

# AUDIO-VISUAL LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION IN THE BROADCAST DOMAIN

S. Basu, C. Neti, N. Rajput\*, A. Senior, L. Subramaniam\*, A. Verma\*  
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598  
\* IBM Solutions Research Center, New Delhi, India.

**Abstract** - We consider the problem of combining visual cues with audio signals for the purpose of improved automatic machine recognition of speech. Although significant progress has been made in machine transcription of large vocabulary continuous speech (LVCSR) over the last few years, the technology to date is most effective only under controlled conditions such as low noise, speaker dependent recognition and read speech (as opposed to conversational speech) etc. On the other hand, while augmenting the recognition of speech utterances with visual cues has attracted the attention of researchers over the last couple of years, most efforts in this domain can be considered to be only preliminary in the sense that unlike LVCSR efforts, tasks have been limited to small vocabulary (e.g., command, digits) and often to speaker dependent training or isolated word speech where word boundaries are artificially well defined.

## INTRODUCTION

The potential for joint audio-visual-based speech recognition is well established on the basis of psychophysical experiments [3, 2]. Efforts have begun recently on experiments with small vocabulary letter or digit recognition tasks (see e.g., [7, 6, 9]). Canonical mouth shapes that accompany speech utterances have been categorized, and are known as visual phonemes or "visemes" [2]. Visemes provide information that complements the phonetic stream from the point of view of confusability. For example, "mi" and "ni" which are confusable acoustically, especially in noise situations, are easy to distinguish visually: in "mi" lips close at onset, whereas in "ni" they do not. The unvoiced fricatives "f" and "s" which are difficult to recognize acoustically belong to two different viseme groups [2].

## METHOD

First a face is located in the first frame of a video sequence and the location is tracked across frames in the video clip. For locating a face, an image pyramid over the permissible scales is generated and for every location in

the pyramid, we score the surrounding area as a face location. Rectangles of pixels from the image pyramid are scored based on a number of criteria, including similarity to skin-tone, and similarity to a diverse training set of face images using the Fisher discriminant analysis and 'distance from face space' [15]. The locations scoring highly on all criteria are determined to be faces. For each high-scoring face locations we consider small translation, scale and rotation changes, rescoreing the face region under each of these changes to optimize the estimates of these parameters. Having found the face, an ensemble of facial feature detectors can be used to determine and verify the locations of the important facial features, including the lip corners and centers. Subsequently, lip parameters are extracted. The technique is an adaptation of a computer vision based face identification method [12] to the present context.

**Visual feature extraction:**

Although previous work has been conducted to define the viseme units derived from human lip-reading experiments [2] and other psychophysical data, more research is necessary to identify the mouth features that are relevant for large vocabulary, speaker independent visual speech recognition.

Candidate features are gray-scale parameters of the mouth region; geometric/model based parameters such as area, height, width of mouth region; lip contours arrived at by curve fitting, spline parameters of inner/outer contour; and motion parameters obtained by 3-D tracking. Gray scale parameters suffer from being sensitive to lighting conditions. Lip contour information, although invariant to lighting conditions, may not provide enough information of the inner articulators such as teeth and tongue. Still another feature set suggested recently to take into account the above factors is the Active Shape model [8].

In this report, we consider grey scale parameters associated with the mouth region of the image. Given the location of the lip corners a rectangular region of normalized scale and rotation, centred on the mouth center is subsampled from the original video frame. Principal Component Analysis (PCA) was used to extract a vector of smaller dimension from this vector of grey-scale values.

**Audio feature extraction:**

Digitized speech sampled at a rate of 16 kHz was considered. A *frame* consists of a segment of speech of duration 25 ms, and produces an 24 dimensional acoustic cepstral vector via the following process, which is standard in speech recognition literature. Frames are advanced every 10 ms to obtain succeeding acoustic vectors.

First, magnitudes of discrete Fourier transform of samples of speech data in a frame are considered in a logarithmically warped frequency scale. Next, these amplitude values themselves are transformed to a logarithmic scale<sup>1</sup>,

---

<sup>1</sup>The later two steps are motivated by logarithmic sensitivity of human hearing to frequency and amplitude.

and subsequently, a rotation in the form of discrete cosine transform is applied. One way to capture the dynamics is to use the delta (first-difference) and the delta-delta (second-order differences) information. An alternative way to capture dynamic information is to append a set of (say *four*) preceding and succeeding vectors to the vector under consideration and then project the vector to a lower dimensional space, which is chosen to have the most discrimination. The latter procedure being known as the Linear Discriminant Analysis (LDA) in standard literature.

