# AN ASYNCHRONOUS DBN FOR AUDIO-VISUAL SPEECH RECOGNITION

*Kate Saenko and Karen Livescu*

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, USA, 02139

## ABSTRACT

We investigate an asynchronous two-stream dynamic Bayesian network-based model for audio-visual speech recognition. The model allows the audio and visual streams to de-synchronize within the boundaries of each word. The probability of de-synchronization by a given number of states is learned during training. This type of asynchrony has been previously used for pronunciation modeling and for visual speech recognition (lipreading); however, this is its first application to audio-visual speech recognition. We evaluate the model on an audio-visual corpus of English digits (CUAVE) with different levels of added acoustic noise, and compare it to several baselines. The asynchronous model outperforms audio-only and synchronous audio-visual baselines. We also compare models with different degrees of allowed asynchrony and find that the lowest error rate on this task is achieved when the audio and visual streams are allowed to desynchronize by up to two states.

***Index Terms***— Speech recognition

## 1. INTRODUCTION

Automatic speech recognition (ASR) has become a key component in efforts to produce more natural human-computer interfaces. Most modern ASR systems use a hidden Markov model (HMM) representation with a single audio signal as their input. Such systems are highly accurate in controlled environments, but their performance degrades rapidly in the presence of noise [14]. This has led researchers to incorporate visual information into ASR systems. Several types of audio-visual speech recognition (AVSR) models using both modalities have been proposed, including those dealing with large vocabulary, continuous speech [14].

One of the goals of AVSR research is to find effective ways of combining video with existing audio-only ASR systems [11]. Several models have been proposed for fusing audio and visual speech input using two parallel HMMs for the two inputs. Most of these models are special cases of dynamic Bayesian networks (DBNs) [10]. A subset of AVSR models allows the audio and visual state streams to de-synchronize to achieve further performance gains. In this paper, we propose and evaluate an asynchronous two-stream DBN for audio-visual speech recognition. This model allows the audio and visual streams to de-synchronize by up to several states within each word. The probability of de-synchronization by a particular number of states is learned during training. This type of asynchrony has been previously used for pronunciation modeling [8, 9] and for lipreading [15]; however, this is its first application to audio-visual speech recognition.

## 2. RELATED WORK

Since this paper focuses on the audio-visual fusion aspect of AVSR, we will not discuss other related topics such as lip tracking or feature extraction. For a comprehensive review of AVSR research we refer the reader to [14].

Several multi-stream models have been developed to take advantage of complementary sources of speech information. The streams can be multi-modal (e.g. audio-visual [7]), different types of features extracted from only the audio [12] or only the video [15], or a mix of both [6]. These models can be thought of as instances of the more general class of DBNs [16]. Here we focus on two-stream models, with one stream emitting audio and the other visual observations.

Asynchronous multi-stream models allow the hidden state streams to de-synchronize, in order to capture the natural asynchrony between sensory modalities. We can think of a continuum of different degrees of coupling: At the one extreme, the streams are completely independent; at the other extreme, they are fully synchronized. Most models, including the Factorial HMM [5] and the Coupled HMM [3], fall somewhere in between. It has been shown (e.g. [7]) that allowing a limited degree of asynchrony is beneficial for AVSR. However, models differ in their implementation of synchrony constraints. For example, the CHMM couples the hidden streams by conditioning the current state of each stream on both the previous state in that stream and the previous state in the other stream(s). In this paper, we propose a different type of synchrony constraint than has been used previously for AVSR. Instead of introducing direct dependencies between each state in a given stream and the states in the other stream(s), we

