Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition

Timothy J. Hazen, *Member, IEEE*

***Abstract*—This paper presents the design and evaluation of a speaker-independent audio-visual speech recognition (AVSR) system that utilizes a segment-based modeling strategy. The audio and visual feature streams are integrated using a segment-constrained hidden Markov model, which allows the visual classifier to process visual frames with a constrained amount of asynchrony relative to proposed acoustic segments. The core experiments in this paper investigate several different visual model structures, each of which provides a different means for defining the units of the visual classifier and the synchrony constraints between the audio and visual streams. Word recognition experiments are conducted on the AV-TIMIT corpus under variable additive noise conditions. Over varying acoustic signal-to-noise ratios, word error rate reductions between 14% and 60% are observed when integrating the visual information into the automatic speech recognition process.**

***Index Terms*—Audio-visual speech recognition, lip-reading, multimodal speech processing.**

I. Introduction

VISUAL information has been shown to be useful for improving the accuracy of speech recognition in both humans and machines [1]–[4]. These improvements are the result of the complementary nature of the audio and visual modalities. For example, many sounds that are confusable by ear are easily distinguishable by eye, such as n and m. The improvements from adding the visual modality are often more pronounced in noisy conditions where the audio signal-to-noise ratio (SNR) is reduced [5], [6].

When developing a speech recognition system that incorporates both the audio and visual modalities, a principled method for integrating the two streams of information must be designed. Because of the success of hidden Markov model (HMMs) in audio speech recognition, most audio-visual speech recognition (AVSR) systems extend HMM techniques to incorporate both modalities. In this paper we describe our efforts in developing an AVSR system which is built upon our existing segment-based speech recognizer [7]. This AVSR system incorporates information collected from visual measurements of the speaker's lip region using an audio-visual integration mechanism that we call a *segment-constrained HMM* [8].

Manuscript received February 4, 2005; revised April 14, 2005. This work was supported in part by ITRI and in part by an industrial consortium supporting the MIT Oxygen Alliance. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ananth Sankar.

The author is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: hazen@csail.mit.edu).

Digital Object Identifier 10.1109/TSA.2005.857572

In describing our system, we focus on three important issues we investigated during the development process. These are as follows.

- What is an appropriate set of visual units for representing the visual information and how does this set map into the underlying phonetic representation used within the speech recognition search engine?
- How important is it to model audio-visual asynchrony within the integration mechanism of the recognizer?
- How important is the use of an adaptive weighting scheme when integrating the audio and visual streams?

The remainder of the paper is organized as follows. In Section II we describe the AV-TIMIT corpus that we used for our experiments. In Section III we discuss the architecture of our AVSR system and the issues we addressed in its development. In Section IV we present our experimental results. Finally, we discuss the conclusions we have drawn from our results in Section V, and propose future work in Section VI.

II. AV-TIMIT Corpus

A. Data Collection

All of the experiments in this paper utilize the Audio-Visual TIMIT corpus collected at MIT [8]. This corpus was collected in January of 2003 in order to provide a large collection of audio-visual speech data from many speakers that was not phonetically constrained. At the time of its collection, many previously collected corpora were limited to only one subject [9] or to a small constrained task such as isolated letters [10], digits [11]–[13], or a short list of fixed phrases [12], [14], [15]. Only two of the A/V corpora that had been published in the literature (including English, French, German, and Japanese) contain both a large vocabulary and a significant number of speakers. The first was IBM's proprietary, 290-subject, large-vocabulary AV-ViaVoice database of approximately 50 h in duration [4]. This corpus is not currently publicly available. The second was the VidTIMIT database,[1] which has been made available for public use by Sanderson [16].

The main design goals for our AV-TIMIT corpus were: 1) continuous, phonetically balanced speech; 2) multiple speakers; 3) controlled office environment; and 4) high-resolution video. To achieve a phonetically balanced data set, speakers were asked to read from a list of TIMIT-SX sentences [17]. The recordings were conducted in a relatively quiet office with controlled lighting, background, and audio noise level. The audio was collected with a far-field array microphone located several feet in

[1]http://rsise.anu.edu.au/~conrad/vidtimit

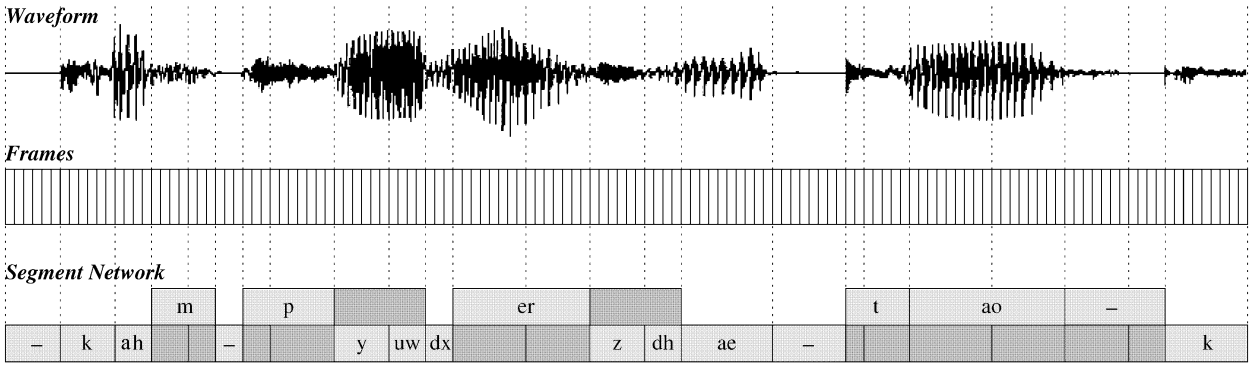

Fig. 1. Illustration of a search network created for segment-based recognition. The best segment path is highlighted in the segment network.

front of the speaker. Full details about the recording conditions and equipment can be found in our previous paper [8].

