A Segment-Based Audio-Visual Speech Recognizer: Data Collection, Development, and Initial Experiments

Timothy J. Hazen, Kate Saenko, Chia-Hao La, and James R. Glass
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street
Cambridge, Massachusetts, USA
{hazen,saenko,chiahao,jrg}@sls.csail.mit.edu

ABSTRACT

This paper presents the development and evaluation of a speaker-independent audio-visual speech recognition (AVSR) system that utilizes a segment-based modeling strategy. To support this research, we have collected a new video corpus, called Audio-Visual TIMIT (AV-TIMIT), which consists of 4 total hours of read speech collected from 223 different speakers. This new corpus was used to evaluate our new AVSR system which incorporates a novel audio-visual integration scheme using segment-constrained Hidden Markov Models (HMMs). Preliminary experiments have demonstrated improvements in phonetic recognition performance when incorporating visual information into the speech recognition process.

Categories and Subject Descriptors

I.2.M [**Artificial Intelligence**]: Miscellaneous

General Terms

Algorithms, Design, Experimentation.

Keywords

Audio-visual speech recognition, audio-visual corpora.

1. INTRODUCTION

Visual information has been shown to be useful for improving the accuracy of speech recognition in both humans and machines [1, 15]. These improvements are the result of the complementary nature of the visual and aural 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 also more pronounced in noisy conditions where the audio signal-to-noise ratio is reduced [1, 17].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.
Copyright 2004 ACM 1-58113-954-3/04/0010 ...$5.00.

In this paper, we describe our efforts in developing an audio-visual speech recognition (AVSR) system. It is hoped that this speech recognition technology can be deployed in systems located in potentially noisy environments where visual monitoring of the user is possible. These locales include automobiles, public kiosks, and offices. Our new AVSR system is built upon our existing segment-based speech recognizer [5]. This AVSR system incorporates information collected from visual measurements of the speaker's lip region using a novel audio-visual integration mechanism which we call a *segment-constrained Hidden Markov Model (HMM).* Our AVSR system is described in detail in Section 3.

To help develop and evaluate our new AVSR system, we have collected a corpus containing 4 total hours of audio-visual speech data collected from 223 different speakers. The corpus contains read-speech utterances of TIMIT SX sentences [22]. The video contains frontal views of the speaker's head in front of a solid blue backdrop under two different lighting conditions. This corpus is described in Section 2. Phonetic recognition experiments using this corpus are reported in Section 4. We summarize our efforts in Section 5 and propose future work in Section 6.

2. DATA COLLECTION

2.1 Existing Audio-Visual Corpora

A variety of audio-visual corpora have been created by researchers in order to obtain experimental results for specific tasks. Corpora available for public use have originated primarily from universities, and tend to be less extensive than the ones collected by private research labs. Many of the former contain recordings of only one subject, e.g. [2]. Even those with multiple subjects are usually limited to small tasks, such as isolated letters [12], digits [3, 16], or a short list of fixed phrases [14]. Only two of the A/V corpora published in the literature (including English, French, German and Japanese) contain both a large vocabulary and a significant number of subjects. The first is IBM's proprietary, 290-subject, large-vocabulary AV-ViaVoice database of approximately 50 hours in duration [8]. The second is the VidTIMIT database [19], which was recently made available by the Linguistic Data Consortium (LDC). It consists of 43 subjects reciting 10 TIMIT sentences each, and has been used in multi-modal person verification research [20]. This corpus was not yet publicly available at the onset of our own data collection effort.

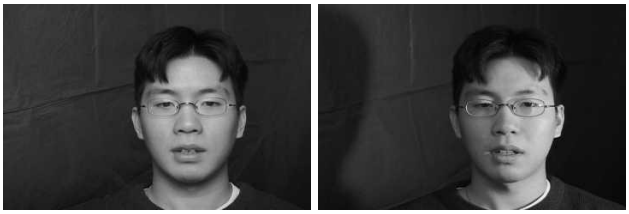

Figure 1: Sample still shots from the AV-TIMIT corpus showing the two lighting conditions used in the recording.

2.2 AV-TIMIT Data Collection

To provide an initial corpus for our research in audio-visual speech recognition we collected a new corpus of video recordings called the Audio-Visual TIMIT (AV-TIMIT) corpus. It contains read speech and was recorded in a relatively quiet office with controlled lighting, background and audio noise level. The main design goals for this corpus were: 1) continuous, phonetically balanced speech, 2) multiple speakers, 3) controlled office environment and 4) high resolution video. The following sections will describe each aspect of the data collection in detail.

