PRODUCTION DOMAIN MODELING OF PRONUNCIATION FOR VISUAL SPEECH RECOGNITION

Kate Saenko, Karen Livescu, James Glass, and Trevor Darrell

{saenko, klivescu, jrg, trevor}@csail.mit.edu
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar Street, Cambridge, MA 02139, USA

ABSTRACT

Articulatory feature models have been proposed in the automatic speech recognition community as an alternative to phone-based models of speech. In this paper, we extend this approach to the visual modality. Specifically, we adapt a recently proposed feature-based model of pronunciation variation to visual speech recognition (VSR) using a set of visually-salient features. The model uses a dynamic Bayesian network to represent the evolution of the feature streams. A bank of SVM feature classifiers, with outputs converted to likelihoods, provides input to the DBN. We present preliminary experiments on an isolated-word VSR task, comparing feature-based and viseme-based units and studying the effects of modeling inter-feature asynchrony.

1. INTRODUCTION

Traditionally, visual speech is modeled as a single stream of contiguous units, each corresponding to a hidden phonetic state. These units are defined by mapping several visually similar phonemes to a single *viseme*. However, a many-to-one mapping does not always exist, as the appearance of the mouth during phone production can be heavily influenced by the surrounding context. This often occurs when articulators not primarily involved in the production of the current phone evolve asynchronously from the primary articulators. Figure 1 shows an example of such de-synchronization in a segment taken from the center of the utterance "promote birth". Note that during the /t/ segment, the lips, which would normally be in a medium-open position, are completely closed due to the upcoming bilabial phoneme.

One way to capture such variability is by using context-dependent units. However, visual coarticulation effects such as the one described above can span three or more phonemes, requiring a large number of models. This leads to an inefficient use of the training data, and cannot anticipate new variations. Alternatively, we can break the assumption that visemes are the basic building blocks of visual speech and instead model articulatory events, which we believe are the more natural visual units.

This research was supported by ITRI and by DARPA under SRI subcontract No. 03-000215.

Fig. 1. *Mouth images aligned with the corresponding phoneme sequence.*

From the point of view of speech production, each sound can be described by a unique combination of several articulator states, or *articulatory features (AFs)*, such as: the presence or absence of voicing, the position of the tongue body and tongue tip, the opening between the lips, and so on. A word consists of a number of (not necessarily synchronous) sequences of articulatory targets. Conventional speech models make the simplifying assumption that a word can be broken up into phonemes, each of which is an atomic unit. The articulatory approach offers a more flexible and parsimonious architecture. For example, the visual speech segment in Figure 1 can be explained as the de-synchronization of the lips from the remaining articulators. Although similar pronunciation models have been used in modeling spontaneous *acoustic* speech [9], to the best of the authors' knowledge, this is the first application of the multi-stream articulatory feature approach in the visual domain. In the following sections, we present a visual speech recognition framework that models visual speech in terms of the underlying articulatory processes.

2. VISUAL ARTICULATORY FEATURE DETECTION

We treat articulatory features as the hidden states underlying the surface visual observations [12], and learn them using a supervised learning approach. An *observed* feature vector is used as the input to a statistical classifier, which outputs the *hidden* articulatory feature labels. A preprocessing step extracts the observed feature vector from the input image. In principle, each articulatory feature classifier could use

different observation-level measurements. For example, the classifier for "lip rounding" could take motion vectors as input, while the "dental" classifier could use color input.

We assume a set of training examples with images of mouths and the corresponding articulatory feature labels; each image has several discrete labels, one for each AF. In preliminary experiments, we have found that support vector machines (SVMs) outperform Gaussian Mixture Models on the task of articulatory feature classification for a single speaker, and have therefore chosen to use SVM classifiers.

In dealing with the visual modality, we are obviously limited to modeling the visible articulators. As a start, we are using features associated with the lips, since they are always visible in the image: LIP-OP (closed, narrow, medium, wide), LIP-RND (rounded, unrounded) and LAB-DEN (labio-dental, not labio-dental). This ignores other articulators that might be distinguishable from the video, such as the tongue and teeth; we plan to incorporate these in the future.

Note that the standard formulation of SVM classification produces a hard decision (the class label). However, in order to not lose information by forcing a decision at this early stage, we produce soft decisions in the form of posteriors $P(F_t = f|X_t = x)$, where X_t is the observation at time t and F_t is a particular AF. SVM outputs are converted to posterior probabilities using the implementation of the sigmoidal mapping described in [3]. Furthermore, since our recognizer uses a generative model, it is more natural to use likelihoods than posteriors, so we convert the posteriors to (scaled) likelihoods using $P(X_t = x|F_t = f) \propto P(F_t = f|X_t = x)/P(F_t = f)$.