The full corpus contains 223 speakers (117 males and 106 females). A majority of the speakers came from our organization's community. All but 12 of the subjects were native speakers of English. Different ages and ethnicities were represented, as well as people with/without beards, glasses, and hats. Each speaker was asked to read 20 or 21 sentences. The first sentence of each round was identical for all speakers and was designed to familiarize the speakers with the data collection application. For the final five sentences in each round, extra side lighting was added in order to provide a second lighting condition for training and testing. In total, 23 different rounds of utterances were created that test subjects were rotated through. Each of the 23 rounds contained a unique set of TIMIT-SX sentences (from the full set of 453 sentences) with no overlap after the first sentence. Each round was spoken by at least nine different speakers. In total, the AV-TIMIT corpus contains 4597 utterances, making it roughly 10 times larger than the VidTIMIT corpus. There are 1793 unique words in the vocabulary of the corpus.

B. Experimental Data Sets

For our experiments we subdivided the AV-TIMIT corpus into three subsets: a training set, a development test set, and a final test set. The training set consisted of 3608 utterances from 185 speakers. To help constrain our initial experiments, we elected to train and evaluate using frontal lighting conditions only (and ignore the side-lighting condition). Under this constraint, the training set for the visual model is reduced to 2751 utterances (though the full 3608 utterances can still be used to train the acoustic model). For evaluation, our development test set contains 284 utterances from 19 speakers in the frontal lighting condition. The final test set contains 285 utterances from another 19 speakers in the frontal lighting condition. There is no overlap in speakers or sentences between any of the three data sets.

C. Additive Noise

Though the AV-TIMIT corpus was recorded in a relatively quiet office, the use of a far-field array microphone allowed for the range of the signal-to-noise ratio of the data to be fairly sizable. Estimates of the average SNR within individual utterances in the corpus varied from 15 dB to nearly 40 dB with an average SNR of approximately 25 dB. Because previous studies have found that the visual channel for speech recognition is especially helpful in noisier audio conditions, we have added additional noise to the AV-TIMIT audio to simulate variable SNR conditions. In particular, we have experimented with both white noise and babble noise data from the NOISEX database for our additive noise conditions [18]. However, in this paper, we only present our results using the babble noise condition. When adding noise, we have varied the average SNR from -10 dB to 20 dB.

III. System Development

A. Segment-Based Speech Recognition

Our audio-visual speech recognition approach builds upon our existing segment-based speech recognition system [7]. One of our recognizer's distinguishing characteristics is its use of segment-based networks for processing speech. Typical speech recognizers use measurements extracted from frames processed at a fixed rate (e.g., every 10 ms). In contrast, segment networks contain variable length segment hypotheses which each correspond to a potential phonetic unit. Our recognizer initially processes the speech using standard frame-based processing. Specifically, 14 Mel-Scale cepstral coefficients (MFCCs) are extracted from the acoustic waveform every 5 ms. However, unlike frame-based hidden Markov models (HMMs), our system hypothesizes points in time where salient acoustic landmarks might exist. These hypothesized landmarks are used to generate the network of possible segments.

The acoustic modeling component of the system scores feature vectors extracted from the segments and landmarks present in the segment network (rather than on individual frames). The search then forces a one-to-one mapping of segments to phonetic events. The end result of recognition is a path through the segment network in which all selected segments are contiguous in time and are assigned an appropriate phone. Fig. 1 illustrates an example segment network constructed for a waveform of the phrase "computers that talk," where the optimal path determined by the recognizer has been highlighted.

B. Visual Feature Extraction

In our system, appearance-based visual features are extracted from the mouth region of each image using the visual front-end component of the Intel AVCSR Toolkit.[2] Each image is first

[2]http://sourceforge.net/projects/opencvlibrary/

normalized for lighting variation using histogram equalization. Next, a principal components analysis (PCA) transform is applied, and the top 32 coefficients are retained as the feature vector for that image. In order to capture information about lip dynamics, three consecutive vectors are concatenated to create one 96-dimensional vector per frame.

C. Audio-Visual Integration

Once the audio and visual streams are processed and converted into feature-based observations, a means for integrating them must be devised. One method is to fuse their observation spaces such that they can be jointly processed within a single set of observation models. This approach, commonly referred to as *early integration* or *feature fusion*, simplifies the integration process by allowing a speech recognizer to use the same search representation for both audio-only and audio-visual recognition [3], [10]. On the negative side, the feature concatenation used in early integration may result in a high-dimensional data space, potentially making a large multimodal database necessary for robust statistical model training. Additionally, early integration makes it difficult to introduce a variety of desirable modeling techniques such as adaptive weighting of the audio and visual classification scores, or asynchronous processing of the audio and visual streams.

The more common approach is to perform *late integration* or *decision fusion*. In this approach, the audio and visual streams are independently processed and classified. Integration occurs at a higher level within the search mechanism of the recognizer. A variety of models have been proposed to perform late integration including multistream HMMs [19], coupled HMMs [20], and product HMMs [21]. One advantage to using a late integration approach is that the audio and visual classifiers are not required to be trained from exactly the same set of data (as alluded to in Section II-B).