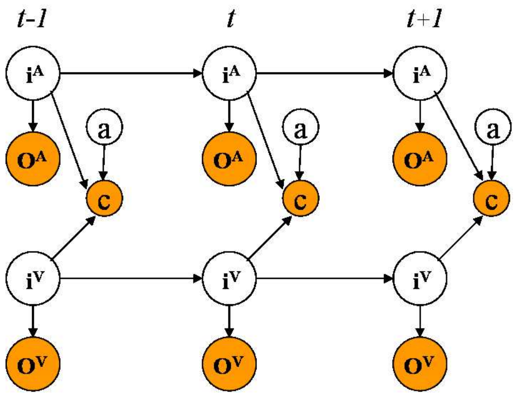

**Fig. 1**. *Asynchronous DBN for AVSR. $t$ is time, and $i_t^F$ is an index into the state sequence of stream $F$, where $F \in \{A, V\}$. Shaded nodes correspond to observed variables. Additional variables associated with synchronization at word boundaries have been omitted from the figure for brevity.*

model the overall probability of the state indices differing by some number of states, regardless of the word or the particular state in the word. This type of asynchrony in a multistream DBN was first proposed for an articulatory feature-based pronunciation model in [8]. The model was also extended to visual-only speech recognition (lipreading) from multiple visual articulatory feature streams in [15]. In this paper, we apply this asynchronous DBN model to audio-visual speech fusion and evaluate its performance on a database of English digits.

## 3. MODEL DESCRIPTION

The proposed model (see Fig. 1) is implemented as a dynamic Bayesian network. For each word in the vocabulary, the model essentially consists of two parallel HMMs, one for the audio stream and one for the video stream, each having $N$ states per word. The joint evolution of the HMM states is constrained by synchrony requirements imposed by additional random variables. This is a modification of the model proposed in [9].

The model allows the states in different streams to proceed through their sequences at different rates (i.e. asynchronously). This asynchrony is not completely unconstrained, however: Sets of state sequences that are more "synchronous" may be more probable than less "synchronous" ones. To make the notion of asynchrony more precise, let the variable $i_t^F$ be the index into the state sequence of stream $F$ at time $t$; i.e., if stream $F$ is in the $n^{th}$ state of a word at time $t$, then $i_t^F = n$. The states are traversed in ascending order from 0 to $N-1$. We define the degree of asynchrony between the streams at time $t$ as $|i_t^A - i_t^V|$. The probabilities of varying degrees of asynchrony are given by the distribution of the variable $a_t$. The variable $c_t$ simply checks that the degree of asynchrony between the two streams is in fact equal to $a_t$. This is done by having $c_t$ always observed with value 1, and defining its distribution as

$$\begin{aligned} P\left(c_t = 1 | a_t, i_t^A, i_t^V\right) &= 1 \\ \iff |i_t^A - i_t^V| &= a_t, \end{aligned}$$

and 0 otherwise. [1] This model therefore has only a few extra parameters in addition to the parameters of the individual streams; if we allow asynchrony by a maximum of $k$ states, then we need only learn an additional $k-1$ probability values.

For the audio stream, the observations $O_t^A$ are features extracted from the audio waveform, and for the video stream, the observations $O_t^V$ are features extracted from the corresponding image sequence of the speaker's mouth. The observation models are mixtures of Gaussians, one mixture per state. Recognition corresponds to finding the most likely settings of the hidden variables, and reading off the word sequence corresponding to the hypothesized states.

To perform recognition with this model, we can use standard DBN inference algorithms [10]. All of the parameters of the distributions in the DBN, including the observation models, the state transition probabilities, and the probabilities of asynchrony between streams, are learned via maximum likelihood using the Expectation-Maximization (EM) algorithm [4].

## 4. DATA AND PROCESSING

We evaluated the above model and several baselines on the isolated digits portion of the Clemson University Audio-Visual Experiments (CUAVE) database [13]. This part of the database consists of 36 speakers speaking 50 English digits each, except for one of the speakers, who only spoke 40 digits. The speakers all faced the camera. Although this is the isolated digits portion of the database, it is a continuous ASR task. We used a training set consisting of 22 speakers, a development set with 6 speakers and a test set with 6 speakers, all chosen at random from a 34-speaker subset of the database. [2] We divided the recording of each speaker into ten-digit utterances, each containing the digits "zero" through "nine" in order. The total number of words in the train, development and test sets was 1090, 300, and 300, respectively.

To evaluate our model at different levels of audio noise, we added babble noise from the NOISEX database to the clean audio, and extracted Mel-frequency cepstral coefficients (MFCCs) from the clean and noisy waveforms. The audio observations consisted of 14 MFCCs, plus first and and second derivatives, minus the first energy coefficient, resulting

[1] A simpler structure without the $a$ variables, as in [8], could be used, but it would not allow for EM training of the asynchrony probabilities.

