Multistream Articulatory Feature-Based Models for Visual Speech Recognition

Short Papers

Multistream Articulatory Feature-Based Models for Visual Speech Recognition

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

Abstract—We study the problem of automatic visual speech recognition (VSR) using dynamic Bayesian network (DBN)-based models consisting of multiple sequences of hidden states, each corresponding to an articulatory feature (AF) such as lip opening (LO) or lip rounding (LR). A bank of discriminative articulatory feature classifiers provides input to the DBN, in the form of either virtual evidence (VE) (scaled likelihoods) or raw classifier margin outputs. We present experiments on two tasks, a medium-vocabulary word-ranking task and a small-vocabulary phrase recognition task. We show that articulatory feature-based models outperform baseline models, and we study several aspects of the models, such as the effects of allowing articulatory asynchrony, of using dictionary-based versus whole-word models, and of incorporating classifier outputs via virtual evidence versus alternative observation models.

Index Terms—Visual speech recognition, articulatory features, dynamic Bayesian networks, support vector machines.

1 Introduction

Visual speech recognition (VSR), also sometimes referred to as automatic lipreading, is the task of transcribing the words uttered by a speaker, given a silent video of the speaker's mouth or face. Human speech perception makes significant use of both the acoustic and visual signals [22], and automatic speech recognition can also benefit from the addition of visual observations [32]. VSR may also be useful as a stand-alone task, when the audio is extremely noisy or not available. Previous work on VSR has largely used hidden Markov model (HMM)-based methods [32], analogously to standard approaches for acoustic speech recognition [14]. In these approaches, a separate HMM is used to model each basic linguistic unit and the HMMs are connected to form a finite-state graph of all possible utterances. The basic linguistic unit is typically either a word or a *viseme*, the visual correlate of an acoustic phoneme, or basic speech sound.

In HMM-based methods, each state can be thought of as a configuration of the vocal tract, or of the visible portion of the vocal tract in the case of visual speech recognition. Each configuration, however, corresponds to a combination of states of multiple speech articulators: the degree of lip opening, lip rounding, the position of the tongue, and so on. Articulatory parameterizations are often used in linguistics to describe phonological structure. For example, Fig. 1 shows one parameterization similar to the one used in articulatory phonology [4]. The parameters are often referred to as *articulatory features (AFs)*. The AFs have been used in a number of models for acoustic speech recognition (e.g., [9], [17], [21]) and a benefit has been found in certain conditions for using separate classifiers of articulatory features rather than a single classifier of phoneme states [17].

In this paper, we explore the use of articulatory feature-based models for visual speech recognition. We use dynamic Bayesian network (DBN) models based in part on those of Livescu and Glass [20], but use discriminative classifiers of feature values to provide either observations or virtual evidence (VE) to the DBNs. In a previous work [34], we described an approach in which each visual frame is independently labeled with a set of classifiers, and showed a benefit of AF classifiers over viseme classifiers at various levels of visual noise. In this paper, we show the benefit of AF-based models in a medium-vocabulary word ranking task and in a small-vocabulary isolated phrase recognition task. Preliminary versions of some of the experiments in this paper have been reported previously [35], [36]. In this paper, we propose a class of models that unify those of [35], [36] and present additional experiments with improved results. Furthermore, we explore the use of dictionary-based models, in which words are broken into predefined subword units, versus whole-word (or, in our case, whole-phrase) models. We also explore the choice of observation model, i.e., the distribution of the visual signal given the hidden state.

2 Background

2.1 Visual Speech Recognition

Automatic visual speech recognition has been studied for over 20 years, both as a stand-alone task and as part of audiovisual systems. The main research issues are visual observation design, the choice of speech units, and decoding, i.e., mapping the sequence of observation vectors to speech units. A comprehensive review can be found in [32].

Visual observations can be categorized as either appearance-based, model-based, or a combination of the two (for an overview, see [32]). Appearance-based observations are based on the intensity and color information in a region of interest (usually the mouth and chin area). The dimensionality of the raw observation vector is often reduced using a linear transform. In contrast, model-based methods assume a top-down model of what is relevant for recognition, such as the lip contours. The parameters of the model fitted to the image are used as visual observations.

The most common decoding model for VSR is the HMM. Several multistream models have been applied to speech processing in recent years [11], [27], [28]. To our knowledge, we are the first to develop multistream DBNs for visual-only data streams.

Although most HMMs/DBNs use a Gaussian mixture model for the state-dependent distributions, several discriminative classification methods have been used, including distance in feature space [30], neural networks [23], and support vector machines (SVMs) [10]. In [10], one SVM was trained to recognize each viseme, and its output was converted to a posterior probability using a sigmoidal mapping [31]. In this work, we use SVMs to classify articulatory features in the video stream and use their outputs as observations in the DBN.

- *K. Saenko is with the MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, 32-D510, Cambridge, MA 02139. E-mail: saenko@csail.mit.edu.*
- *K. Livescu is with the Toyota Technological Institute at Chicago, University Press Building, 1427 East 60th Street, Second Floor, Chicago, IL 60637. E-mail: klivescu@uchicago.edu.*
- *J. Glass is with the MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, 32-G444, Cambridge, MA 02139. E-mail: glass@mit.edu.*
- *T. Darrell is with the UC Berkeley CS Division and the International Computer Science Institute (ICSI), 1947 Center Street, Suite 600, Berkeley, CA 94704. E-mail: trevor@eecs.berkeley.edu.*

Manuscript received 31 Mar. 2008; revised 30 Aug. 2008; accepted 25 Nov. 2008; published online 23 Dec. 2008.
Recommended for acceptance by A. Martinez.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-03-0185.
Digital Object Identifier no. 10.1109/TPAMI.2008.303.

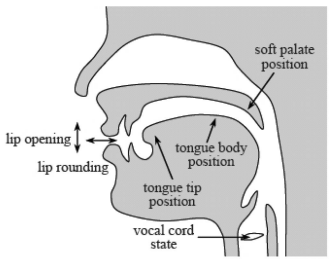

Fig. 1. A midsagittal diagram of the vocal tract, showing articulatory features in human speech production.