In our system we have taken a late integration approach to combining the audio and visual information. As mentioned earlier, our system performs this integration using a segment-constrained HMM. This modeling approach is implemented with three primary steps. First, fixed-length video frames are mapped to hypothesized variable-length audio segments from the segment network. The mapping is performed such that any path through the segment network will incorporate each video frame exactly once. Second, each context-dependent phonetic segment defined in the acoustic model stream is mapped to a context-dependent segment-constrained visual HMM. Finally, the segment-constrained visual HMM uses a frame-based Viterbi search over video frames in the segment to generate a segment-based score for the visual model. Full details of this process can be found in our previous work [8]. The decision fusion between the audio and visual models is performed via a weighted linear combination of the segment-level scores generated from each model.

One advantage of using a late integration strategy is that the audio and visual streams are independently classified and hence can be adaptively weighted. In most systems, including ours, the relative weighting between the audio and visual streams is fixed within an utterance, but can be preset based on an estimate of the expected SNR [22], [23]. Using this scheme, Dupont, and Luettin observed that the optimal relative visual model weighting in their system exhibited a near linear relationship to the SNR [19]. Algorithms which dynamically alter the audio-visual weights for local regions within an utterance [4], [24], or allow different weightings for different word models [25], have also shown promise.

D. Visual Units Determination

When a late integration strategy is employed, the model structures devised for the audio and visual classifiers can each be created independently. As a result, there is a great deal of freedom in constructing the classifier used in the visual component and selecting its unit set. Typical audio-only speech recognition systems use phones as the basic units for speech recognition. When incorporating visual information into the process, one is confronted with the problem that the visual signal only provides partial information about the underlying sequence of phones. This is because, in general, one can only see a speaker's lips and jaw, while the other articulators (e.g., the tongue and the glottis) are typically hidden from sight. As a result, various sets of phones that are acoustically distinct may be visually indistinguishable. For example, the phones [b] and [p] differ from each other only in voicing, which is not visually apparent because it occurs at the glottis.

A system can take advantage of the visual similarities between different phonetic units by clustering these units together within the visual classifier. By increasing the number of training examples in each model class, clustering phonetic events into classes whose members are (supposedly) visually indistinguishable can improve the robustness of the visual models without (presumably) harming their discriminative ability.

The simplest form of clustering is to map the phones to visual units called *visemes*. Visemes are generally defined as the set of linguistically minimal units which are visually distinguishable [26], [27]. While many researchers have utilized the practice of clustering phonetic elements into viseme classes, there is no definite consensus about how the set of visemes is constituted [28]. A study of the literature reveals a variety of different viseme sets being used within AVSR systems [29]–[31]. Typically AVSR systems have used viseme sets containing between 12 and 20 different viseme classes. In our initial baseline system, we too manually crafted a set of 15 visemes [8]. To help us determine a useful set of visemic units for our AVSR system, we performed bottom-up clustering experiments using models created from phonetically labeled visual frames. The clustering of phones into visemes in our baseline system is shown in Table I.

Though the use of visemes in AVSR tasks is a common practice, a cursory examination of our set of visemes reveals several classes with obvious deficiencies. For example, within the rounded vowel viseme class (RV), it should be immediately obvious that the phones [aw], [ow], and [uw] should be easily distinguishable from the phones [w] and [oy] based on the dynamics of their lip rounding. The [aw], [ow], and [uw] all become more rounded as the phone progresses, while the [w] and [oy] start from a rounded position and become less rounded as the phone progresses. Within the same class the vowel [uh] and [ao] are not diphthongs or semivowels and likely retain relatively static rounding by comparison.

Supporting the notion that typical visemes sets are too restrictive, an early study by Finn and Montgomery found that their

TABLE I
MAPPING OF PHONETIC UNITS TO VISEMES FOR OUR EXPERIMENTS

Viseme Label	Phone Set
Sil	h# pau
OV	ax ih iy dx
BV	ah aa
FV	ae eh ay ey hh
RV	aw uh uw ow ao w oy
L	el l
R	er axr r
Y	y
LB	b p
LCl	bcl pcl m em
AlCl	s z epi tcl dcl n en
Pal	ch jh sh zh
SB	t d th dh g k
LFr	f v
VlCl	gcl kcl ng

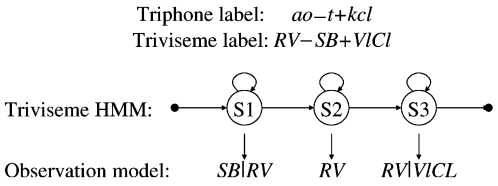

Fig. 2. Example segment-constrained triviseme HMM from our system.

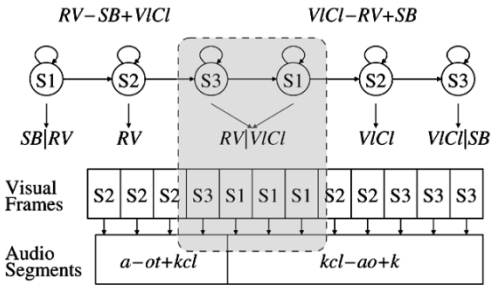

Fig. 3. Using segment-constrained HMMs to represent audio-visual asynchrony for a given audio segment sequence.

lip reading system could distinguish between different phonetic elements even when they belonged to the same viseme class [32]. Perhaps reflecting a feeling that typical viseme classes were too general, Bregler *et al.* used a much richer set of 42 visemes in their work [33]. Similarly, in the complementary field of audio-visual speech synthesis, Sannier *et al.* constructed a talking face synthesis system based on 44 viseme units [34].