2.2.1 Linguistic Content

Because of size and linguistic flexibility requirements, we decided to create a corpus of phonetically rich and balanced speech. We used the 450 TIMIT-SX sentences originally designed to provide a good coverage of phonetic contexts of the English language in as few words as possible [22]. Each speaker was asked to read 20 sentences. The first sentence was the same for all speakers, and is intended to allow them to become accustomed to the recording process. The other 19 sentences differed for each round. In total, 23 different rounds of utterances were created that test subjects were rotated through. Each of the 23 rounds of utterances was spoken by at least nine different speakers.

2.2.2 Recording Process

Recording was completed during the course of one week. The hardware setup included a desktop PC, a GN Netcom voice array microphone situated behind the keyboard, and a high-quality SONY DCR-VX2000 video camcorder. The camera was mounted on a tripod behind the computer display to record a frontal view of each subject. A blue curtain was hung behind the chair to reduce image background noise; however, users were not told to restrict their movements. The audio quality was generally clean, but the microphone did pick up some noise from a computer fan. The average signal-to-noise ratio within individual utterances was approximately 25 dB, with a standard deviation of 4.5 dB.

After being seated in front of the computer, the user was instructed to press and hold the "Record" button on the interface while reading each prompted utterance from the screen. Upon button release, the program echoed the recorded waveform back, so that the user could hear his/her own recording. To help ensure that the speech matched the orthographic transcription, an observer was present in the room to ask the user to re-record a sentence if necessary. For the last five sentences, extra side lighting was added in order to simulate different lighting conditions (see Figure 1).

Figure 2: Examples of tracked mouth regions from the AV-TIMIT corpus. The bottom row shows tracking failures.

2.2.3 Database Format

Full color video was stored in uncompressed digital video (DV) AVI format at 30 frames per second and 720x480 resolution. In addition to the audio track contained in the video files, the audio was also saved into separate WAV files, sampled at 16 KHz. The total database duration is approximately 4 hours.

2.2.4 Demographics

The majority of volunteers came from our organization's community. The final audio-visual TIMIT corpus contained 223 speakers, of which 117 were male and 106 were female. 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.

2.3 AV-TIMIT Annotation

2.3.1 Audio Processing

Time-aligned phonetic transcriptions of the data were created automatically using a word recognition system configured for forced-path alignment. This recognizer allowed multiple phonetic pronunciations of the spoken words. Alternate pronunciation paths could result either from a set of phonological variation rules or from alternate phonemic pronunciations specified in a lexical pronunciation dictionary.

The acoustic models for the forced-path alignment process were seeded from models generated from the TIMIT corpus [22]. Because the noise level of the AV-TIMIT corpus was higher than that of TIMIT (which was recorded with a noise-canceling close-talking microphone), the initial time-aligned transcriptions were not as accurate as we had desired (as determined by expert visual inspection against spectrograms). To correct this, the acoustic models were iteratively retrained on the AV-TIMIT corpus from the initial transcriptions. After two re-training iterations, a final set of transcriptions were generated and deemed acceptable based on expert visual inspection. These transcriptions serve as the reference transcriptions used during the phonetic recognition evaluation presented in Section 4.

2.3.2 Video Processing

The video was annotated in two different ways. First, the face region was extracted using a face detector. In order to eliminate any translation of the speaker's head, the face sequence was stabilized using correlation tracking of the nose region. However, since we also needed a reliable way to locate the speaker's mouth, we then used a mouth tracker to extract the mouth region from the video. The mouth tracker is part of the visual front end of the open source AVCSR toolkit available from Intel [9].

