INVESTIGATIONS WITH MULTISTREAM ARTICULATORY MODELS FOR SMALL-VOCABULARY VISUAL SPEECH RECOGNITION

Kate Saenko, Karen Livescu and Trevor Darrell

Computer Science and Artificial Intelligence Laboratory, MIT
Cambridge, MA, USA

ABSTRACT

We report on investigations with a small-vocabulary visual speech recognition (lipreading) system based on modeling the joint evolution of multiple state streams, each of which is associated with an articulatory feature. We extend and improve upon earlier preliminary experiments with such a system. In this model, represented as a dynamic Bayesian network, the observations consist of outputs of SVM classifiers for each articulatory feature, and the state emission distributions are Gaussian. The articulatory features correspond to the positions of the lips and teeth and have 2-4 possible values each. The state streams are essentially hidden Markov chains, with some asynchrony allowed between streams to account for co-articulation. The asynchrony constraints are probabilistic, with the probabilities of different degrees of asynchrony learned from data via the Expectation-Maximization algorithm along with the remaining model parameters. We describe experiments on a recognition task consisting of 20 isolated phrases. We investigate the behavior of the multi-stream model, and compare it to single-stream hidden Markov models using either viseme classifier outputs or image appearance parameters, such as PCA coefficients of image DCTs, as observations. We find that increasing the number of streams from three to four improves performance, and that the multi-stream model outperforms the single-stream baselines.

1. INTRODUCTION

The task of visual speech recognition (VSR), also sometimes referred to as automatic lipreading, is useful either as a part of an audio-visual speech recognition (AVSR) system, or as a stand-alone system when the audio is extremely noisy or not available. We have recently begun investigating the role of articulatory feature (AF) based models in VSR [9, 10, 11]. In general, articulatory features correspond to the positions of the speech articulators (lips, tongue, vocal fold vibration state, and so on) during speech production. In particular, the articulatory features relevant to VSR are those that are visible: lip rounding, lip opening, and positions of the teeth and tongue when they are visible.

One interesting characteristic of articulatory features is that their trajectories are "semi-independent", in the sense that different articulators may move at different rates between the target positions necessary to produce a given utterance. Our work is based on [6], which introduced an articulatory feature-based DBN for acoustic pronunciation modeling, and [9], which introduced visual articulatory features.

Automatic lipreading is a very challenging task, usually not admitting high accuracy except in constrained scenarios. In previous experiments, we have shown the benefit of modeling articulatory features in a medium-vocabulary word *ranking* task [10]. More recently, we have applied this idea to a complete, small-vocabulary isolated phrase recognizer. We could imagine using such a system to control a car stereo, a situation in which hands-free control is useful and in which the acoustic signal may be highly degraded. Using articulatory features corresponding only to the position of the lips, we showed improved performance of a multi-stream articulatory model over a single-stream viseme-based hidden Markov model (HMM) [11]. Here we extend this model to use an additional feature corresponding to the position of the teeth, and investigate the performance of our models in comparison with both viseme-based HMMs and HMMs using more standard image appearance features as observations.

In the next section, we describe our model and its implementation as a dynamic Bayesian network with SVM outputs as observations. We then describe experiments in the car stereo control domain using a small two-speaker data set.

2. AN ARTICULATORY FEATURE-BASED MODEL FOR SMALL-VOCABULARY VSR

Our model of visual speech as several streams of articulatory features is motivated in part by the desire to capture the underlying physical process of speech production. We will present the model in detail after describing the advantages of AF-based visual speech recognition.

Conventional approaches to visual speech recognition separate the space of mouth positions into linguistic units called "visemes", which are phonemes clustered into visually distinguishable classes. For example, the phonemes /p/, /b/, and /m/ may be clustered to produce a "bilabial closure" viseme. However, the appearance of a viseme can be heavily influenced by context. This often occurs when articulatory gestures not primarily involved in the production of the current sound evolve asynchronously from the ones that are. Classifying multiple articulatory features, such as lip rounding and lip opening, captures more information than just classifying a single viseme. Furthermore, allowing the features to sometimes proceed through their trajectories asynchronously accounts for co-articulation effects. Therefore, we assign several AF labels to each speaking mouth image, instead of just a single viseme label. These labels produce multiple streams, each corresponding to a speech production gesture, essentially factoring the viseme state space. An alternative is to use context-dependent viseme units. However, visual coarticulation effects can span three or more visemes, requiring a large number of context-dependent models. This leads to an inefficient use of training data, and cannot anticipate new variations. In contrast, an asynchronous AF approach offers a more flexible and parsimonious architecture.