3. A DYNAMIC BAYESIAN NETWORK FOR FEATURE-BASED VSR

Our recognition model is based on the work described in [8, 9]. The model generates, for each word in its vocabulary, all sequences of AF values that are possible realizations of that word, along with the probabilities of those realizations. While a word may have only one or two baseform (dictionary) pronunciations, such as *either* → {/iy dh er/, /ay dh er/}, there may be thousands of AF combinations that are possible realizations of the word. In order to take advantage of the semi-independent evolution of the AF streams—in other words, the factorization of the AF state space—we implement the model as a dynamic Bayesian network (DBN). Figure 2 shows (a slightly simplified version of) one frame of the DBN used in our experiments. Conditioned on the identity of the word—or, in the case of words with more than one baseform, conditioned on the word and the baseform—the model essentially consists of three parallel HMMs, one per AF, where the joint evolution of the HMM states is constrained by synchrony requirements as we describe below.

Given a word, the model generates its realizations as follows. First, a baseform is drawn from the set of allowed baseforms for the word. This baseform pronunciation defines a set of target feature value trajectories, one for each AF F. The AFs then proceed through their trajectories, possibly at different rates (i.e. asynchronously). In Figure 2, i_t^F is an index into the trajectory of feature F at time frame t; i.e., if F is in the n^{th} state of its trajectory at time t, then $i_t^F = n$. U_t^F is the underlying target value corresponding to this state. We define the degree of asynchrony between two features F_1 and F_2 at time t as $|i_t^{F_1} - i_t^{F_2}|$. This asynchrony is not completely unconstrained: Sets of trajectories that are more "synchronous" may be more probable than less "synchronous" ones, and we impose a limit on the maximum asynchrony between sets of features. The probabilities of varying degrees of asynchrony are given by the distributions of the $async^j$ variables. $checkSync_t^j$ simply checks that the degree of asynchrony between its parent features is in fact equal to $async_t^j$: It is always observed with value 1 and its distribution is

$$\begin{aligned} P\left(checkSync_t^j{=}1|async_t^j, i_t^{F_1}, i_t^{F_2}\right) &= 1 \\ \iff |i_t^{F_1} - i_t^{F_2}| &= async_t^j, \end{aligned}$$

and 0 otherwise, where $i_t^{F_1}$ and $i_t^{F_2}$ are the indices of the features corresponding to $checkSync_t^j$. [1] In the model of Figure 2, $async_t^1$ is the degree of asynchrony between LIP-RND and LIP-OP, and $async_t^2$ is the degree of asynchrony between LIP-OP and LAB-DEN.

Finally, the *surface* value S_t^F that is actually produced by the speaker at time t for feature F may differ from the underlying value U_t^F, usually due to undershoot (e.g. incomplete lip closure for a /b/) or context effects. For the experiments in this paper, however, we focus on asynchrony modeling and assume that $S_t^F = U_t^F$ for all t, F.

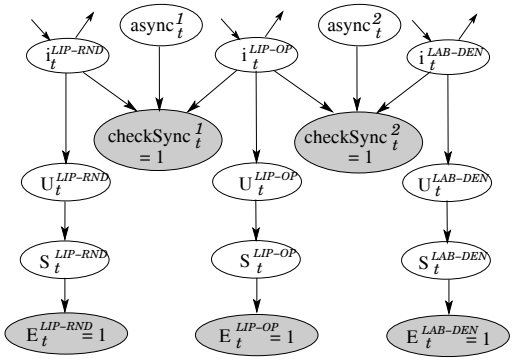

Fig. 2. *One frame of a DBN for feature-based VSR. All variables are discrete-valued. U_t^F are the underlying feature values, and S_t^F are the surface values. i_t^F is an index into the state sequence of feature F. Edges without parents/children in the figure connect the i_t^F in adjacent frames.*

[1] A simpler structure could be used for decoding, but it would not allow for EM training of the asynchrony probabilities (see [8]).

In order to incorporate the likelihoods computed from the SVM outputs, we use the Bayesian network mechanism of *soft evidence* [1]. This is used when a variable is not observed but we have some information that causes us to favor some values over others; this is exactly what the SVM outputs tell us about the AF values. Soft evidence allows us to combine a generative model with likelihoods computed by any means, including discriminative classifiers such as SVMs. This can be done by adding, for each articulatory feature F, a "dummy" evidence variable E_t^F, whose value is always 1 and whose distribution is constructed so that $P(E_t^F = 1|S_t^F = f)$ is proportional to the likelihood $P(X_t = x|S_t^F = f)$ computed from the SVM discriminant values.