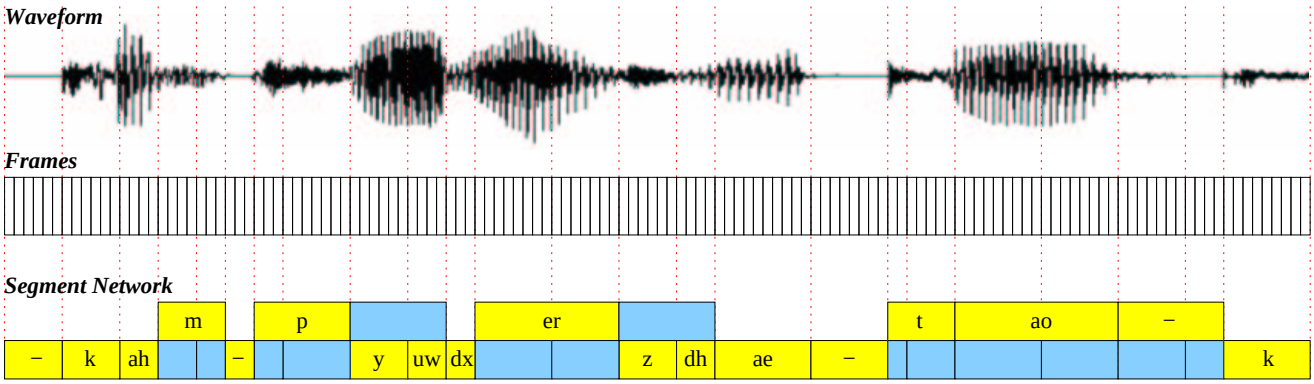

Figure 3: Illustration of a search network created for segment-based recognition. The best segment path is highlighted in the segment network.

Although the front end algorithms were trained on different corpora than our own, they performed relatively well on the AV-TIMIT corpus. The mouth tracker uses two classifiers (one for mouth and the other for mouth-with-beard) to detect the mouth within the lower region of the face. If the mouth was detected successfully in several consecutive frames, the system entered the tracking state, in which the detector was applied to a small region around the previously detected mouth. Finally, the mouth locations over time were smoothed and outliers were rejected using a median filter. For more details about the algorithm, see [11]. The system performed well on most speakers; however, for some it produced unacceptable tracking results (see Figure 2). Two possible reasons for such failures are side lighting and rotation of the speaker's head, both of which the system had difficulty handling. Facial expressions, e.g. smiling, also seemed to have a negative effect on tracking. Another possibility is that the fixed parameters used in the search did not generalize well to some speakers' facial structure. To obtain better tracking in such cases, the search area for the mouth in the lower region of the face was adjusted manually, as were the relative dimensions of the mouth rectangle.

With these measures, most of the remaining tracking failures were in the first few frames of the recording, before the speaker started reading the sentence. The final tracking results, consisting of a 100x70 pixel rectangle centered on the mouth in each frame, were saved to a separate file in raw AVI format.

3. SEGMENT-BASED AVSR

3.1 Segment-Based Recognition

Our audio-visual speech recognition approach builds upon our existing segment-based speech recognition system [5]. 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 milliseconds). In contrast, segment networks are based on the idea that speech waveforms can be broken up into variable length segments that each correspond to an acoustic unit, such as a phone.

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 milliseconds. 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 a 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. Figure 3 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.

3.2 Visual Feature Extraction

There are two main approaches to visual feature extraction for speech recognition. The first is an *appearance-based*, or bottom-up, approach, in which the raw image pixels are compressed, for example, using a linear transform, such as a discrete cosine transform (DCT), principle component analysis (PCA) projection, or a linear discriminant analysis (LDA) projection. The second is a *model-based*, or top-down, approach, in which a pre-determined model, such as the contour of the lips, is fitted to the data. Some approaches combine both appearance and model-based features. It has been found that, in general, bottom-up methods perform better than top-down methods, because the latter tend to be sensitive to model-fitting errors [15].

For this experiment, appearance-based visual features were extracted from the raw images of the mouth region using the visual front end of the AVCSR Toolkit. First, each image was normalized for lighting variation using histogram equalization. Then, a PCA transform was applied, and the top 32 coefficients retained as feature vectors. Figure 4 shows the lip images re-constructed from the means of the PCA coefficient distributions for the middle frame of each phoneme. In order to capture the dynamics of the signal, three consecutive vectors were concatenated together, resulting in one 96-dimensional vector per frame.