We now describe our model in detail, starting with the articulatory feature set. Previously, we have used three features associated with the lips: lip opening (LO), with values 'closed', 'narrow', 'medium' and 'wide'; lip rounding (LR), with values 'rounded' and 'unrounded', and labio-dental (LD), with values 'yes' and 'no'. Although these are sufficient to differentiate between the phrases in our test vocabulary, in this paper, we have added a fourth AF describing the position of the teeth (TP), with three values: 'unknown', 'neutral' and 'open'. The goal was to determine whether using more features than the minimum needed to distinguish our vocabulary would improve performance.

There are several ways to classify the input observations as articulatory features. We chose to use support vector machines (SVMs), which have been shown to perform well on lipreading tasks [4]. However, an SVM is an inherently static classifier. Therefore, we combine the outputs of the SVMs over time using hidden Markov chains. Since we want to allow the articulatory features to proceed in a semi-independent fashion, we cannot use a single-chain HMM. Instead, we use a dynamic Bayesian network (DBN) including separate streams for different features, with some synchrony constraints imposed on pairs of streams.

Figure 1 shows three frames of the DBN used in our experiments. The model consists of M parallel HMMs, one per AF, where the joint evolution of the HMM states is constrained by K synchrony requirements imposed by the variables $c^j, j = 1...K$. In this paper, we concentrate on a small-vocabulary task, and therefore have a separate DBN for each of the V phrases in the vocabulary, with an equal number of states S for each phrase. The total number of states is therefore VSM. For comparison, Figure 2 shows a conventional single-stream HMM, which we use as a baseline in our experiments.

Fig. 1. DBN for feature-based VSR. t is time, i_t^F is an index into the state sequence of feature stream F, where $F \in \{$LO,LR,LD,TP$\}$.

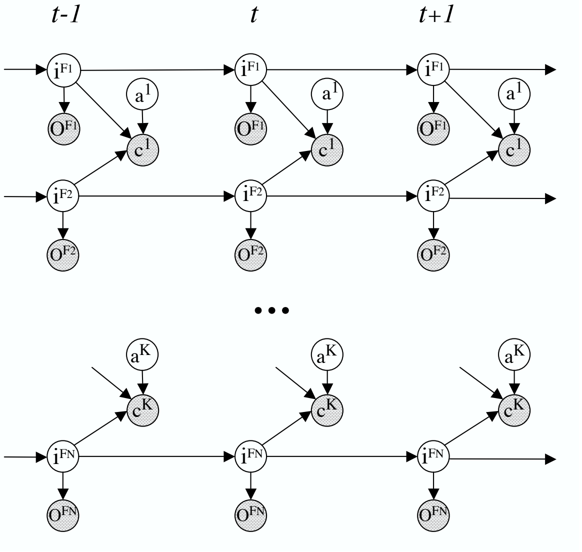
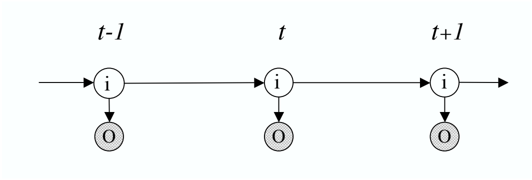


Fig. 2. Single-stream HMM. i is the index into the state sequence, and O is the observation.



To make the notion of asynchrony more precise, let the variable i_t^F be the index into the state sequence of feature stream F at time t; i.e., if stream F is in the n^{th} state of a given phrase at time t, then $i_t^F = n$ (see Figure 1). We define the degree of asynchrony between two feature streams F_1 and F_2 at time t as $|i_t^{F_1} - i_t^{F_2}|$. The probabilities of varying degrees of asynchrony are given by the distributions of the a^j variables. Each c_t^j variable simply checks that the degree of asynchrony between its parent feature streams is in fact equal to a_t^j. This is done by having the c_t^j variable always observed with value 1, with distribution

$$P\left(c_t^j = 1 | a_t^j, i_t^{F_1}, i_t^{F_2}\right) = 1 \iff |i_t^{F_1} - i_t^{F_2}| = a_t^j,$$

and 0 otherwise, where $i_t^{F_1}$ and $i_t^{F_2}$ are the indices of the feature streams corresponding to c_t^j. [1] For example, for c_t^1, $F_1 = LR$ and $F_2 = LO$.

Rather than use hard AF SVM decisions, or probabilistic outputs as in [4] and [10], we use the outputs of the decision function directly, as we have found this to produce the best results in initial experiments [11]. For each stream, the observations O^F are the SVM margins for feature F and the observation model $P(O^F|i^F)$ is a Gaussian mixture.