[2] Speakers 25 and 18 were left out because of problems with their visual processing.

| SNR | -4dB | 4dB | 6dB | 10dB | 12dB |
|---|---|---|---|---|---|
| AHMM | 70.7 | 32.3 | 23.3 | 13.0 | 7.3 |
| VHMM | 56.7 | 56.7 | 56.7 | 56.7 | 56.7 |
| MHMM | 47.3 | 23.7 | 18.7 | 6.0 | 3.3 |
| aDBN(1) | 50.0 | 23.3 | 18.0 | 7.0 | 3.7 |
| aDBN(2) | **47.3** | **19.3** | **13.0** | **5.0** | **3.0** |
| aDBN(3) | 49.3 | 19.7 | 15.3 | 6.3 | 3.3 |

**Table 1**. Word error rate (WER), in percent, on the test set for various models.

in a 41-dimensional vector in each frame. The visual observations consisted of 35 discrete cosine transform (DCT) coefficients of a 16-by-16 grayscale mouth subregion, plus first derivatives, for a total of 70 dimensions.[3] Mean and variance normalization was applied to the audio and visual features on a per utterance basis. The original visual features were sampled at 29.97Hz; however, to enable state-synchronous audio-visual fusion as a baseline method, they were interpolated to 100Hz to match the audio frame rate.

## 5. EXPERIMENTS

Each model was implemented as a whole-word recognizer, with $N$=16 states per word. The vocabulary consisted of the 10 digit words “zero” through “nine”, a 16-state silence word, and a 1-state short pause word. Decoding was restricted to 10 digit words per utterance. The development set was used to tune the number of Gaussians in the mixtures for the audio-only and the video-only models, and the stream weights for the audio-visual models.

We implemented and evaluated all of the models using GMTK, the Graphical Models Toolkit [2, 1]. First, we evaluated audio-only and visual-only HMM-based models, which we refer to as AHMM and VHMM, respectively. The best number of Gaussian components per mixture was found to be 1 for the AHMM and 4 for the VHMM. The top two rows of Table 1 show the performance of these single-stream models on the test set, measured in terms of the word error rate (WER). The columns in the table correspond to the different signal-to-noise ratios (SNRs) in the noisy audio. All models were trained on clean audio. The AHMM achieved 0.0% error rate on the clean test set. However, its WER degraded significantly in noise, up to 70.7% for the noisiest condition. The visual-only model achieved a 56.7% error rate.

Next, we compared synchronous and asynchronous audio-visual models. To achieve the best balance between input streams at different audio noise levels, we made use of stream weights: an audio stream weight, $\lambda$, and a video stream weight $(2 - \lambda)$. [4] The models were trained with both weights set to 1.0. During decoding of the development set, $\lambda$ was varied from 0.0 to 2.0 with a step size of 0.1. The weight that produced the best results on the development set is shown in Table 2 for each model. Results on the test set were obtained using these weights.

| SNR | -4dB | 4dB | 6dB | 10dB | 12dB |
|---|---|---|---|---|---|
| MHMM | 0.8 | 1.3 | 1.4 | 1.9 | 1.8 |
| aDBN(2) | 1.2 | 1.5 | 1.6 | 1.8 | 1.8 |

**Table 2**. The best audio stream weight ($\lambda$) obtained on the development set for each audio-visual model.

The synchronous model we compare to is the multi-stream HMM (MHMM), which consists of a single stream of hidden states with two observation streams. The observation log likelihood at each state is a weighted combination of the audio and visual observation log likelihoods, using the stream weights found above. The MHMM observation model parameters were initialized using the Gaussian parameters from the AHMM for the audio stream, and from the VHMM for the video stream. The number of Gaussians in the mixtures was set to one. These initial parameters were trained to weak convergence (2% relative difference in log-likelihood). Then, the MHMM was trained to stronger convergence (.5%). As shown in Table 1, the multi-stream HMM achieved lower WER than the audio-only baseline across all noise conditions.