3.3 Visual Recognition Units

Typical audio-only speech recognition systems use phones (i.e., the acoustic realizations of phonemes) as the basic units

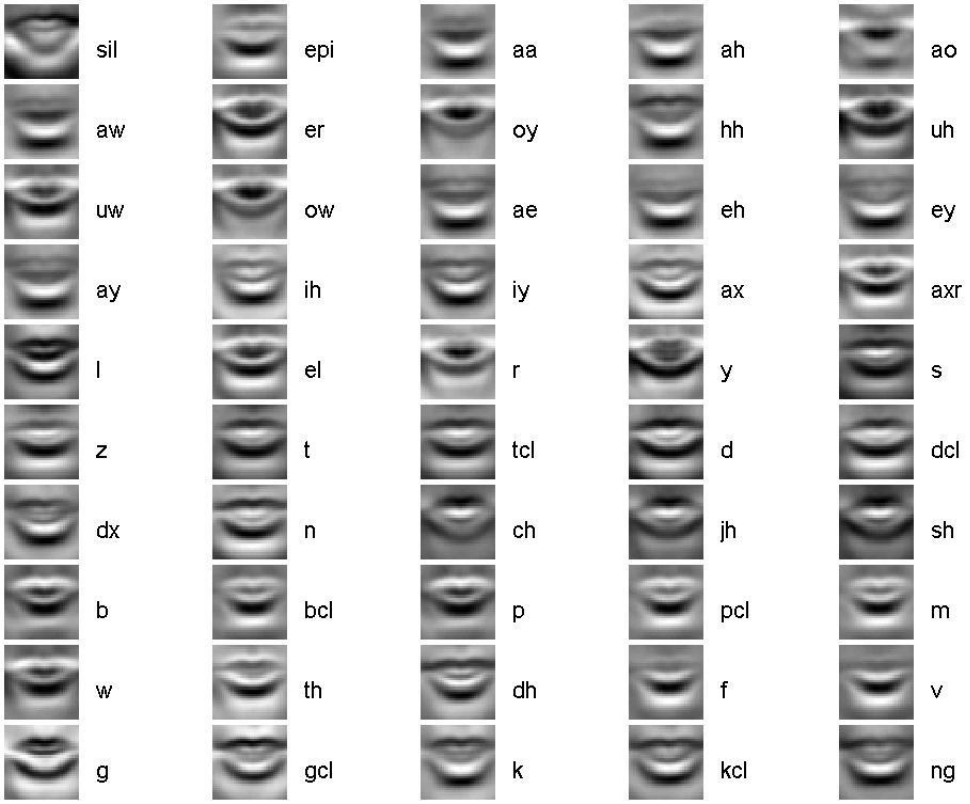

Figure 4: Image representation generated from mean value of the 32-dimension feature vector extracted from video frames for each of 50 different phonetic labels.

for speech recognition. Words are represented as strings of phones, which can then be mapped to acoustic observations using context-dependent acoustic models. When adding visual information, most systems incorporate visual units called *visemes*, which correspond roughly to phones, and describe the visual realization of phonemes.

In general, one can only see a speaker's lips and jaws, while the other articulators (e.g., the tongue and the glottis) are typically hidden from sight. Therefore, some visemes can actually correspond to more than one phone, resulting in a one-to-many mapping. For example, the phones [b] and [p] differ from each other only in that [b] is voiced. Since voicing occurs at the glottis, which is not visible, these two phones are visually indistinguishable, and can thus be mapped to the same viseme unit. From a probabilistic modeling viewpoint, the use of viseme units is essentially a form of model clustering that allows visually similar phonetic events to share a tied model.

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. We use the Bhattacharyya distance as a measure of similarity between two Gaussian distributions, which is defined as follows:

$$\frac{1}{8}(\mu_1-\mu_2)^t\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_1-\mu_2)+\frac{1}{2}\ln\frac{|(\Sigma_1+\Sigma_2)/2|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}}$$

We started from 50 visual feature distributions corresponding to 50 phonetic units (there are 54 total phonetic units used in our recognizer, but during this clustering we pre-clustered the two silence units into one unit, as well as merging [em] with [m], [en] with [n], and [zh] with [sh]). We then used a standard agglomerative hierarchical clustering algorithm to successively merge clusters based on the maximum distance between them. The resulting cluster tree based on the 96-dimensional stacked PCA feature vectors is shown in Figure 5. In almost all cases the clusterings are obvious and expected. For example the cluster of [s], [z], [n], [tcl] and [dcl] represents all of the phones with coronal closures or constrictions.

Table 1 shows the list of viseme units used in our system, along with the set of phonetic units corresponding to each viseme (as roughly based on the clusterings in Figure 5). As mentioned early, the entire phonetic unit set contains 54 different units. This set of units roughly mimics the set of units used in TIMIT.