2.2 Articulatory Feature-Based Models in Speech Recognition

Articulatory models are an active area of work in automatic speech recognition (a survey can be found in [15]). Most of the work focuses on classifying articulatory feature values from the acoustic signal (e.g., [16], [17]). In this work, the classifier outputs are typically combined into either a phonetic likelihood (e.g., [17]) or a feature vector to be modeled with a Gaussian mixture distribution (e.g., [5]). These likelihoods or feature vectors are then used in a model in which each word or sub-word unit (e.g., a phoneme) is modeled as a hidden Markov model, identical to the standard (nonarticulatory) approach to speech recognition.

A smaller number of research efforts have been aimed at explicitly modeling the states of articulators, allowing them to stray from their target values or desynchronize [9], [20], [33]. More recently, the use of articulatory knowledge has also been proposed (though not implemented) for visual speech recognition [26]. This is not a straightforward application of acoustic speech recognition techniques because, unlike the acoustic signal, the visual signal provides direct evidence for certain articulatory gestures (e.g., lip closing) that are ambiguous in the acoustic signal.

2.3 Dynamic Bayesian Networks for Speech Recognition

All of our models are represented as dynamic Bayesian networks [7], [24]. To fix notation, Fig. 3 shows two frames of the DBN representing an HMM. Square and circular nodes denote discrete and continuous variables, respectively. Shaded nodes denote observed variables and unshaded ones denote hidden variables.

DBNs have been increasingly becoming popular in recent speech recognition research [3], [25], [38]. The HMM of Fig. 3 is a standard model for recognizing isolated words, consisting of a separate HMM for each word, which we refer to as a "whole-word" model. The variable s is the subword state; for a word with n phonemes, a typical choice is to allocate $3n$ states. HMMs for speech recognition are typically left-to-right, that is, it is assumed that there is a sequence of states that must be traversed from first to last. The variable o is typically a vector of acoustic observations, and its distribution is most frequently modeled as a mixture of Gaussians. This serves as our baseline for whole-word recognition experiments. For recognition of continuous word strings, additional variables would be needed to represent the language model (distribution over word strings).

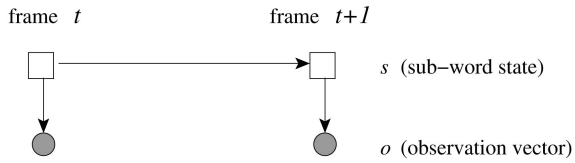

Fig. 3. A hidden Markov model represented as a DBN, which serves as our baseline single-stream whole-word model.

The model in Fig. 3 is useful for small-vocabulary recognition, where it is feasible to collect sufficient statistics for each word. For larger vocabularies, each word is typically broken into subword units such as phonemes, or visemes in the case of VSR, and each unit is represented with an HMM (typically with three states per phoneme/viseme). Observation models are shared among words with identical subword states. Fig. 4 shows such a model, which we refer to as a "dictionary-based" model. Here, p refers to the state within the phoneme-specific HMM, which is typically deterministic given the subword state s (unless the word has multiple pronunciations). Nodes with thick outlines denote variables that are deterministic given their parents. Now, s at time $t+1$ depends on p at time t since the transition probabilities are phoneme-dependent. The subword state s still encodes sequencing information (e.g., "first phone state in the word *Bob*"), while p encodes the actual phonetic unit (e.g., [b]).

Instead of a generative observation model, a discriminative phoneme or viseme classifier may be used. In this case, the DBN is identical except that the observation variable is replaced with a "virtual evidence" (or "soft evidence") node [1], [29], corresponding to a scaled likelihood estimate derived from postprocessed classifier outputs. For acoustic speech recognition, this is most often done using multilayer perceptron classifiers [23] (here we use support vector machines).

Thus far, we have described several single-stream models—whole-word, dictionary-based, and dictionary-based with virtual evidence—that serve as baseline, viseme-based recognizers. We next describe the proposed articulatory models. Both the baseline and proposed models can be trained via expectation-maximization (EM) [8], as is standard in speech recognition, and decoding (testing) can be done using standard DBN inference algorithms [3].

3 Proposed Approach

In the multistream models we propose, each articulatory feature is associated with a separate sequence of hidden states. We allow for possible differences between the target and actual positions of the articulators, as well as for possible asynchrony between the state sequences. The proposed class of models is an extension of the approach of [20], which introduced a general model of pronunciation variation using articulatory variables. We extend

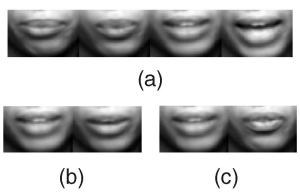

Fig. 2. Average feature appearance (lip opening: closed, narrow, mid, and wide; lip rounding and labiodental: no, yes; see Section 5.1): (a) lip opening, (b) lip rounding, and (c) labiodental.

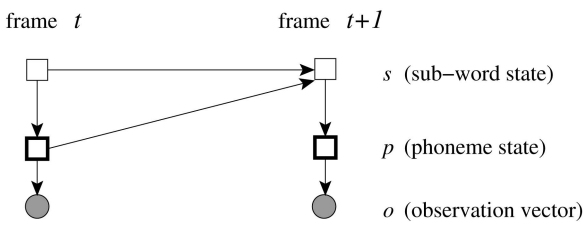

Fig. 4. A baseline single-stream dictionary model.

these models to the task of lipreading, including lipreading-specific articulatory features and observation models based on the outputs of articulatory classifiers.

We consider two ways of incorporating the outputs of articulatory feature classifiers. Classifier outputs may not match the values in the dictionary, for example, the target articulation may involve closed lips, but in the actual production, the lips may only be narrowed. We experiment with two ways to handle this discrepancy: 1) by explicitly modeling the distribution of actual given target articulation, treating the (postprocessed) classifier outputs as "virtual evidence" (see below) and 2) by modeling the distribution of classifier outputs conditioned directly on the state (a generative model). We also consider both "dictionary-based" and "whole-word" models, as described in Section 2.3. In dictionary-based models, we explicitly map the current subword state to the intended articulation and share observation models among words with identical subword states. Dictionary-based models may use either the generative or VE-based observation models described above. For whole-word models, only generative observation models can be used since we do not have classifiers for each possible subword state.