[1] A simpler structure, as in [5], could be used, but as pointed out there, it would not allow for EM training of the asynchrony probabilities.

Table 1. The mapping from visemes to AFs.

Viseme	LO	LR	LD	TP
1	closed	any	no	any
2	any	any	yes	any
3	narrow	rounded	no	any
4	medium	unrounded	any	neutral
5	medium	unrounded	any	open
6	medium	rounded	any	neutral
7	medium	rounded	any	open
8	wide	any	any	any

We have previously also investigateed dictionary models [10], which cluster together states that should have the same articulatory positions. However, for the current small-vocabulary task, we have found that we obtain better performance with phrase-specific models [11]. Therefore, our current approach uses a separate DBN for each phrase in the vocabulary, with i^F ranging from 1 to the maximum number of states S. Recognition corresponds to finding the phrase whose DBN has the highest Viterbi score.

To train and test this model, we can use standard DBN inference algorithms [7]. All of the parameters of the DBNs, including the observation models, the per-feature state transition probabilities, and the probabilities of asynchrony between streams, are learned simultaneously via maximum likelihood using the Expectation-Maximization (EM) algorithm [3].

3. EXPERIMENTS

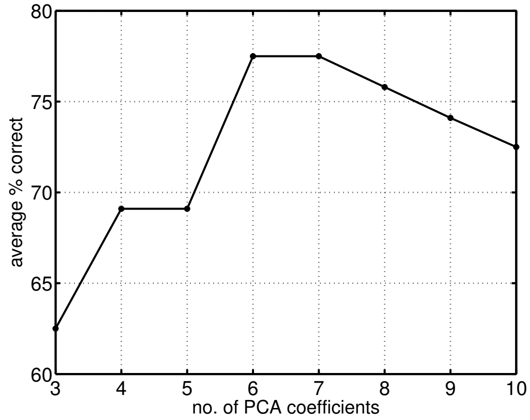
In this section, we evaluate our recognizer on a set of short phrases, and compare it to several baselines. In all of the following experiments, the LIBSVM package [2] was used to implement the SVM classifiers, and the Graphical Models Toolkit [1] was used for all DBN computations.

3.1. Data and experimental setup

We collected approximately 2.5 minutes of video, consisting of two speakers reading phonetically-balanced English sentences, recorded at 30 frames per second, for training of viseme and AF classifiers. A subset of the frames was labeled manually with AF values: 500 labels for the LO feature, 414 labels for LR, 232 labels for LD, and 399 labels for TP. Labels for eight visemes were obtained from combinations of AF labels (see Table 1), for a total of 669 labeled frames. These visemes correspond to the combinations of AF values that occur in the training data, making for a fair comparison between AF-based and viseme-based models.

Each frame of video was processed as follows. First, a face was detected and a mouth localized within the lower region of the face using the visual front end described in [11]. Then, the mouth region was resized to 32 by 16 pixels. Two types of observation vectors were extracted from the resized mouth. The first was obtained by applying a 2-dimensional discrete cosine transform (DCT) to the image, and retaining the top D coefficients. The second was obtained by applying a principal component analysis (PCA) transform to the full set of 512 DCT coefficients, and retaining the top C PCA components. Both of the above fall under the category of "appearance features", and have been studied extensively in the context of visual speech recognition [8]. In the following experiments, we evaluate HMMs using these two types of appearance features as observations.

Fig. 3. Performance of the PCA baseline in terms of average % correct phrases as a function the number of PCA coefficients.



AF and viseme SVM classifiers were trained using 75 of the PCA observations per frame, along with the manual labels described above. A one-vs.-all strategy was employed for multi-class SVMs, with a total of nine SVM classifiers trained for the four AFs: four for LO, one for LR, one for LD, and three for TP. Also, one SVM was trained for each of the eight visemes. All SVMs used Radial Basis Function (RBF) kernels. Since the choices of the free parameters of the SVMs—the error penalty and the RBF parameter—are crucial to their performance, a four-fold cross-validation was performed to find their optimal values.

We evaluate the baseline and proposed recognizers on a small-vocabulary task, consisting of video of the same two speakers saying short command phrases. The chosen 20 phrases could be used to control a hypothetical car stereo, e.g. "turn on the radio" [11]. Each command was recorded three times at different speaking rates, in order to get different amounts of co-articulation, for a total of 60 phrases per speaker. The speakers clearly enunciated the phrases during the first repetition (slow condition), then spoke successively faster during the second and third repetitions (medium and fast conditions). Each recognition experiment was conducted three times, training the system on two speed conditions and testing it on the remaining condition. The accuracies were averaged over the three trials to produce the final result.