Finally, we evaluated the proposed asynchronous DBN model (aDBN). Again, the initial parameters were those of the weakly trained single-Gaussian AHMM and VHMM, and the aDBN model was trained for several more EM iterations. We also varied the maximum degree of asynchrony allowed between streams. This was achieved by setting the initial probabilities of asynchrony, i.e. of values of the $a$ variable, to 0 for all values greater than $a^{MAX}$, for $a^{MAX} = 1, 2, 3$. The results are shown in Table 1 in rows titled aDBN, with $a^{MAX}$ in parentheses. The results show that, for this task, the best performance is achieved by allowing up to 2 states of asynchrony. Allowing 3-state asynchrony did not further improve the results, and the learned probability $p(a = 3)$ was less than 0.01. Looking at the learned parameters of aDBN(2), the probability of the streams being completely synchronous was 0.23, being asynchronous by 1 state 0.54 and being off by 2 states 0.23. Comparing the results for the MHMM and the aDBN(2), we see that the aDBN achieves a lower WER than the MHMM on test sets for all SNRs except -4 dB. The same was true for the development set.

[3] The visual observations were provided by Amarnag Subramanya. We are grateful for his assistance.

[4] The weights sum to 2 to match the effective observation weight of a model with no explicit weights.

## 6. SUMMARY AND FUTURE WORK

In this paper, we applied a two-stream asynchronous DBN model to audio-visual digit recognition and compared it to an audio-only HMM, video-only HMM, and synchronous multi-stream HMM. All audio-visual models improved over the audio-only baseline across all SNR levels, as expected on this task. The aDBN achieved lower word error rates on the test set than the MHMM, for all SNRs except -4 dB. We evaluated several versions of the aDBN model corresponding to different degrees of maximum allowed asynchrony between the audio and visual streams. We found that the best choice for this task is a maximum of 2 states of asynchrony.

Currently, state asynchrony is only allowed within word boundaries in our model. That is, all state counters are reset to 0 after a word transition. However, to account for anticipatory co-articulation which occurs when the mouth starts moving in anticipation of the upcoming word, we plan to incorporate cross-word asynchrony into the model. In addition, the current model is extremely simple, assuming that the degree of asynchrony is context-independent and that each stream is equally likely to outpace the other. This results in a small number of additional parameters, but it would also be interesting to investigate relaxation of these restrictions. Finally, we would like to compare our model with other asynchronous DBN-based models.

## 7. REFERENCES

[1] J. Bilmes, "The Graphical Models Toolkit", http://ssli.ee.washington.edu/ bilmes/gmtk/.

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

[3] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, pp. 994-999, San Juan, Puerto Rico, June 1997.

[4] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, **39**:1–38, 1977.

[5] Z. Ghahramani and M. Jordan, "Factorial hidden Markov models," in Proc. Conference Advances in Neural Information Processing Systems, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds., vol. 8, pp. 472-478, MIT Press, Cambridge, MA, USA, 1995.

[6] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "DBN based multi-stream models for audio-visual speech recognition," in Proc. ICASSP, 2004.

[7] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in Proc. Human Language Technology Conference, San Diego, 2002.

[8] K. Livescu and J. Glass, "Feature-based pronunciation modeling for speech recognition," in Proc. HLT/NAACL, 2004.

[9] K. Livescu and J. Glass, "Feature-based pronunciation modeling with trainable asynchrony probabilities," in Proc. ICSLP, 2004.

[10] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, U.C. Berkeley CS Division, 2002.

[11] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop," in Proc. Works. Signal Processing, pp. 619-624, Cannes, France, 2001.

[12] H. Nock and S. Young, "Loosely-Coupled HMMs for ASR," in Proc. ICSLP, 2000.

[13] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in Proc. ICASSP, 2002.

[14] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech", in Proc. IEEE, 2003.

[15] K. Saenko, M. Siracusa, K. Wilson, K. Livescu, J. Glass, and T. Darrell, "Visual Speech Recognition with Loosely Synchronized Feature Streams," in Proc. International Conference on Computer Vision, 2006.

[16] Y. Zhang, Q. Diao, S. Huang, W. Hu, C. Bartels, and J. Bilmes, "DBN based multi-stream models for speech", in Proc. ICASSP, 2003.