams while working with synchronous MSHMM models. The proposed technique aligns the visual stream to the temporal evolution of the audio in terms of the speech-recognition system. Such an approach is interesting because it enables the use of MSHMMs, for which more weighting and fusion strategies have been developed.

References

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004, ch. Audio-visual automatic speech recognition: An overview.

[2] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, and et al, "Articulatory feature-based methods for acoustic and audio-visual speech recognition," *ICASSP Proceedings*, 2007.

[3] S. Dupont and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition," *IEEE Transactions on Multimedia*, 2000.

[4] J. N. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "DBN based multi-stream models for audio-visual speech recognition," *ICASSP Proceedings*, 2004.

[5] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, 2002.

[6] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," *Conference on Human Language Technology*, 2002.

[7] Patterson, Gurbuz, Tufekci, and Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *ICASSP Proceedings*, 2002.

[8] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Readings in speech recognition*, 1990.

[9] G. Zweig and S. Russell, "Speech recognition with dynamic Bayesian networks," in *AAAI Proceedings*, 1998.

[10] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," in *ICASSP Proceedings*, 2002.

[11] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *ICASSP Proceedings*, 1994.

[12] K. Murphy, "Dynamic Bayesian networks: representation, inference and learning," Ph.D. dissertation, University of California, 2002.

[13] K. Kumatani, S. Nakamura, and K. Shikano, "An adaptive integration based on product HMM for audio-visual speech recognition," in *ICME Proceedings*, 2001.

[14] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[15] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMMsystems," in *ICASSP Proceedings*, 2000.

[16] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," *Cambridge University Press*, 1995.