3.4 Audio-Visual Integration

3.4.1 Early vs. Late Integration

One of the key questions to ask when integrating audio and visual data is "When should the information be combined?" In *early integration*, feature vectors from both modalities are concatenated into one large vector. This resulting vector is then processed by a joint audio-visual classifier, which uses the combined information to assign likelihoods to the recognizer's phonetic hypotheses.

In *late integration*, the audio data and video data are analyzed by separate classifiers. Each classifier processes its

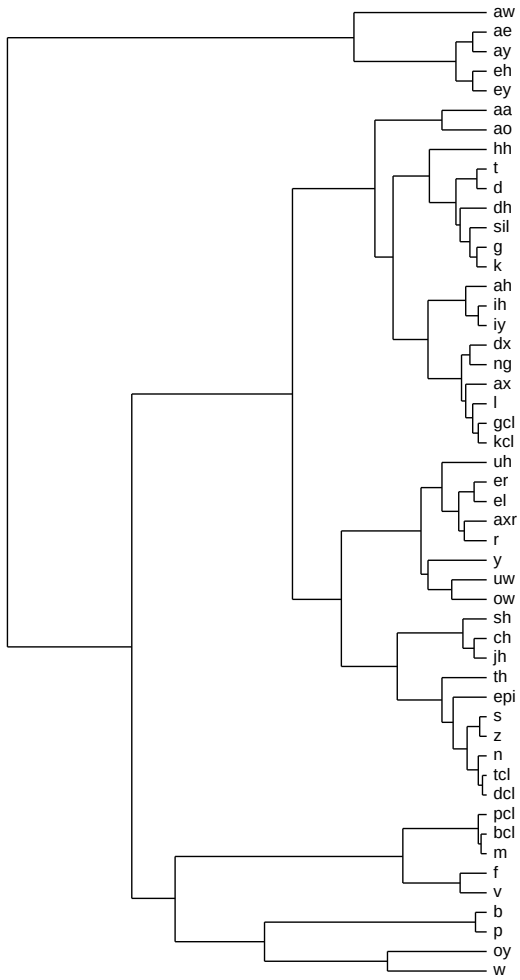

Figure 5: Bottom-up clustering of phonetic events based on similarity of Gaussian models constructed from the central visual frame of phonetic segments using 96-dimension stacked PCA feature vectors.

own data stream, and the two sets of outputs are combined in a later stage to produce the final hypothesis. We have chosen to integrate two independent audio and visual classifiers at the segment-level. In other words, the audio and visual streams are processed independently until they are merged during the segment-based search to produce joint audio-visual scores for each segment.

We chose to perform late integration instead of early integration for two primary reasons. First, the feature concatenation used in early integration would result in a high-dimensional data space, making a large multi-modal database necessary for robust statistical model training.

Second, late integration provides greater flexibility in modeling. With late integration, it is possible to train the audio and visual classifiers on different data sources. Thus, when training the audio classifier, we could use audio-only data sources to supplement the available joint audio-visual data. Also, late integration allows adaptive channel weighting between the audio and visual streams based on environmental conditions, such as the signal-to-noise ratio. Additionally, late integration allows asynchronous processing of the two streams, as discussed in the next subsection.

Viseme Label	Phone Set
Sil	- _
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

Table 1: Mapping of phonetic units to visemes for our experiments.

3.4.2 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 preciously synchronized at an 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.

A variety of methods for modeling audio-visual asynchrony have been proposed in the literature including multi-stream hidden Markov models [4], product hidden Markov models [4], and multi-state time-delay neural networks [13]. In all of the previous work we have examined, frame-based modeling is used on both streams, and asynchrony is controlled via joint constraints placed on the finite-state networks processing each stream.

3.4.3 Segment-Constrained HMMs

In our AVSR recognizer, we have implemented asynchronous modeling of the audio and visual streams using an approach we call segment-constrained HMMs. This modeling is implemented with three primary steps. First, fixed-length video frames are mapped to variable-length audio segments defined by the audio recognition process. 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 is mapped to a

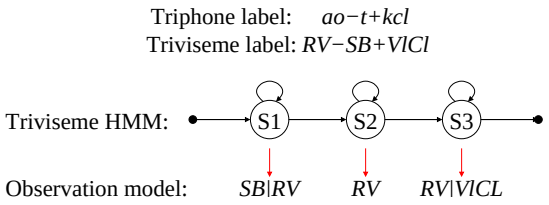

Figure 6: Example segment-constrained triviseme HMM from our system.