In the following sections, we present: the sets of articulatory features we use for visual recognition, support vector machine-based classifiers of these features, and a more formal description of our models in terms of dynamic Bayesian networks.

3.1 Articulatory Feature Sets

Various articulatory feature sets have been proposed for acoustic speech recognition. In dealing with the visual modality, we limit ourselves to modeling the visible articulators. As a start, we choose a feature set based on the one in [20]. Specifically, we use features associated with the lips, since they are always visible in the image. The features are: lip opening (LO, with values *closed, narrow, medium*, and *wide*), lip rounding (LR, with values *yes, no*), and the labiodental feature, corresponding to an /f/ or /v/ articulation (LD, with values *yes, no*). This ignores other articulators that might be discernible from the video, such as the tongue and teeth. We will later show that adding a fourth feature associated with the teeth can improve the three-feature model.

We do not, in general, have access to ground-truth articulatory feature labels for training data; obtaining them would require tracking the state of the vocal tract, which is very challenging. In the experimental section, we compare two ways of obtaining labels: manually, using a human labeler, and automatically, by mapping phoneme labels to feature values. (Phoneme labels, in turn, can be obtained automatically through forced alignment with the known transcription.) Some combinations of feature values can occur with manual labeling but not with automatic labeling, e.g., $\{\mathrm{LO} = narrow, \mathrm{LR} = yes$, and $\mathrm{LD} = yes\}$, if a speaker strays from the target articulation. Manual labels tend to be more consistent and, as we will show, produce better classifiers.

3.2 Articulatory Feature Classifiers

We convert the images to a sequence of appearance-based observation vectors, and classify each feature separately in each frame of the sequence using an SVM classifier. The DBN models require an estimate of the distribution of observations given each possible state of each articulatory feature, i.e., the likelihood.

We propose two ways of estimating the observation distributions. The first is to convert the output of each SVM to a probability. Given an observation vector $x \in \mathbb{R}^n$ and the unthresholded, raw decision value of a binary SVM $f(x)$, Platt [31] proposed to fit a sigmoidal function that maps from $f(x)$ to an estimate of the posterior probability of the positive class: $P(Y = 1|X = x) = (1 + e^{(af(x)+b)})^{-1}$, where a, b are the estimates using maximum likelihood. We use the multiclass extension of Platt's method described in [6] to produce posterior probabilities over all values of a feature $F, P(F = f|X = x)$ and map the posteriors to (scaled) likelihoods using $P(X = x|F = f) \propto P(F = f|X = x)/P(F = f)$. The scaled likelihoods are used as virtual evidence in the DBN (see Section 3.3).

The second method we consider is to use the decision value $f(x)$ directly as an observation. This can also be viewed as a nonlinear transform method for extracting visual observations.

We next describe the DBN-based models used to recognize words or phrases given the outputs of the articulatory classifiers.

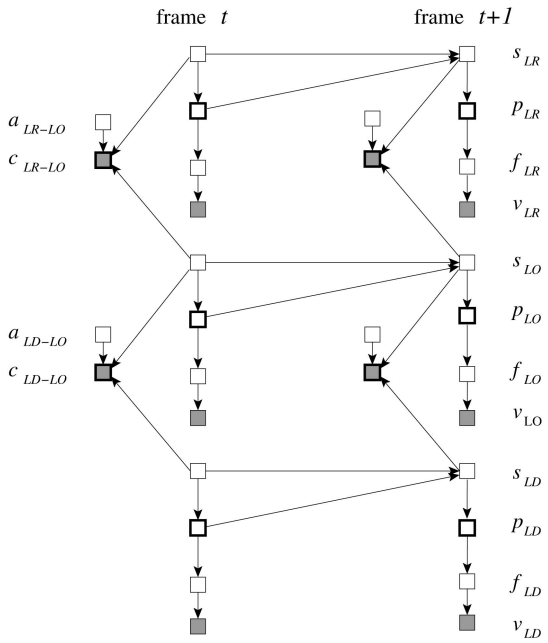

Fig. 5. Articulatory dictionary-based model with virtual evidence provided by articulatory classifiers. The variables are described in the text.

3.3 Dictionary Models

Fig. 5 shows the DBN for one word in a dictionary-based articulatory model, using an observation model based on virtual evidence from AF classifiers. This model uses the three articulatory features described previously: LR, LO, and LD. However, the structure can be straightforwardly generalized to an arbitrary set of articulators. The variables in the model are as follows:

- s_{LR}, s_{LO}, s_{LD}: The subword state corresponding to each articulatory feature, analogously to s in Fig. 4.
- p_{LR}, p_{LO}, p_{LD}: The phonetic state corresponding to the current subword state for each articulator, analogously to p in Fig. 4.
- f_{LR}, f_{LO}, f_{LD}: The value of each feature, with distribution given by $p(f_F|p_F)$.
- v_{LR}, v_{LO}, v_{LD}: Virtual evidence provided by classifiers for LR, LO, and LD. The distribution $p(v_F|f_F)$ is the postprocessed output of the classifier for feature F (see Section 3.2).
- $a_{F_1-F_2}$: These variables encode the asynchrony constraints. We define the degree of asynchrony between any two features F_1 and F_2 as $|s_{F_1} - s_{F_2}|$. The variable $a_{F_1-F_2}$ takes on values in $\{0, 1, 2, \ldots, m\}$, where m is the maximum state index in the current word. The distribution $p(a_{F_1-F_2})$ is the probability of the two features F_1 and F_2 desynchronizing by $a_{F_1-F_2}$ states.

- $c_{F_1-F_2}$: These binary variables, always observed with value 1, enforce the asynchrony constraints. The observed value is allowed only if $|s_{F_1} - s_{F_2}| = a_{F_1-F_2}$.[1]

We note that the virtual evidence variables in this model can also be replaced with observation vectors modeled with a Gaussian mixture distribution, analogously to the single-stream model of Fig. 4. We compare these variants in Section 5.3.