An alternative to using visemes as the basic visual units is to retain the standard phonetic labels, but to use top-down decision tree clustering to create *tied* models (or HMM states) within the recognizer [35]. Decision-tree clustering can be performed using a variety of different constraints, and can be tailored to the specific topology of a given visual model. In our experiments, discussed later, we compare the use of our baseline viseme set against automatically generated clusters.

E. Audio-Visual Asynchrony

There is an inherent asynchrony between the visual and audio cues of speech. Speech is produced via the closely coordinated movement of several articulators. In some cases, such as the [b] burst release, the visual and audio cues are well synchronized. However, due to co-articulation effects and articulator inertia, the audio and visual cues may not be precisely synchronized at any given time. The articulators such as the lips and tongue sometimes move in anticipation of a phonetic event tens or even hundreds of milliseconds before the phone is actually produced [1]. In these cases, the visual evidence of the phonetic event may be evident before the acoustic evidence is produced.

To provide an example, consider the /g/to/m/ transition in the word *segment*. Typically, the/g/in this context is unreleased with only the voiced velar closure [gcl] being realized. Because this closure is produced with the tongue, the lips are free to form the closure for the [m] during the [gcl] segment. The labial closure for [m] does not affect the acoustics of the velar closure [gcl] because velar closures precede labial closures in the vocal tract. As a result, the visual evidence of the [m] can be present before its acoustic evidence.

Many AVSR approaches provide for asynchronous processing of the audio and visual streams [19]–[21], [36]. In all cases, a set of synchrony constraints are used to restrict the degree of allowable asynchrony. For example, phone synchronous HMM approaches allow fully asynchronous processing of the states of the audio and visual HMMs within a phone, but force synchronous state transitions at phone boundaries. Looser synchronization constraints can allow greater asynchrony at the expense of greater search complexity.

In our system we use audio segment boundaries as anchor points for our synchronization constraints. Fig. 2 shows an example visual HMM used for the triphone $ao-t+kcl$ (where [ao] is the current phone, [t] is the left context, and [kcl] is the right context). In this example, the triphone $ao-t+kcl$ is mapped to the triviseme $RV-SB+VlCl$ based on the viseme clustering in Table I. The figure also shows the mapping from each triviseme state to the label of the observation density function it uses. In this example, the left state of every triviseme HMM is mapped to a diviseme model (e.g., $SB \mid RV$) based on its left context, the middle state uses a context-independent model for that viseme (e.g., *RV*), and the right state is handled by the right side diviseme (e.g., $RV \mid VlCl$). Having established a model structure for aligning visual frames with an audio segment, the optimal frame alignment is determined using a Viterbi search over the frames in a segment.

When using the model structure employed above we can allow asynchronous processing across phonetic boundaries by applying appropriate state tying constraints. In our example, note that the third state $(S3)$ of a triviseme model will use the same diviseme observation density function as the first state $(S1)$ of the triviseme model for the following segment. This is illustrated in Fig. 3, where state 3 of the triviseme $RV-SB+VlCl$ and state 1 of the triviseme $VlCl-RV+SB$ both use the same diviseme observation model, $RV \mid VlCl$, for their output probability function. For any sequence of two triphone segments, the diviseme observation model capturing the visual transition between these audio segments is allowed to extend an arbitrary number of visual frames into either the

preceding or following audio segments. Our HMM topology also allows the system to skip any of the three states in a triviseme model. An example of this is shown in Fig. 3 where the first state of the triviseme $RV - SB + VlCl$ is skipped and the segment begins immediately with state 2 of the HMM.

The specific example in Fig. 3 demonstrates how our approach can model asynchrony during the [ao] to [kcl] transition in the word $talk$. In the visual signal there will be a smooth and gradual transition from the rounded vowel [ao] into the velar closure [kcl]. However, in the audio signal an abrupt acoustic transition occurs at the moment the velar closure is realized. The acoustic signal then contains silence during the velar closure [kcl] until it is released with a [k] burst. Thus, the transitional movement of the lips from a rounded to an unrounded position partially occurs during the velar closure when no acoustic change is evident. This asynchrony is handled in our model by allowing the visual frames assigned to $RV \mid VlCl$ viseme transition to straddle the acoustic segment boundary separating the [ao] acoustic segment from the [kcl] acoustic segment.

IV. Experiments and Results

A. Recognition Task

In our previous work we evaluated our audio-visual recognizer using a phonetic recognition paradigm [8]. In this work, we evaluate using a word recognition paradigm which allows us to incorporate a level of lexical constraint into the task that is lacking in the phonetic recognition paradigm.

Because the AV-TIMIT corpus is comprised of artificial hand-crafted sentences, the corpus does not lend itself well to standard word recognition language modeling techniques. For our experiments, we do not use a standard statistical language model but instead employ an unweighted word-pair grammar. In this grammar, a transition from one word to another can occur only if that particular word pair sequence occurs in at least one of the AV-TIMIT sentences. This grammar is very constraining because 1411 words of the 1793 word vocabulary in the corpus occur in only one of the 453 AV-TIMIT sentences. If converted into a statistical model (with all arcs leaving a node given a uniform likelihood), the perplexity of the word pair grammar is approximately 3.