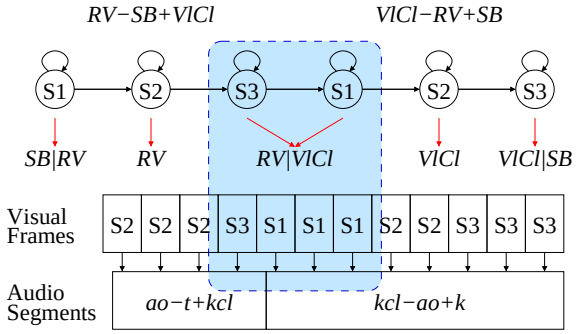

Figure 7: Using segment-constrained HMMs to represent audio-visual asynchrony for a given audio segment sequence.

context-dependent segment-constrained viseme HMM. Finally, the segment-constrained viseme HMM uses a frame-based Viterbi search over visual frames in the segment to generates a segment-based score for the visemic model. These steps are described in the following paragraphs.

The first step of our audio-visual integration is identifying the visual frames corresponding to each segment. Since the audio sampling rate is often not an integer multiple of the video sampling rate, a convention must be defined to systematically map frames to segments. In our approach, the beginning and ending times of any visual frame are averaged to obtain the frame's midpoint. Then, the frame is assigned to any segment whose start time lies before the frame's midpoint, and whose end time lies after it. This scheme has the desirable property that, for any given path of non-overlapping segments covering all times in the utterance, each video frame is mapped to exactly one segment.

The audio recognizer uses triphone-based context-dependent modeling for each segment. The mapping of segment-based triphone acoustic models to frame-based viseme models is accomplished through our segment-constrained HMM approach. Each triphone acoustic model is mapped to a corresponding triviseme visual model. Each visual model is represented by a three state, left-to-right HMM model which allows any of the three states to be skipped. Figure 6 shows an example triviseme HMM model used for the phonetic triphone *ao-t+kcl* (where *ao* is the current phone, *t* is the left context, and *kcl* is the right context). This triphone is mapped to the triviseme *RV-SB+VlCl* (where *RV* is a rounded vowel, *SB* is a stop burst, and *VlCl* is a velar closure.)

Figure 6 also shows the mapping from each triviseme state to the label of the observation density function it uses. In our approach, the left state of every triviseme HMM is mapped to a diviseme model (e.g., *SB|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|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.

At this point, it is important to note that the third state (*S*3) of a triviseme model will use the same diviseme observation density function as the first state (*S*1) of the triviseme model for the following segment. This is illustrated in Figure 7, 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|VlCl*, for their output probability function. This figure demonstrates how a constrained form of asynchronous modeling is introduced by our implementation of the viseme HMMs. 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 of the preceding or following audio segments. Also note that the segment constrained HMM is allowed to skip any of the three states in a triviseme model. An example of this is shown in Figure 7 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 Figure 7, provides an example of 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 occurs during a period of time when no acoustic change is evident. This asynchrony is handled in our model by allowing the visual frames assigned to *RV|VlCl* viseme transition to straddle the acoustic segment boundary separating the [ao] acoustic segment from the [kcl] acoustic segment.

4. EXPERIMENTAL RESULTS

For our initial experiments with our AVSR design, we have implemented a phonetic AVSR recognizer that we have trained and evaluated on the AV-TIMIT corpus. The following sections provide the details and results of our experiments.

4.1 Experimental Data Sets

For our experiments we sub-divided 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 evaluate using frontal lighting conditions only (and ignore the side-lighting condition). Under this constraint, the training set is reduced to 2751 utterances. The full 3608 utterances are still used for training the acoustic models, while the visual classification models are trained from only the reduced set of 2751 frontal lighting utterances.

For evaluation, our development test set contains 284 ut-

terances from 19 speakers. The final test set contains 285 utterances from another 19 speakers. There is no overlap in speakers between any of the three data sets.

4.2 Modeling Details

For acoustic modeling our system utilizes two types of acoustic measurements, segment features and landmark features. Details about these measurement types can be found in [5] and [21]. These models were trained from time-aligned phonetic transcriptions described in Section 2.3.1. In all, the segment model contains 834 context-dependent triphone-based segment models, and 629 diphone- and triphone-based landmark models.