3.4 Dictionary Models Using Viseme Classifiers

The previous section described an articulatory feature-based dictionary model in which the evidence is provided by AF classifiers. It is also possible to combine an AF-based pronunciation model with viseme classifiers, by replacing the multiple variables v_{LR}, v_{LO}, v_{LD} with a single-viseme virtual evidence variable v_{VIS}, whose parents are f_{LR}, f_{LO}, f_{LD}. In such a model, the viseme classifier output is converted to virtual evidence $p(v_{VIS}|f_{LR}, f_{LO}, f_{LD})$ using a (many-to-one) mapping from features to visemes. If the streams in this model are constrained to be synchronous ($P(a_{F1-F2} = n) = 0$ for all $n > 0$), then it is equivalent to a baseline single-stream model, except that it has multiple (identical) transition probabilities.

3.5 Whole-Word/Small-Vocabulary Models

Analogously to the single-stream case, articulatory models may be dictionary-based or whole-word. Fig. 6 shows the DBN for a word in a whole-word recognizer. The structure is similar to that of the dictionary-based model of Section 3.3, but there are no explicit variables for the phonetic or articulatory state. For this reason, the outputs of the articulatory classifiers cannot be used directly as virtual evidence in this model. Instead, the classifier outputs are used as observations, o_F for each feature F, and the observation distribution $p(o_F|s_F)$ is modeled as a mixture of Gaussians. The advantage of a whole-word model is that it is not necessary to craft a good dictionary. In addition, no matter how good the dictionary, the same subword units may appear different in different words. For this reason, whole-word models are typically preferred for small-vocabulary tasks.

4 Experimental Details

We perform two sets of experiments using two audiovisual data sets of American English speech. The first set of experiments (Sections 5.1 and 5.2) is a medium-vocabulary word ranking task (with 1,793 words) and uses a subset of an existing corpus of read sentences. We refer to this data set as MED-V (for MEDIUM Vocabulary). The second set of experiments (see Section 5.3) is based on a small-vocabulary phrase recognition task (with 20 short phrases), and uses a corpus that we refer to as SMALL-V (for SMALL Vocabulary). This task is intended to simulate a more realistic stand-alone application.

The MED-V data set consists of words excised from continuously spoken sentences in the AVTIMIT corpus. AVTIMIT [13] consists of audiovisual recordings of phonetically balanced sentences from the TIMIT corpus [37] recorded in an office with controlled lighting and background. We use a subset consisting of 10 speakers reading the same sentences ("Set 02"). We use forced transcriptions of the audio to obtain word and phone boundaries, with the latter converted to canonic AF labels.

For the second set of experiments, we collected the new data set SMALL-V, consisting of two parts: 1) a first part similar to MED-V, containing about 2.5 minutes of video of two male native English speakers reading TIMIT sentences, used only to train AF classifiers and 2) a second part consisting of the 20 isolated short phrases, read by the same two speakers, with each phrase repeated three times. The phrases are shown in the Appendix, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.303, and are sample commands that can be used to control an audio system (e.g., "mute the volume"). The resulting 120 phrase recordings were used to test the AF classifiers and DBNs.

The front end of our system extracts the visual observations from the input video. Given a sequence of grayscale images, we first perform face detection and tracking, followed by lip detection and tracking, extraction of a region of interest (ROI) around the lips, and, finally, extraction of observations from the pixels inside the ROI [36]. The lip detection was initialized manually in the first frame of each video sequence.

We use a set of appearance features similar to ones that achieved state-of-the-art performance in prior work [32]. The extracted ROIs are resized to $height$ by $width$ pixels, and a discrete cosine transform (DCT) is applied to each image to obtain N^{DCT} coefficients. Finally, the dimensionality is further reduced using principal components analysis (PCA), with the top N^{PCA} coefficients retained. For the MED-V data set, $height = 16, width = 32$, $N^{DCT} = 512$, and $N^{PCA} = 75$. For SMALL-V, $height = 37, width = 54, N^{DCT} = 900$, and $N^{PCA} = 100$. The dimensionalities were chosen to give the best classification performance, using cross-validation on the training sets.

We use the LIBSVM [6] toolkit to implement the SVM classifiers, and the Graphical Models Toolkit (GMTK) [2] to implement the DBNs. We use a radial basis function (RBF) kernel in all SVMs, with the kernel width parameter and the error penalty parameter optimized by cross-validation on the training set. The sigmoidal function mapping SVM decision values to probabilities is also trained by cross-validation.

A final experimental detail concerns the relative weighting of different variables in the DBN. As is common in speech recognition, we use an exponential weight λ on the observation models. For example, in the model of Fig. 5, we use $p(v_F|f_F)^\lambda$ instead of $p(v_F|f_F)$ for each feature F. In the experiments, λ is tuned only for the baseline models, and the remaining models use the same λ; this gives an advantage to the baselines, but the AF-based models still outperform them. We also note that the results were roughly constant over a large range of λ in each experiment.

5 Experiments

We present two sets of experiments: The first (Sections 5.1 and 5.2) evaluates dictionary models on the medium-vocabulary word ranking task (using MED-V), while the second (Section 5.3) applies dictionary and whole-word models to the more practical scenario of short-phrase recognition (using SMALL-V). All of the experiments follow the outline given in Algorithm 1.

The main goals of the experiments are: 1) to compare the effects of using *AF-based versus viseme-based observation models* (classifiers) and 2) to compare the effects of using *synchronous versus asynchronous pronunciation models* (DBNs), independent of which classifiers are used. A synchronous pronunciation model is the special case of our models in which the features are constrained to be completely synchronous (i.e., $a_{F_1-F_2} = 0$). Using viseme classifiers with a synchronous pronunciation model results in a model almost identical to the conventional viseme-based HMM that has been used previously for VSR (e.g., [10]) and we consider this our baseline model.

1. We note that the structure could be simplified somewhat, as in [19], replacing $a_{F_1-F_2}$ and $c_{F_1-F_2}$ with a single variable; however, this simpler structure does not allow learning of the asynchrony probabilities via a straightforward application of expectation-maximization.