Table 2. Number of phrases (out of 40) recognized correctly by various models. The first column lists the held-out speed condition used to test the model. The remaining columns show results for three baseline models and for the synchronous and asynchronous versions of our model with 3 and 4 articulatory features. The 6-vis and 3-AF results are the initial results reported in [11].

test	HMM baselines			3-AF-based		4-AF-based	
condition	PCA	6-vis	8-vis	sync	async	sync	async
fast	23	16	19	23	25	23	25
med.	34	19	29	29	30	36	37
slow	36	27	22	27	25	34	33
avg. %	**77.5**	**51.6**	**58.3**	**64.1**	**65.8**	**77.5**	**79.2**

3.2. Results

We evaluated the HMM baseline with both the DCT and PCA-based appearance features. The graph in Figure 3 shows performance for the PCA baseline as the number of coefficients C is varied from 3 to 10. The best PCA-based model (for $C = 6$) outperformed all DCT-based models, so we consider it the appearance-based baseline.

Table 2 summarizes the results of our experiments, comparing our model to both the PCA and viseme-based HMMs on the command phrase task. In all of the systems shown in Table 2, we used single Gaussians with tied diagonal covariance matrices as observation models. Increasing the number of mixtures to two in the PCA-based HMM resulted in degraded performance, most likely because of the limited size of the data set. The second and third columns show the performance of the HMM with viseme SVM outputs as observations. Although both models, in principle, use enough visemes to distinguish all of the phrases in the vocabulary, they did not do as well as the appearance-based HMMs. One reason for this may be that some visemes occur infrequently, and thus have too little training data. On the other hand, AFs are able to utilize training data more efficiently. Another possibility is that there are too many classes in the viseme-based multiclass SVMs, suggesting that investigations with alternative, inherently multiclass, classifiers may be useful.

The fourth and fifth columns show that our initial 3-AF DBN outperforms either viseme-based HMM but not the PCA HMM. This is not surprising, since the three features carry less information than the full image on which the PCA coefficients are based. The sixth column shows that the 4-AF DBN performs as well as the PCA HMM. Finally, the asynchronous version of the 4-AF model, in the last column, achieved the best overall performance, although the difference is not statistically significant on this data set. In this model, three pairs of streams were allowed to de-synchronize by up to one state[2]—LO and LR, LO and LD, and LO and TP—with the three asynchrony probabilities $p(a^j = 1)$ learned from the training data. Adding the fourth (TP) feature therefore improved the accuracy of the synchronous DBN from 64.1% to 77.5%, and of the asynchronous DBN from 65.8% to 79.2%. Note that the four AFs arguably still do not capture all of the relevant information in the image; for example, some aspects of tongue motion may be visible and independently informative.

4. CONCLUSIONS

We conducted investigations of a small-vocabulary VSR system based on articulatory-feature modeling of speech. We extended and improved upon a previous 3-feature model [11] by introducing a teeth position feature. We also compared our model with an HMM using viseme classifier outputs as observations, as well as HMMs using more standard appearance-based features. We found that the asynchronous AF-based DBN outperforms these models on a realistic and challenging real-world recognition task. However, the data set is rather small, and statistical significance results can only be stated regarding the improvement from viseme-based to AF-based models and from 3-AF to 4-AF models.

In the future, we plan to explore a more comprehensive set of articulatory features—for example, including information about the visible portion of the tongue—and extend the model to AVSR. We would also like to explore unsupervised learning of AF classifiers, in order to avoid manual data labeling, as well as to evaluate on larger datasets.

5. REFERENCES

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

[2] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," 2001. http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[3] 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.

[4] M. Gordan, C. Kotropoulos, and I. Pitas, "A support vector machine-based dynamic network for visual speech recognition applications," in *EURASIP Journal on Applied Signal Processing*, 2002.

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

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

[7] K. Murphy, *Dynamic Bayesian networks: representation, inference and learning*. Ph.D. thesis, U.C. Berkeley CS Division, 2002.

[8] 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.

[9] K. Saenko, J. Glass, and T. Darrell, "Articulatory features for robust visual speech recognition," in *Proc. ICMI*, 2005.

[10] K. Saenko, K. Livescu, J. Glass, and T. Darrell, "Production domain modeling of pronunciation for visual speech recognition," in *Proc. ICASSP*, 2005.

[11] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual Speech Recognition with Loosely Synchronized Feature Streams," in *Proc. ICCV*, 2005.

[2] Note that there is no restriction on the amount of *time* that the streams spend in asynchronous states.