For visual modeling, the time-aligned phonetic transcriptions were mapped to viseme labels using the mapping contained in Table 1. Visual frames were mapped to the time-aligned segments using each frame's mid-point. To initialize the triviseme observation models, each frame was mapped to a specific triviseme state (either the first, second or third state) based on its relative position in the segment (i.e., in the first, second, or final third). Each triviseme state was then mapped to its appropriate observation model (i.e., a diviseme transition model for the first and third states, or a context-independent viseme model for the middle state). The trained models obtained from this initial mapping were then subjected to two rounds of iterative Viterbi retraining. In total the viseme observation models contained 15 context-independent viseme models and 203 diviseme transition models. In addition to the viseme observation models, the transition probabilities in the triviseme segment-constrained HMMs were also estimated during the Viterbi training process. In total, there are 2690 unique trivisemes resulting in 8070 states in the segment-constrained HMMs.

In addition to the audio and visual classifiers, a phonetic bigram was estimated from the training utterances to provide the language model constraint. The recognizer also controls the trade-off between phonetic insertions and phonetic deletions using a single segment transition weight. The scores from the visual classifier, audio classifiers, and phonetic bigram are integrated using linear weighted combination. The weighting factors for the linear combination were tuned to maximize phonetic recognition performance on the development test.

4.3 Results

Our initial AVSR results are reported in Table 2. Results are reported in phonetic recognition error rates over the standard 39 phonetic classes typically used in TIMIT experiments [10]. As can be seen in the table, when incorporating the visual information into the recognizer, a relative reduction of 5% in phonetic error rate was obtained on the development set, when the system's scaling weights were tuned appropriately. On the final test set, our AVSR system produced a relative 2.5% reduction in errors from our ASR system. Although these gains appear small, previous studies in quiet environments have also found relatively small improvements in performance when adding visual information [17]. With an average signal-to-noise ratio of 25 dB, the AV-TIMIT corpus is relatively noise-free. We expect a bigger benefit from the visual channel when we move to noisier environments. We should also note that previous "small" improvements in phonetic recognition have translated into larger improvements in word recognition when our modeling techniques are transitioned from phonetic recognition to word recognition tasks [6, 7].

Test Set	Type of Recognizer	Phonetic Error Rates (%)			
		Subs.	Ins.	Del.	Total
Dev	Audio Only	23.6	6.8	9.9	39.8
Dev	Audio-Visual	21.9	6.4	9.4	37.8
Test	Audio Only	20.9	4.9	10.4	36.1
Test	Audio-Visual	20.2	4.6	10.4	35.2

Table 2: Phonetic recognition error rates using audio-only and audio-visual recognition.

Phone Pair	Errors Corrected
m n	16
k p	16
l r	10
ae ax	10
ae ao	7

Table 3: Phone pairs which accounted for the largest number of ASR substitution errors corrected by the AVSR system.

To examine what types of errors are corrected when the visual information is added, we extracted the list of phonetic confusions that were most often corrected when moving from our audio-only recognizer to our AVSR system. Table 3 shows the five phonetic confusions must commonly corrected by the AVSR system. In each of these five phone pairs, it is clear that visual information can be exploited to improve recognition between the pairs. For example, the [m]/[n] and [k]/[p] pairs can be distinguished by the presence or absence of a labial closure.

5. SUMMARY

In this paper we have presented details about our initial efforts in developing an audio-visual speech recognition system. This research contains two primary contributions. First, we have collected a new audio-visual speech corpus based on TIMIT utterances. This corpus contains 4 hours of video from 223 different speakers. It is our plan to make all (or at least portions) of this corpus publicly available to the research community once its annotation has been fully completed and verified. As of the publication time of this paper, we are in the final stages of verifying the video annotations (i.e., the face and lip tracking results), and are beginning to investigate the potential avenues for distributing this data. The most obvious mechanism is to distribute the data via the Linguistic Data Consortium, but no final decision has yet been determined in this regard.

The second major component of this work is the development of an audio-visual speech recognition system which utilizes a new approach to audio-visual integration, which we call segment-constrained hidden Markov modeling. A preliminary implementation of this approach using standard acoustic and visual processing techniques yielded a 2.5% reduction in phonetic error rate during our experiments with the AV-TIMIT corpus.

6. FUTURE WORK

As with all development work in its initial phases, there is still much to be done on our project. From the scientific point of view, we are currently exploring new visual features for efficiently and robustly representing relevant lip information [18]. For real-world applications, such as kiosks, visual features need to be robust to additional difficulties such as visual occlusions, rotated or tilted head positions, and general visual noise (e.g., variable lighting conditions, video compression artifacts, etc.).