TABLE 1
Raw and Per-Class Classifier Accuracies (in Percent) for the Feature and Viseme SVMs on the MED-V Test Set

	Labels	LO	LR	LD	Viseme
per-class acc.	automatic	44 *(25)*	64 *(50)*	49 *(50)*	29 *(17)*
	manual	59 *(25)*	81 *(50)*	93 *(50)*	51 *(17)*
raw acc.	automatic	86 *(87)*	84 *(84)*	98 *(99)*	74 *(73)*
	manual	83 *(69)*	89 *(78)*	99 *(99)*	72 *(58)*

Algorithm 1. Overview of experimental approach
Input : A video stream of a speaker's face
Output : Probability of each word or phrase in the vocabulary
1 *Initialization: estimate parameters of SVM AF classifiers and DBN models on training data.*
2 Detect the face and track the mouth region.
3 Extract visual observations from each frame.
4 Classify AFs in each frame using a bank of SVMs.
5 Postprocess SVM outputs for use as observations in DBN.
6 For each word or phrase w, compute posterior probability of w using inference on corresponding DBN.

5.1 Single-Speaker Experiments on MED-V

To investigate how sensitive our models are to the quality of the training labels, we have manually labeled the 21 utterances read by one speaker (Speaker "03"). We compare models using manual labels to models using a mapping from phonemes to canonic feature values (see Section 3.1). The mapping used for canonic labels is shown in the Appendix. Fig. 2 shows the mean image for each feature value, reconstructed from its PCA coefficients.

5.1.1 Classifiers

We first compare the behavior of classifiers (of both features and visemes) trained on automatic and manual labels. The data is split into a training set of 10 utterances and a test set of 11 utterances. For the viseme classifier, there are six classes, consisting of those combinations of feature values that occur in the automatically labeled training set. When mapping manual AF labels to viseme labels, we use the same set of visemes as in the canonic labels, even though there are some combinations that are not allowed in the canonic mapping.[2]

We note that classifier accuracy measures are of limited utility; the true test is recognition performance using each set of classifiers, which we discuss in the next section. It is nevertheless useful to understand the behavior of the classifiers. Table 1 shows the raw classifier accuracies, (the percentage of frames classified correctly), as well as the average per-class accuracies (the percentage of correctly classified frames for each class, averaged over the N classes). Chance performance is given in parentheses ($\frac{100}{N}$ for per-class accuracy: the percentage of training frames corresponding to the most frequent label for raw accuracy). In these results, the correct labels are taken to be manual labels for the manual-train classifiers and automatic labels for the automatic-train classifiers. What we are testing is, therefore, the "learnability" of each labeling by this type of classifier.[3]

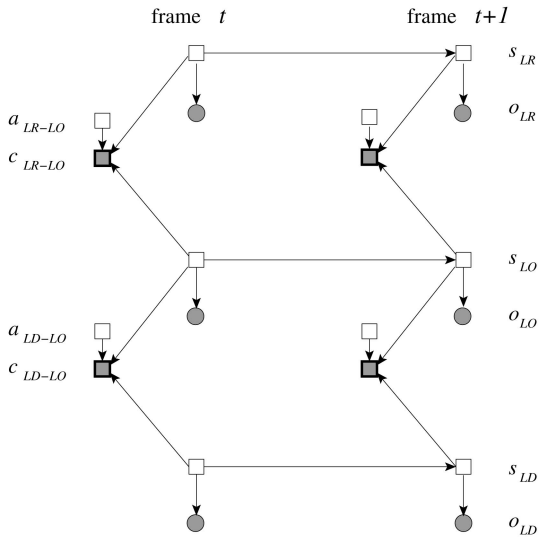

Fig. 6. Articulatory whole-word model.

We hypothesize that the canonic labels are less consistent or "noisier" and, therefore, expect lower accuracies from the automatic-train classifiers. This is, indeed, what we find: The manual-train classifiers have higher accuracies than the corresponding automatic-train ones, in most cases, by a wide margin. The two cases in which the manual-train classifiers have lower accuracies—the raw accuracies of the LO and viseme classifiers—correspond to much lower chance performance (and therefore, a more difficult task) for the manual labels. The LD classifier's per-class accuracy almost doubles, when manual labels are used, from chance to 93 percent. The poor performance of the automatic-train LD classifier is most likely due to the noisiness of the canonic labels, which often miss the appearance of labiodental closure. Note that the raw accuracies are not very informative in this case since the more frequent label, *no*, occurs 99 percent of the time.

Overall, the classifier performance results demonstrate that, as expected, the manual labels are more easily learned by the SVM classifiers. In the following section, we show that there are also large differences in recognizer performance.

5.1.2 Word Ranking Results

Here, the task is to recognize isolated words excised from continuous speech, an extremely difficult lipreading task even for humans. The overall error rate is, therefore, not a meaningful performance metric. Instead, we perform a word-ranking experiment. For each sequence 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 (where the highest rank is 1 and the lowest 1,793). We evaluate performance by the mean rank of the correct word (see [35] for distributions of correct word ranks). Since the vocabulary is too large to learn a separate model for each word, we use dictionary-based models for this task (Fig. 5).

We compare the effects of the following on the word ranking results: viseme versus AF classifiers, synchronous versus

2. We could add these combinations as extra viseme classes; however, they occur extremely rarely and would have insufficient training data. (This is, of course, one of the motivations for using articulatory features.)

3. We note that it is arguable what the ground-truth labels should be: If manual labels are used, it may not be a fair test of the automatic-label classifiers, and vice versa. The use of both labelings does not test true correctness, but does test the consistency of the labels.

TABLE 2
Mean Rank of the Correct Word for Several Recognizers on Speaker 03 from the MED-V Corpus

Classifier (mapping)	Mean rank, sync model	Mean rank, async model
Viseme (canonic)	232.8*	216.3 (0.2)
Feature (canonic)	232.7 (0.5)	200.1 (0.1)
Viseme (manual)	232.6*	225.8 (0.3)
Feature (manual)	135.8 (4e-05)	125.6 (5e-05)
Feature (oracle classifiers)	113.0 (4e-04)	109.7 (4e-04)

*Numbers in parentheses are p-values relative to viseme baselines (marked *).*

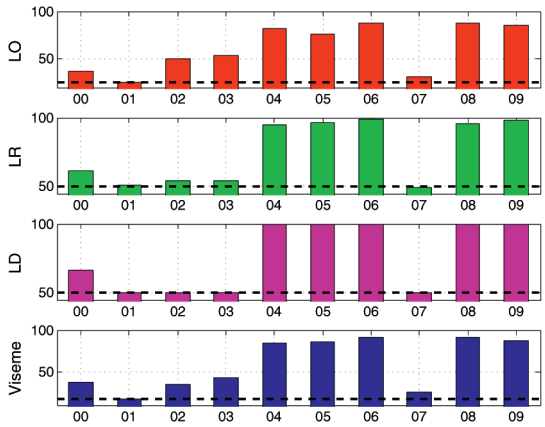

Fig. 7. Classifier per-class accuracy for each speaker in MED-V. The dashed lines show chance performance.