From a word constraint perspective, our word recognition task is easier than the connected digit tasks which are commonly used for audio-visual recognition experiments. However, because the AV-TIMIT vocabulary is much larger, the acoustic and visual models in this task cannot be constructed in a word-dependent manner as they typically are in connected digit tasks. Thus, developing a robust set of models for both the acoustic and visual components is far more difficult because the range of contextual phonetic variability is far greater. It is important to investigate audio-visual speech recognition on larger vocabulary tasks in order to determine whether the techniques developed on small vocabulary tasks can be extended to more general speech recognition tasks.

B. Visual Model Sets

In this work we have generated three different visual model sets for our experiments. The first model set is based on the viseme set shown in Table I. All triphone labels in the recognizer are mapped to a three-state triviseme HMM model. Within the AV-TIMIT word recognition task, there are a total of 13 985 unique triphones that the recognizer may hypothesize. Each triphone label is mapped to one of 2690 different triviseme HMMs. As discussed in Section III-E and shown in Fig. 2, each state in a visual HMM is then mapped to a corresponding observation density function. The center states of each HMM are mapped to one of 15 different context-independent viseme models. The left and right states in each HMM are mapped to diviseme observation models for the transitions into the phone from the left context and out of the phone into the right context. There is a single observation density function for each of 203 possible diviseme transitions. Each density function is modeled with a mixture Gaussian model with a maximum of 50 components. In total the full model set uses 1915 total Gaussian components across the 218 observation models. This set of visual models will be referred to as the *Triviseme* set.

The second visual model set disposes of the notion of visemes and uses decision tree clustering to cluster the visual HMM states directly from the triphone labels. In this case, each triphone is mapped to a two-state HMM model. The left states of each model represents the left-side diphone of the triphone and the right state represents the right-side diphone of the triphone. Decision tree clustering is performed to cluster similar left-side diphones together based on their visual feature vectors. Decision tree clustering is also performed on the right-side diphones independently. Unlike the *Triviseme* unit set, this unit set does not share a single diphone model across the phonetic boundary. Instead, it forces the visual observation model to transition synchronously with the transition of the observation model of the acoustic stream. In total, the full triphone label set is clustered down to 147 left-side diphone clusters and 154 right-side diphone clusters used for observation modeling. These 301 observation models use a total of 2462 Gaussian components. From these observation model clusters, each triphone is mapped to a unique pair of left-side and right-side diphone clusters yielding a total 2928 visual HMM models. This visual model set will be referred to as the synchronous decision tree (or *Sync-DT*) model set.

The third visual model set also clusters the visual HMM states directly from the triphone labels. However, this model uses the same three-state HMM asynchronous modeling scheme used by the *Triviseme* model set. The middle HMM state remains a context-independent model, but in this case the decision tree clustering algorithm generated 44 different context-independent center states. Because the recognizer only uses 54 different phonetic labels, this means that most phonetic units are assigned their own context-independent visual model. The decision tree clustering also generated 197 diphone clusters. In total these 241 clusters are realized with observation models containing a total of 2332 Gaussian components. From these observation model clusters, each triphone is mapped to a unique trio of HMM state clusters resulting in a total of 10 125 visual HMM models. This visual model set will be referred to as the asynchronous decision tree (or *Async-DT*) model set.

In comparing the three models sets, the structure of the *Triviseme* set and the *Async-DT* set are very similar. The big difference is the added complexity of the *Async-DT* set

TABLE II
WORD ERROR RATES FOR DIFFERENT AVSR SYSTEMS TESTED ON SPEECH WITH ADDED BABBLE NOISE USING AUDIO MODELS TRAINED UNDER CLEAN AUDIO CONDITIONS. ERROR RATE REDUCTIONS ARE RELATIVE TO THE AUDIO-ONLY SYSTEM. UNDERLINED VALUES REPRESENT THE TOP PERFORMING SYSTEM FOR THAT PARTICULAR SNR

SNR	Word Error Rates (%) / Error Rate Reduction			
(dB)	Audio Only	Trivisemes	Sync-DT	Async-DT
clean	2.27 / -	1.54 / 32%	1.36 / 40%	0.91 / 60%
20	1.81 / -	1.81 / 0%	1.41 / 22%	1.09 / 40%
15	2.22 / -	2.04 / 8%	1.68 / 24%	1.77 / 20%
10	3.81 / -	3.22 / 15%	2.67 / 30%	2.90 / 24%
5	12.5 / -	9.11 / 27%	8.11 / 35%	9.47 / 24%
0	54.1 / -	39.2 / 28%	37.9 / 30%	37.1 / 31%
-5	103.8 / -	90.5 / 13%	87.8 / 15%	89.3 / 14%
average	25.8 / -	21.1 / 18%	20.1 / 22%	20.4 / 21%
visual-only	N/A	96.3 / N/A	95.4 / N/A	96.6 / N/A

TABLE III
WORD ERROR RATES FOR DIFFERENT AVSR SYSTEMS TESTED ON SPEECH WITH ADDED BABBLE NOISE USING AUDIO MODELS TRAINED UNDER MATCHED NOISE CONDITIONS. ERROR RATE REDUCTIONS ARE RELATIVE TO THE AUDIO-ONLY SYSTEM. UNDERLINED VALUES REPRESENT THE TOP PERFORMING SYSTEM FOR THAT PARTICULAR SNR