From an engineering point of view, we plan to explore a variety of techniques for improving performance, including replacing our current manually determined set of viseme clusters with a clustering determined using a top-down decision tree approach. We also plan to investigate the effect of different noise conditions on our approach, and specifically how it affects the weighting of the audio and visual scores.

From a development point of view, we plan to migrate this system to word recognition tasks, and eventually to deploy this system within real applications in difficult environments such as automobiles, public kiosks, and noisy offices. To this end, we have already undertaken a data collection effort within moving automobiles, and future data collections in other challenging environments are anticipated.

7. ACKNOWLEDGMENTS

This work was sponsored in part by the Ford-MIT Alliance. The authors also wish to thank Marcia Davidson for her efforts during the data collection process, and all of the subjects who contributed data for the AV-TIMIT corpus.

8. REFERENCES

[1] C. Benoit. The intrinsic bimodality of speech communication and the synthesis of talking faces. In *Journal on Communications of the Scientific Society for Telecommunications*, Hungary, number 43, pages 32–40, September 1992.

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

[3] S. Chu and T. Huang. Bimodal speech recognition using coupled hidden Markov models. In *Proc. of the International Conference on Spoken Language Processing*, vol. II, Beijing, October 2000.

[4] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. In *IEEE Transactions on Multimedia*, number 2, pages 141–151, September 2000.

[5] J. Glass. A probabilistic framework for segment-based speech recognition. To appear in *Computer Speech and Language*, 2003.

[6] A. Halberstadt and J. Glass. Heterogeneous measurements and multiple classifiers for speech recognition. In *Proceedings of ICSLP 98*, Sydney, Australia, November 1998.

[7] T. J. Hazen and A. Halberstadt, "Using aggregation to improve the performance of mixture Gaussian acoustic models," In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Seattle, May, 1998.

[8] IBM Research - Audio Visual Speech Technologies: Data Collection. Accessed online at http://www.research.ibm.com/AVSTG/data.html, May 2003.

[9] Intel's AVCSR Toolkit source code can be downloaded from http://sourceforge.net/projects/opencvlibrary/.

[10] K. F. Lee and H. W. Hon. Speaker-independent phone recognition using hidden Markov models. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641-1648, November 1989.

[11] L. H. Liang, X. X. Liu, Y. Zhao, X. Pi and A.V. Nefian. Speaker independent audio-visual continuous speech recognition. In *Proc. of the IEEE International Conference on Multimedia and Expo*, vol.2, pp. 25–28, 2002.

[12] I. Matthews, J. A. Bangham, and S. Cox. Audio-visual speech recognition using multiscale nonlinear image decomposition. In *Proc. of the International Conference on Spoken Language Processing*, pp. 38–41, Philadelphia, PA, 1996.

[13] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel. Towards unrestricted lip reading. In *International Journal of Pattern Recognition and Artificial Intelligence*, number 14, pages 571–585, August 2000.

[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*, pages 72–77, Washington, D.C., March 1999. 16 IDIAP–RR 99-02.

[15] C. Neti, *et al.* Audio-visual speech recognition. In *Technical Report, Center for Language and Speech Processing*, Baltimore, Maryland, 2000. The Johns Hopkins University.

[16] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database. In *Proc. of the Audio- and Video-based Biometric Person Authentication Workshop*, Germany, 1997.

[17] G. Potamianos and C. Neti. Audio-visual speech recognition in challenging environments. In *Proc. Of EUROSPEECH*, pp. 1293–1296, Geneva, Switzerland, September 2003.

[18] K. Saenko, T. Darrel, and J. Glass. Articulatory features for robust visual speech recognition In these proceedings, *ICMI'04*, State College, Pennsylvania, 2004.

[19] C. Sanderson. The VidTIMIT Database. *IDIAP Communication 02-06*, Martigny, Switzerland, 2002.

[20] C. Sanderson. Automatic Person Verification Using Speech and Face Information. PhD Thesis, Griffith University, Brisbane, Australia, 2002.

[21] N. Ström, L. Hetherington, T.J. Hazen, E. Sandness, and J. Glass. Acoustic modeling improvements in a segment-based speech recognizer. In *Proc. 1999 IEEE ASRU Workshop*, Keystone, CO, December 1999.

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