asynchronous pronunciation models, automatic versus manual labels, and "oracle" virtual evidence versus classifier outputs. The last of these is intended to test how well the recognizers could perform with "perfect" classifiers; this is explained further below. For this pilot experiment with a single speaker, the DBN parameters (the asynchrony and transition probabilities) are set by hand. (See Section 5.2 for multispeaker experiments in which all parameters are learned.) In the models with asynchrony, LR and LO are allowed to desynchronize by up to one state (one phoneme-sized unit), as are LO and LD.

Table 2 summarizes the mean rank of the correct word in a number of experimental conditions.[4] We make several observations. First, for the same pronunciation model and labels, using feature classifiers always improves the word ranks over using viseme classifiers (for example, from 232.6 to 135.8 using the synchronous model with manual labels). The advantage of articulatory features may stem from the fact that they each have fewer values and, therefore, more training data per class, than do visemes on the same training set. The second observation is that asynchronous pronunciation models consistently outperform synchronous ones, regardless of classifier choice, although this difference is not statistically significant in any one test.

Next, the automatic versus manual labels 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 retranscribe the training set.

To show how well the system could be expected to perform if we had ideal classifiers, we replaced the SVM virtual evidence with "likelihoods" derived from the manual transcriptions. In this "oracle" test, we assigned a very high likelihood (≈ 0.95) to feature values matching the transcriptions and the remaining likelihood to the incorrect feature values. We see that systems using the best classifiers (trained with manual labels) do not quite reach oracle performance, but are much closer to it than systems using the automatic labels for classifier training.

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). The differences between each synchronous model and the corresponding asynchronous model are not significant ($p \geq 0.1$ on this test set), but most feature-based models are significantly better than the baseline.

5.2 Multiple-Speaker Experiments on MED-V

Since we can easily produce automatic feature labels for more than one speaker, in this section, we perform canonic-label experiments on multiple speakers from MED-V. The experimental setup is identical to the canonic-label experiments in Section 5.1, except that the classifiers are trained on multiple speakers and the DBN parameters are learned from data. We experiment with a group of ten speakers, all of which read the same set of sentences.

5.2.1 Classifiers

For each speaker, we use the even-numbered utterances for training and the odd-numbered utterances for testing. Fig. 7 shows the per-class accuracies for the LO, LR, LF, and VIS classifiers. It is clear that, for speakers 00, 01, 02, 03, and 07, the classifiers perform very poorly, sometimes at chance levels. For the other five speakers, accuracy is quite good. Speakers in the former group have some tracking problems, especially 01 and 07, for which classification rates are the lowest and near chance. We exclude these two speakers from the following experiments, using the remaining eight. Sample tracking results for all 10 speakers are shown in the Appendix, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.303.

5.2.2 Word Ranking Results

We performed word ranking experiments on the eight speakers 00, 02, 03, 04, 05, 06, 08, and 09, using the same procedure as in Section 5.1.2. We find small but not statistically significant improvements for the proposed models. Over all words in the test set, the mean rank of the correct word is 123.3 using the baseline model (synchronous pronunciation model, viseme-based observation model). Switching to an asynchronous pronunciation model produces a mean rank of 121.5, a synchronous model with a feature-based observation model gives 119.0, and an asynchronous model with a feature-based observation model gives 118.5.

The three speakers with poorer classifier performance (00, 02, and 03) also have significantly worse ranking performance. If we consider only the speakers with average classification accuracy > 50 percent (speakers 04, 05, 06, 08, and 09), there is a statistically significant improvement in mean rank from using feature-based over viseme-based models (from 64.6 to 52.1, p-value 0.02), but not from using asynchronous models (from 64.6 to 66.2 for viseme-based models, 52.1 to 52.5 for feature-based models). On the speakers with poor classifiers (00, 02, and 03), feature classifiers do not help over viseme classifiers (from 210.5 to 218.4); asynchronous models improve over synchronous ones ($210.5 \rightarrow 203.7, 218.4 \rightarrow 216.7$), but not significantly so. We conclude that allowing asynchrony on this task does not make a statistically significant difference overall, perhaps because the labels were mapped from inherently synchronous phoneme labels. However, when the classifiers perform reasonably, AF-based models significantly outperform viseme-based ones.

4. The mean ranks are better than in earlier work [35] due to improvements to the classifiers.

TABLE 3
Average Percentage of Phrases Correctly Recognized with Various Models in the SMALL-V Task

dictionary			whole-word						
6V	3AF +VE	3AF	PCA	6V	8V	3AF	async +3AF	4AF	async +4AF
31	**34**	**43**	**69**	**52**	**58**	**64**	**66**	**78**	**79**

5.3 Phrase Recognition Experiments on SMALL-V

The word ranking experiments have shown that articulatory feature-based models have an advantage over conventional viseme-based ones, at least when the classifiers perform reasonably. Next, we investigate whether this advantage applies in the more practical setting of the small-vocabulary command-and-control recognition task of the SMALL-V data set. Since the vocabulary is small, we consider both dictionary and whole-word models. We also compare Gaussian observation models with the VE-based models used thus far. In addition, we experiment with adding a fourth articulatory feature (which is, in principle, not needed to distinguish the phrases). Finally, we compare our models against another baseline, which uses linear transform image features directly as observations.

5.3.1 Classifiers

In these experiments, we use only manual AF and viseme labels. A subset of the frames in the classifier training set was labeled manually with AF values. We train AF and viseme SVM classifiers, now using a one-versus-all strategy for multiclass SVMs, with six SVMs trained for the three AFs (four for LO, one for LR, and one for LD) and six SVMs for the six visemes. The average per-class accuracies of the classifiers on the test set are 79 percent for LO, 78 percent for LR, 57 percent for LD, and 63 percent for visemes.