SNR	Word Error Rates (%) / Error Rate Reduction			
(dB)	Audio Only	Trivisemes	Sync-DT	Async-DT
clean	2.27 / -	1.54 / 32%	1.36 / 40%	0.91 / 60%
20	1.90 / -	1.45 / 24%	0.95 / 50%	1.04 / 44%
15	1.99 / -	1.50 / 25%	1.00 / 50%	1.00 / 50%
10	2.49 / -	2.18 / 13%	1.77 / 29%	1.41 / 44%
5	5.26 / -	4.90 / 7%	4.26 / 19%	4.26 / 19%
0	16.5 / -	14.1 / 15%	14.6 / 12%	12.3 / 26%
-5	62.4 / -	48.7 / 22%	46.3 / 26%	46.4 / 26%
-10	90.9 / -	79.0 / 13%	78.9 / 13%	78.3 / 14%
average	22.9 / -	19.2 / 16%	18.6 / 19%	18.2 / 21%

resulting from use of 44 context-independent center HMM states instead of the 15 context-independent viseme states used by the *Triviseme*. In comparing the *Async-DT* set with the *Sync-DT* set, there are two big differences. First, the *Sync-DT* set separates each transitional diphone into two models, one for the left-side of the phonetic boundary and one for the right-side of the boundary. This increases the number of diphone classes by more than 50%. However, the *Sync-DT* set also eliminates the use of a context-independent center state in each HMM which reduces its complexity. We eliminated the center states in order to make the complexity of the *Sync-DT* and *Async-DT* models as similar as possible (for the purpose of providing a fairer comparison of the models). Overall, the *Async-DT* model set has only 5% fewer Gaussian components (spread over 20% fewer observations models) than the *Sync-DT* model set.

C. Visual Modeling Results

Our final results on the AV-TIMIT test set when adding varying amounts of babble noise to the audio are shown in Tables II and III. Table II shows the results when the acoustic models are trained using only the clean speech condition. Table III shows the results when the acoustic models are trained using the same noise condition as the test data. These two tables represent the best case and worst case scenarios for speech recognition under variable unseen noise conditions. In practice, a system would likely use some form of noise compensation or acoustic model adaptation to account for the environmental noise condition, and the results would be expected to fall somewhere between the mismatched condition results in Table II and the matched condition results in Table III. All results in these tables use a fixed set of audio-visual fusion weights, as will be discussed in Section IV-D. For reference, the results when using only the visual information are also shown in Table II.

Within the two tables the best performing visual model set for each SNR is underlined. The results show that the visual information can significantly improve the speech recognition accuracy even under high SNR conditions. Using the *Async-DT* models over variable acoustic SNR conditions, relative reductions in word error rate of between 14% and 60% are obtained. The results also show that the automatically clustered models sets *Sync-DT* and *Async-DT* generally outperform the manually crafted *Triviseme* model set. There is not a single SNR value in either the mismatched or matched model conditions in which the *Triviseme* model set is the top performer. A discussion of these results will follow in Section V.

D. Stream Weighting Results

In this paper we utilize a static weighting scheme where the audio visual classifier weights are fixed for the duration of the utterance. However, the weights can be preset based on the expected SNR of the audio stream. The AV-TIMIT development test set is used to determine the optimal weighting factors for the different recognizer components. In total, there are four weights to tune (the audio boundary model weight, the audio segment model weight, the visual model weight, and a word transition weight for controlling the tradeoff between word insertions and deletions). We fixed the audio boundary model weight at a value of 1 and then optimized the audio segment model weight and word transition weight using an audio-only recognition paradigm under various different noise conditions. Over all of the conditions we tested, the optimal weights for the audio-only recognizer remained virtually identical. As a result, we elected to set these weights to fixed values that were approximately optimal on the development set for all future experiments.

When we added the visual stream to the task, we optimized the visual model weight for each SNR level relative to the fixed audio model weights. As expected, there is a correlation between the SNR and the optimal weighting factor. Fig. 4 shows the optimal visual model weight values for different SNR values in babble noise on the development test set using the *Triviseme* visual models. A linear best fit approximation of the optimal weights is also shown. As can be seen in the figure, the optimal visual weight nearly doubles as the SNR is varied from the clean condition to -10 dB.

Despite the correlation between the visual weight and the SNR observed in Fig. 4, we also discovered that the visual weight at any given SNR can be varied significantly from

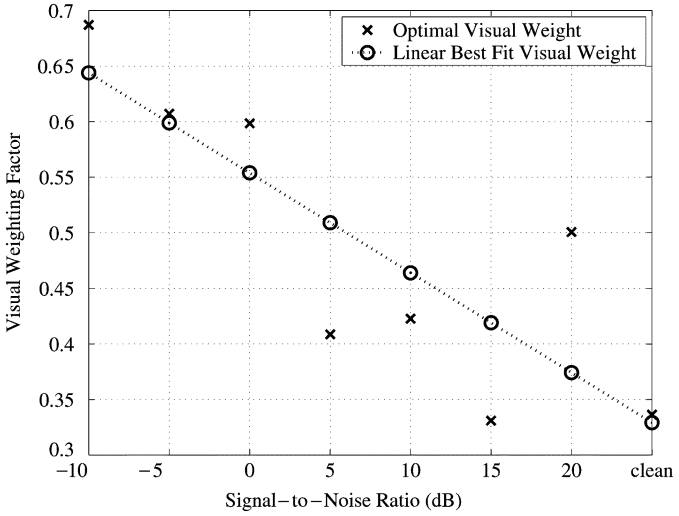

Fig. 4. Optimal weighting of visual stream for speech recognition using the *Triviseme* visual model tested on the AV-TIMIT development data over varying SNR levels. The linear best fit weighting function is also shown.