Thus far we have used this model for isolated-word recognition, which amounts to finding the word that maximizes the probability of the observations. The parameters of the distributions in the DBN can be learned from SVM soft evidence outputs for a set of training data, for example using the Expectation-Maximization (EM) algorithm [5]. Both tasks can be accomplished using standard DBN inference algorithms [10]. In the proof-of-concept experiments described below, however, a very small data set was used and no DBN parameter learning was done; the DBN parameters were set manually, using linguistically plausible values.

4. EXPERIMENTS AND RESULTS

We have conducted pilot experiments to investigate several questions. First, we would like to compare the effects of using feature-based versus viseme-based *classifiers*, as well as of using a feature-based versus viseme-based *pronunciation model*. A viseme-based pronunciation model is a special case of our DBN, in which the features are constrained to be completely synchronous (i.e. $async_t^j$ is identically 0) and no feature changes are allowed (i.e. $S_t^F = U_t^F$). Using viseme classifiers with a viseme-based pronunciation model results in essentially the conventional viseme-based HMM that is used in most VSR systems. In order to use a feature-based pronunciation model with viseme classifiers, we use a many-to-one mapping from surface features (S_t^F) to visemes. Also, since we do not have ground truth articulatory feature labels, we investigate how sensitive the system is to the quality of the training labels in terms of both feature classification and word recognition. In order to facilitate quick experimentation, these initial experiments focus on an isolated-word recognition task using a small data set and, as previously mentioned, manual settings for the (small number of) DBN parameters.

4.1. Data and Visual Signal Preprocessing

For these initial experiments, we used 21 utterances taken from a single speaker in AVTIMIT [7], a corpus of audio-visual recordings of subjects reading phonetically balanced sentences with a vocabulary of 1793 words. Of these, 10 utterances were used for training and 11 for testing. To simulate the isolated-word task, utterances were split into words, resulting in a 70-word test set. Each visual frame was also manually transcribed with the three AFs.

The raw video stream was preprocessed by first extracting 37x54 pixel mouth regions from the image sequence and converting them to grayscale (see Figure 1). Then, a DCT transform was applied to each image to obtain a set of 1998 coefficients, of which the 900 highest-frequency coefficients were retained. The dimensionality was further reduced via PCA, with the top 100 PCA coefficients used as the final observation vector.

SVM type	LIP-OP	LIP-RND	LAB-DEN	viseme
Forced train(%)	44 (84)	63 (84)	50 (99)	33 (71)
Manual train(%)	59 (83)	78 (87)	87 (99)	N/A

Table 1. *Classifier accuracies for the feature and viseme SVMs, averaged over the N classes for each SVM: $acc = \frac{1}{N}\sum_{i=1}^{N} acc(class\ i)$. Chance performance is $\frac{1}{N}$. The numbers of classes are: 4 for LIP-OP, 2 for LIP-RND and LAB-DEN, and 6 for the viseme SVM (consisting of those combinations of feature values that occur in the forced transcriptions).*

4.2. Classification

A radial basis function (RBF) kernel SVM classifier was trained for each of the three AFs using LIBSVM [3]. To find optimal values of the SVM parameters, cross-validation was performed on the training set. In order to study the effects of training label accuracy, we considered two cases. In one case, phoneme labels from an acoustic forced transcription were converted to AF training labels using a deterministic mapping. In the other case, manual articulatory feature transcriptions were used. The first three columns of Table 1 show the resulting classification rates for each feature. Average per-class accuracies are reported to account for the uneven distribution of classes in the data, with the total percentage of correctly classified frames shown in parentheses. As we can see, manual labels result in higher per-class accuracy. The last column shows the performance of a viseme SVM trained from forced transcriptions.

4.3. Word ranking experiments

Because of the extreme difficulty of this task—lipreading isolated words excised from continuous speech with a relatively large vocabulary—we cannot expect to obtain reasonable word recognition error rates. Instead, we perform a word ranking experiment: For each spoken word in the test set, we compute the probability of each word in the vocabulary and rank the words based on their relative probabilities. Our goal is to obtain as high a rank as possible for the correct word. Performance is evaluated using both the mean rank of the correct word over the test set and the entire distribution of the correct word ranks.