5.3.2 Adding a New Feature

The feature set so far has included three features associated with the lips. Although these are sufficient to differentiate between the test phrases, we now add a fourth AF describing the position of the teeth (TP), with three values: *open* (used when there is a space between the teeth), *neutral* (used when there is no visible space), and *unknown* (used when the teeth are not visible). To determine whether using more features than the minimum needed to distinguish the vocabulary improves performance, we conduct experiments with both the original three-feature set and the new four-feature set. To make for a fair comparison between AF-based and viseme-based models, we also expand the viseme set to eight, corresponding to those combinations of AF values that occur in the training data.

5.3.3 Recognition Results

Table 3 summarizes the phrase recognition accuracies. Recall that each phrase was recorded three times; each recognition experiment is conducted three times, training on two repetitions of each phrase and testing on the remaining repetition. The table shows the average accuracies over the three trials.

The first column under "dictionary-based" in Table 3 corresponds to a viseme dictionary baseline model using a Gaussian distribution over SVM margin outputs for each viseme state. This model correctly recognizes only 30 percent of test phrases. The feature-based model with the three original AFs improves performance to 43 percent. The same model using a virtual evidence observation model (second column) has worse performance than the Gaussian one; we use the latter exclusively for the remaining experiments. Still, the accuracy is quite low; as we discuss next, we obtain much better results with whole-word models.

The right side of the table, under "whole-word," gives results for three baseline models: "6V," the viseme baseline corresponding to the 3-AF model; "8V," the viseme baseline for the 4-AF model; and "PCA," an HMM-based model with Gaussian observation distributions over raw PCA visual features (with the number of PCA coefficients set at 5 by cross-validation on the training data). Of these, the "PCA" baseline has the best performance. Comparing the 3-AF synchronous model to its 6-viseme equivalent, and the 4-AF synchronous model to its 8-viseme equivalent, we again find that feature-based models outperform viseme-based ones.

The "3-AF" and "async+3AF" columns show that AF-based models with the original 3-feature set outperform either viseme-based baseline, but not the PCA baseline. This is not surprising since the three lip features presumably carry less information than the full image on which the PCA coefficients are based. However, the "4-AF" column shows that, with four features, the AF-based model outperforms the PCA baseline. Finally, the last column gives the performance of an asynchronous 4-AF model, in which three pairs of streams were allowed to desynchronize by up to one state—LO and LR, LO and LD, and LO and TP. This model achieves the best overall performance, although the difference between it and the synchronous version is not statistically significant on this data set. Adding the fourth feature (TP) improved the accuracy of the synchronous DBN from 64 percent to 78 percent, and of the asynchronous DBN from 66 percent to 79 percent. 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.

6 Conclusion

We have presented a class of models that use multiple streams of articulatory features to model visual speech. This paper unifies previously presented work [35], [36] and includes additional model and experimental variants. In our approach, dynamic Bayesian networks are used to represent streams of hidden articulator states and to allow for asynchrony between them. A bank of support vector machine classifiers provides input to the DBN, in the form of either virtual evidence (probabilities) or raw margin outputs. We have presented experiments conducted on two visual speech tasks, a medium-vocabulary word-ranking task and a small-vocabulary phrase recognition task. The main findings are: 1) AF-based models outperform conventional single-stream viseme-based models on both tasks and 2) models allowing for asynchrony between streams usually outperform synchronous models, but not at a statistically significant level on our data. One reason for the improved performance with AF-based models may be that some visemes occur infrequently and, thus, have too little training data, while AF classifiers can utilize training data more efficiently. It is also possible that there are too many classes in the viseme-based multiclass SVMs, suggesting that investigations with alternative, inherently multiclass, classifiers may be useful.

A few additional aspects of the results are noteworthy and suggest possible directions for future work. We have found that, although not always visible in the image, articulators other than the lips can provide important information and help to improve performance (Section 5.3.3). Additional features, such as tongue position, may improve performance even further. We have also found (in Section 5.2) that classifier (and, therefore, recognition) performance varies widely across speakers; one area for future work is, therefore, the investigation of the causes of this variability. The word-ranking experiments with manual labels for a single speaker (Section 5.1.2) have underscored the importance of accurate training labels. Manually labeling large data sets may

not be feasible, but future work may include an iterative procedure to refine the automatic labels.

Finally, two important directions for future work are the application of the proposed models to more casual (conversational) speech styles and the addition of acoustic modality. The degree of asynchrony between articulators may be more pronounced in conversational speech and the types of models we have used were originally motivated by conversational speech phenomena [19]. In the case where we include the acoustic modality, our models allow the combination of audio and visual speech to be done at the articulatory feature level, as opposed to the phoneme/viseme level. Preliminary work in this direction has recently begun [12], [21]. We believe this may be a more appropriate model for audio-video fusion, since it accounts for the apparent asynchrony among the acoustic and visual signals [11] naturally via the mechanism of asynchronous AF streams.

ACKNOWLEDGMENTS

This work was supported by the US Defense Advanced Research Projects Agency (DARPA) and ITRI.