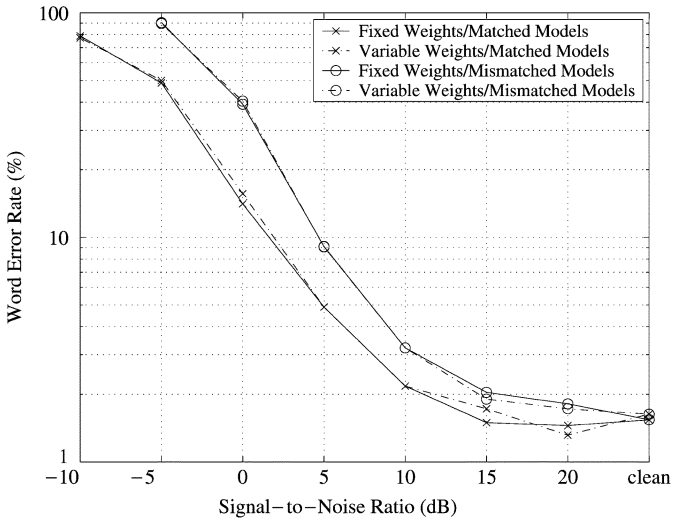

Fig. 5. AVSR performance of *Triviseme* visual models when using matched versus mismatched acoustic models, and when using a fixed visual model weight versus a variable weight this is adapted to the SNR.

the optimal weight without causing a severe degradation in recognition accuracy. This observation led us to believe that there may be little difference in performance on unseen data between setting the visual weight to a single fixed value for all SNRs versus using the variable weights from the linear best fit function shown in Fig. 4. To verify this we tested the two weighting schemes on the AV-TIMIT test data using the *Triviseme* visual models. In this test, the single fixed visual weight was set to the average optimal weight observed in Fig. 4. The results of this test are shown in Fig. 5 where we can observe that there is no advantage in our system to using a variable weighting scheme based on the SNR. This figure also shows the significant difference in performance between the matched and mismatched acoustic model conditions. This shows that adaption of the acoustic models to the current noise condition is far more important than adaptation of the audio-visual weighting factor.

V. Conclusion

The experiments presented in this paper have yielded several results that are worth comment. First, although Potamianos *et al.* claim that using "different speech classes in the audio- and visual-only components complicates audiovisual integration" [4], we have found no added complications in our system from taking such an approach. No alterations in the audio components or primary search mechanism needed to be made when adding the visual components. Further more, allowing different speech classes among the different components lets each component be independently optimized without forcing any constraints upon it from other components.

Second, it is our conclusion that visemes, though useful for expository and educational purposes, are a suboptimal representation when used within the model structure of a speech recognizer. Our experiments indicate that it is better to retain the phonetic labels within the visual model structure and allow data-driven clustering techniques to perform any label classing that may be necessary.

Next, our results do not show that asynchronous modeling of the audio and visual streams is better than synchronous modeling. Tables II and III fail to show an advantage for either the asynchronous *Async-DT* model or the synchronous *Sync-DT* model. Although more study of this issue is needed, we believe that the combination of context dependent modeling with a sufficiently large enough amount of training data may allow a synchronous multistream model the ability to implicitly learn the effects of articulator asynchrony without requiring any explicit method for modeling it.

Next, though we did observe a correlation between the SNR of an utterance and the optimal audio-visual fusion weighting, we did not observe a significant difference in performance on unseen data when an adaptive weighting scheme based on the SNR was used in place of a single pre-set weighting. This is partially explained by the fact that the SNR levels in our experiments represent the *average* SNR. Thus, even in noisy conditions, there will be some regions of the signal where the SNR is considerably higher than the average and other regions where it is considerably lower than the average. This variance of the local SNR across an utterance could explain why a wide range of weighting ratios between the audio and visual streams yielded similar results in our system. Further study of this issue is needed in order to determine if a locally adaptive weighting scheme can give further improvements.

Finally, our experiments have demonstrated the difficulty of automatic speech recognition for larger vocabulary tasks based only on visual information. While lip-reading systems have typically been able to achieve speaker-independent error rates of 40% on digit strings recorded under studio conditions [3], [19], large vocabulary recognition using only visual information remains an extremely difficult problem in need of further study. Like the large vocabulary system of Potamianos *et al.* [3], our AV-TIMIT system also has a word error rate greater than 90% when using only visual information. The high error rate of the visual-only system may also partially explain why very little is gained from adapting the stream weights of the audio and visual components of the system as the SNR gets worse. Because there are no SNR conditions in which the visual system is significantly better than the audio system, there is little to be gained from increasing the relative weight of the visual model in low-SNR conditions.

VI. Future Work

In this work we focused on the issues of visual-model unit selection, audio-visual asynchrony, and audio-visual stream weighting. However, there are still many interesting problems left to investigate. One open question is the effect of lexical and language model constraint within AVSR systems. Because the visual signal typically does not provide as much phonetic disambiguation as the audio signal, strong lexical constraint may be needed in order to take full advantage of the visual information. In future work we plan to explore the effect of language model perplexity on the speech recognition performance improvements observed from the addition of the visual signal. Also, because our segment-based approach has significant differences with the more traditional coupled-HMM approaches used by others, experiments directly comparing these approaches could be illuminating.