We used the Graphical Models Toolkit [2] for all DBN computations. In the models with asynchrony, LIP-RND and LIP-OP were allowed to desynchronize by up to one

Classifier unit	Mean rank, sync model	Mean rank, async model
Viseme	281.6	262.7 (.1)
Feature, forced train	216.9 (.03)	209.6 (.02)
Feature, manual train	165.4 (.0005)	149.4 (.0001)
Feature, oracle	113.0 (2×10^{-5})	109.7 (3×10^{-5})

Table 2. *Mean rank of the correct word in several conditions.*

index value (one phoneme-sized unit), as were LIP-OP and LAB-DEN. Table 2 summarizes the mean rank of the correct word in a number of experimental conditions, and Figure 3 shows the entire empirical cumulative distribution functions (CDFs) of the correct word ranks in several of these conditions. In the CDF plots, the closer the distribution is to the top left corner, the better the performance. We consider the baseline system to be the viseme-based HMM, i.e. the synchronous pronunciation model using the viseme SVM.

In these experiments, the asynchronous pronunciation model always outperforms the synchronous one, regardless of the type of classifiers used. This may seem counterintuitive when viseme classifiers are used; however, certain apparently visemic changes may be caused by feature asynchrony; e.g. a /k/ followed by an /uw/ may look like an /ao/ because of LIP-OP/LIP-RND asynchrony. Next, the forced train vs. manual train comparison suggests that we could expect a sizable improvement in performance if we had more accurate training labels. While it may not be feasible to manually transcribe a large training set, we may be able to improve the accuracy of the training labels using an iterative training procedure, in which we alternate training the model and using it to re-transcribe the training set. To show how well the system could be expected to perform if we had ideal classifiers, we replaced the SVM soft evidence with likelihoods derived from our manual transcriptions. In this "oracle" test, we assigned a very high likelihood ($\approx$0.95) to feature values matching the transcriptions and the remaining likelihood to the incorrect feature values. Table 2 also gives the significance (p-value) of the mean rank differences between each model and the baseline (according to a one-tailed paired t-test [13]). The differences between each synchronous model and the corresponding asynchronous model are not significant ($p \geq .1$ on this test set), but all feature-based models are significantly better than the baseline.

5. SUMMARY AND FUTURE WORK

We have shown, for a limited VSR scenario, that a recognizer that models the articulatory asynchrony inherent in the human speech production system can outperform one that does not. We plan to continue testing this model on more data and in comparison with more realistic viseme-based baselines. We are also interested in applying this model to the problem of audio-visual fusion. Most state-of-the-art audio-visual speech recognizers model the asynchrony between the audio and visual streams [6]. However, the fusion

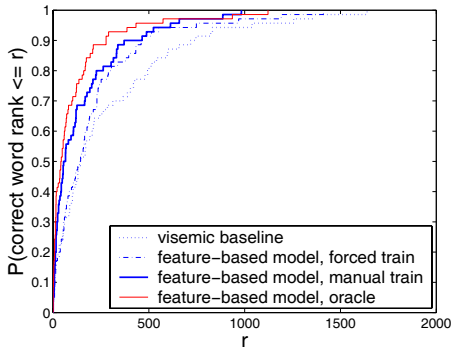

Fig. 3. *CDF of the correct word's rank, using the visemic baseline and the proposed feature-based model. The rank* r *ranges from 1 (highest) to the vocabulary size (1793).*

is done at the level of the phoneme/viseme. We believe that the feature is a more natural level for audio-visual fusion. This has been previously suggested [11], but to our knowledge has not been attempted. The structure we have used can be naturally extended to perform this type of fusion; all that is required is a complementary set of classifiers for the acoustically-salient features, such as voicing and nasality, and the corresponding additional variables in the DBN.

6. REFERENCES

[1] J. Bilmes, "On soft evidence in Bayesian networks," U. Washington Dept. of EE Tech. Report No. UWEETR-2004-0016, 2004.

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

[4] T. Dean and K. Kanazawa, "A model for reasoning about persistence and causation," *Computational Intelligence*, **5**:142–150, 1989.

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

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

[7] T. J. Hazen, K. Saenko, C. H. La, and J. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proc. ICMI*, 2005.

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

[11] P. Niyogi, E. Petajan, and J. Zhong, "A feature based representation for audio visual speech recognition," in *Proc. Aud. Vis. Sp. Conf.*, 1999.

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

[13] E. W. Weisstein, "Paired t-Test," from *MathWorld*–A Wolfram Web Resource. http://mathworld.wolfram.com/Pairedt-Test.html