REFERENCES

[1] J. Bilmes, "On Soft Evidence in Bayesian Networks," Technical Report UWEETR-2004-00016, Electrical Eng. Dept., Univ. of Washington, 2004.
[2] J. Bilmes, "The Graphical Models Toolkit," http://ssli.ee.washington.edu/people/bilmes/gmtk/, 2009.
[3] J.A. Bilmes and C. Bartels, "Graphical Model Architectures for Speech Recognition," *IEEE Signal Processing Magazine,* vol. 22, no. 5, pp. 89-100, Sept. 2005.
[4] C.P. Browman and L. Goldstein, "Articulatory Phonology: An Overview," *Phonetica,* vol. 49, nos. 3/4, pp. 155-180, 1992.
[5] O. Cetin et al., "An Articulatory Feature-Based Tandem Approach and Factored Observation Modeling" *Proc. Int'l Conf. Acoustics, Speech, and Signal Proc.,* pp. IV-645-IV-648, Apr. 2007.
[6] C.-C. Chang and C.-J. Lin, "LIBSVM A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
[7] T. Dean and K. Kanazawa, "A Model for Reasoning About Persistence and Causation," *Computational Intelligence,* vol. 5, no. 2, pp. 142-150, Feb. 1989.
[8] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the *EM* Algorithm," *J. Royal Statistical Soc. Series B,* vol. 39, no. 1, pp. 1-38, 1977.
[9] L. Deng, G. Ramsay, and D. Sun, "Production Models as a Structural Basis for Automatic Speech Recognition," *Speech Comm.,* vol. 22, nos. 2/3, pp. 93-111, Aug. 1997.
[10] M. Gordan, C. Kotropoulos, and I. Pitas, "A Support Vector Machine-Based Dynamic Network for Visual Speech Recognition Applications," *EURASIP J. Applied Signal Processing,* vol. 2002, no. 11, pp. 1248-1259, 2002.
[11] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony Modeling for Audio-Visual Speech Recognition" *Proc. Human Language Technology Conf.,* p. 1006, Mar. 2002.
[12] M. Hasegawa-Johnson, K. Livescu, P. Lal, and K. Saenko, "Audiovisual Speech Recognition with Articulator Positions as Hidden Variables," *Proc. Int'l Congress on Phonetic Sciences,* Aug. 2007.
[13] T.J. Hazen, K. Saenko, C.-H. La, and J.R. Glass, "A Segment-Based Audio-Visual Speech Recognizer: Data Collection, Development, and Initial Experiments" *Proc. Int'l Conf. Multimodal Interfaces,* pp. 235-242, Oct. 2004.
[14] F. Jelinek, *Statistical Methods for Speech Recognition.* MIT Press, 1998.
[15] S. King et al., "Speech Production Knowledge in Automatic Speech Recognition," *J. Acoustical Soc. of Am.,* vol. 121, no. 2, pp. 723-742, Feb. 2007.
[16] S. King and P. Taylor, "Detection of Phonological Features in Continuous Speech Using Neural Networks," *Computer Speech and Language,* vol. 14, no. 4, pp. 333-353, Oct. 2000.
[17] K. Kirchhoff, G.A. Fink, and G. Sagerer, "Combining Acoustic and Articulatory Feature Information for Robust Speech Recognition," *Speech Comm.,* vol. 37, nos. 3/4, pp. 303-319, July 2002.
[18] G. Krone, B. Talle, A. Wichert, and G. Palm, "Neural Architectures for Sensor Fusion in Speech Recognition," *Proc. European Speech Comm. Assoc. Workshop Audio-Visual Speech Processing,* pp. 57-60, Sept. 1997.
[19] K. Livescu and J. Glass, "Feature-Based Pronunciation Modeling for Speech Recognition," *Proc. Human Language Technology Conf. North Am. Chapter of the Assoc. for Computational Linguistics,* May 2004.
[20] K. Livescu and J. Glass, "Feature-Based Pronunciation Modeling with Trainable Asynchrony Probabilities" *Proc. Int'l Conf. Spoken Language,* pp. 677-680, Oct. 2004.
[21] K. Livescu et al., "Articulatory Feature-Based Methods for Acoustic and Audio-Visual Speech Recognition: JHU Summer Workshop Final Report," Johns Hopkins Univ., Center for Language and Speech Processing, 2007.
[22] H. McGurk and J. McDonald, "Hearing Lips and Seeing Voices," *Nature,* vol. 264, no. 5588, pp. 746-748, Dec. 1976.
[23] N. Morgan and H. Bourlard, "Continuous Speech Recognition," *IEEE Signal Processing Magazine,* vol. 12, no. 3, pp. 24-42, May 1995.
[24] K. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," PhD dissertation, Computer Science Division, Univ. of California, 2002.
[25] A.V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A Coupled HMM for Audio-Visual Speech Recognition" *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 2013-2016, May 2002.
[26] P. Niyogi, E. Petajan, and J. Zhong, "A Feature Based Representation for Audio Visual Speech Recognition," *Proc. Int'l Conf. Auditory-Visual Speech Processing,* Aug. 1999.
[27] H. Nock and S. Young, "Modelling Asynchrony in Automatic Speech Recognition Using Loosely Coupled Hidden Markov Models," *Cognitive Science,* vol. 26, no. 3, pp. 283-301, May/June 2002.
[28] H. Pan, S.E. Levinson, T.S. Huang, and Z. Liang, "A Fused Hidden Markov Model with Application to Bimodal Speech Processing," *IEEE Trans. Signal Processing,* vol. 52, no. 3, pp. 573-581, Mar. 2004.
[29] J. Pearl, *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.
[30] E. Petajan, "Automatic Lipreading to Enhance Speech Recognition," *Proc. Global Telecomm. Conf.,* pp. 265-272, 1984.
[31] J. Platt, "Probabilities for SV Machines," *Advances in Large Margin Classifiers,* A.J. Smola, P.L. Bartlett, B. Schoelkopf, and D. Schuurmans, eds., pp. 61-73, MIT Press, 2000.
[32] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent Advances in the Automatic Recognition of Audiovisual Speech," *Proc. IEEE Int'l Conf. Image Processing,* vol. 91, no. 9, pp. 1306-1326, Sept. 2003.
[33] M. Richardson, J. Bilmes, and C. Diorio, "Hidden Articulator Markov Models for Speech Recognition," *Speech Comm.,* vol. 41, nos. 2/3, pp. 511-529, Oct. 2003.
[34] K. Saenko, T. Darrell, and J.R. Glass, "Articulatory Features for Robust Visual Speech Recognition" *Proc. Int'l Conf. Multimodal Interfaces,* pp. 152-158, Oct. 2004.
[35] K. Saenko, K. Livescu, J. Glass, and T. Darrell, "Production Domain Modeling of Pronunciation for Visual Speech Recognition," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. v/473-v/476, Mar. 2005.
[36] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual Speech Recognition with Loosely Synchronized Feature Streams" *Proc. Int'l Conf. Computer Vision,* pp. 1424-1431, Oct. 2005.
[37] V. Zue, S. Seneff, and J. Glass, "Speech Database Development: TIMIT and Beyond," *Speech Comm.,* vol. 9, no. 4, pp. 351-356, Aug. 1990.
[38] G. Zweig, "Speech Recognition Using Dynamic Bayesian Networks," PhD dissertation, Computer Science Division, Univ. of California, 1998.