Acknowledgment

The author would like to thank C.-H. La, K. Saenko, and J. Glass for their efforts during the collection of the AV-TIMIT corpus and the development of the baseline recognition system used in this work.

References

[1] C. Benoit, "The intrinsic bimodality of speech communication and the synthesis of talking faces," in *The Structure of Multimodal Dialogue II*, M. M. Taylor, F. Nel, and D. Bouwhuis, Eds. Amsterdam, The Netherlands: John Benjamins, 2000, pp. 485–502.

[2] C. Chibelushi, F. Deravi, and J. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 23–37, Mar. 2002.

[3] G. Potamianos, C. Neti, G. Iyengar, and E. Helmuth, "Large-vocabulary audio-visual speech recognition by machines and humans," in *Proc. Eurospeech*, Aalborg, Denmark, Sep. 2001, pp. 1293–1296.

[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[5] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, no. 2, pp. 212–215, Mar. 1954.

[6] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1293–1296.

[7] J. Glass, "A probabilistic framework for segment-based speech recognition," *Comput. Speech Lang.*, vol. 17, no. 2–3, pp. 137–152, Apr./Jul. 2003.

[8] T. Hazen, K. Saenko, C. La, and J. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proc. Int. Conf. Multimodal Interfaces*, PA, Oct. 2004.

[9] M. Chan, Y. Zhang, and T. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in *Proc. Workshop on Multimedia Signal Processing*, Redondo Beach, CA, 1998, pp. 65–70.

[10] I. Matthews, J. Bangham, and S. Cox, "Audio-visual speech recognition using multiscale nonlinear image decomposition," in *Proc. Int. Conf. Spoken Language Processing*, Philadelphia, PA, 1996, pp. 38–41.

[11] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," in *Proc. Audio- Video-based Biometric Person Authentication Workshop*, 1997.

[12] S. Chu and T. Huang, "Bimodal speech recognition using coupled hidden Markov models," in *Proc. Int. Conf. Spoken Language Processing*, vol. II, Beijing, China, Oct. 2000.

[13] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multi-modal human-computer interface research," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Orlando, FL, May 2002, pp. 2017–2020.

[14] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "XM2VTSDB: The extended M2VTS database," in *Audio- and Video-based Biometric Person Authentication, AVBPA'99*, Washington, DC, Mar. 1999, pp. 72–77. 16 IDIAP-RR 99-02.

[15] S. Chu and T. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Orlando, FL, May 2002, pp. 2009–2012.

[16] C. Sanderson, "The VidTIMIT database," in *IDIAP Communication 02–06*. Martigny, Switzerland: IDIAP, Aug. 2002.

[17] V. Zue, S. Seneff, and J. Glass, "Speech database development: TIMIT and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, 1990.

[18] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[19] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[20] A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Orlando, FL, May 2002, pp. 2013–2016.

[21] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. Human Language Technology Conf.*, San Diego, CA, Mar. 2002, pp. 1–6.

[22] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, "Toward unrestricted lip reading," *Int. J. Pattern Recognit. Artif. Intell.*, no. 14, pp. 571–585, Aug. 2000.

[23] M. Heckmann, F. Berthommier, and K. Kroschel, "Optimal weighting of posteriors for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, Salt Lake City, UT, May 2001, pp. 161–164.

[24] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, Salt Lake City, UT, May 2001, pp. 165–168.

[25] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Montreal, QC, Canada, May 2004, pp. 857–860.

[26] C. Fisher, "Confusions among visually perceived consonants," *J. Speech Hearing Res.*, vol. 11, pp. 796–804, 1968.

[27] K. Berger, *Speechreading: Principles and Methods*. Baltimore, MD: National Educational, 1972.

[28] T. Chen and R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.

[29] P. Silsbee, "Sensory integration in audiovisual automatic speech recognition," in *Proce. 28th Annual Asilomar Conf. Signals, Systems, and Computers*, vol. 1, Pacific Grove, CA, Oct./Nov. 1994, pp. 561–565.

[30] A. Rogozan, "Discriminative learning of visual data for audiovisual speech recognition," *Int. J. Artif. Intell. Tools*, vol. 8, no. 1, pp. 43–52, Mar. 1999.

[31] C. Neti *et al.*, "Audio-visual speech recognition," Center Lang. Speech Process., The Johns Hopkins Univ., Baltimore, MD, Tech. Rep., 2000.

[32] K. Finn and A. Montgomery, "Automatic optically-based recognition of speech," *Pattern Recognit. Lett.*, vol. 8, no. 3, pp. 159–164, Oct. 1988.

[33] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, Minneapolis, MN, Apr. 1993, pp. 557–560.

[34] G. Sannier, S. Balcisoy, N. Magnenat-Thalmann, and D. Thalmann, "VHD: A system for directing real-time virtual actors," *Vis. Comput.*, vol. 15, no. 7/8, pp. 320–329, Nov. 1999.

[35] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, Princeton, NJ, Mar. 1994, pp. 307–312.

[36] S. Bengio, "An asynchronous hidden Markov model for audio-visual speech recognition," in *Advances in Neural Information Processing Systems, NIPS 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 1237–1244.

Timothy J. Hazen (M'04) received the S.B., S.M., and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1991, 1993, and 1998, respectively.

He is a Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory where he works in the areas of automatic speech recognition, automatic person identification, multimodal speech processing, and conversational speech systems.

Dr. Hazen is an Associate Editor of IEEE Transactions on Speech and Audio Processing.