## Mode Fusion

Once the feature vectors for the lip movements are available from the video data, in principle techniques similar to those used in standard speech recognition can be applied to the classification and subsequent recognition of video features. The more important issue is how to integrate the audio and visual data for the purpose of recognition. This is a form of data fusion problem, which we address via the following considerations.

Using visual information to augment the audio signal for speech recognition involves the ability to fuse different representations of the same underlying production process. Such a mode-fusion or multi-modal integration involves the following categories of sensory data fusion [10]. These constitute: (1) feature fusion — features are extracted from the raw data and subsequently combined. This involves, for example, fusing speech features with lip and facial features; (2) decision fusion — this is the fusion at the most advanced stage of processing, after independent classification of each modality and can happen at the sub-word level, word-level, utterance level or at the action level. In the following section, we describe some preliminary experiments using feature fusion. In feature fusion, we first extract audio features and video features and then concatenate the two to generate a single audio-video feature vector. We use LDA (as described earlier) on the combined vector to generate a lower dimensional discriminant feature representation.

## EXPERIMENTS

The audio-visual data we experiment with is provided by the LDC [11]. The audio part of the data is a subset of the standard speech recognition evaluation conducted by the DARPA community (known as the HUB4 effort). The speech database consists of large vocabulary (approximately 60,000 words) continuous speech drawn from a variety of news broadcasts. The entire database includes television (e.g. CNN, CSPAN) as well as radio shows. We focus on segments of this data, the video part of which primarily consists of the “talking head” type images (e.g. an anchor-person in a CNN newscast). The audio-video data available from LDC [11] in the analog SVHS format is digitized in MPEG2 format (at a rate of 5Mb/sec). The audio and

video streams are then de-multiplexed and decompressed. The resulting decompressed audio is sampled at a rate of 16 kHz and the video at a standard rate of 30 frames/second.

The audio sampling rate of 16 kHz is chosen so as to be able to compare the joint audio-visual recognition results with the audio-only HUB4 evaluation experiments. While this is an ongoing data collection effort, at the present time we have about 700 video clips of approximately 10-15 seconds duration each (the entire HUB4 database is approximately 200 hrs. of speech data, not all of which is usable for our purpose).

In summary, we use a database of large vocabulary continuous visual speech transmitted over a broadcast channel. The fact that it is real-life data (as opposed to data collected in controlled environments) distinguishes this from existing databases. While making the system applicable to real problem domains, this does make visual feature extraction and subsequent processing a more challenging task.

The need for controlled data is not to be underplayed and, in our view, may indeed have an important role to play in this general area of research. For purposes of validation of results we also collected "read" large vocabulary continuous visual speech. This data was collected in acoustically quiet, controlled conditions and the resolution of the lip region in the video image was much larger than in the LDC data mentioned above — thus making video based recognition a more tractable task. For the purpose of fair comparison with the LDC data, the video digitization parameters and audio sampling frequency were kept the same. We label this data the 'ViaVoice Audio-Visual' (VVAV) data.

We report specific results on the joint audio-video phonetic classification and its comparison with audio-only and video-only classification. For video we experiment with both the phonetic classification and a 'viseme' based approach as described above. One approach to labeling the video feature vectors is to label the speech data from a Viterbi alignment and to subsequently use a phoneme to viseme mapping. To produce phonetic alignments of the audio data we use the acoustic models trained using the DARPA HUB4 speech recognition data. The video frame rates are typically lower than the audio frame rate. This is circumvented by inter-frame interpolation. In all experiments the HUB4-video database of continuous large vocabulary speech mentioned in Section [11] is used.

In the following experiments 672 audio-video clips of VVAV data was used as the training set. The test set consisted of 36 different clips taken from the same database. All the experiments use LDA features. In the phonetic/visemic classification each phone/viseme is modeled as a mixture of 5 gaussians.

A comparison of Tables 1 and 2 shows that audio-visual recognition in acoustically degraded conditions is better than either of the two streams processed independently. An approximate improvement of 14% is obtained

compared to audio-only classification scheme.

We used the following grouping of phonemes into viseme classes. For a detailed explanation of the symbols used for phoneme classes we refer to [13].

(AA, AH, AX), (AE), (AO), (AW), (AXR, ER), (AY), (CH), (EH), (EY), (HH), (IH, IX), (IY), (JH), (L), (OW), (OY), (R), (UH, UW), (W), (X, D\$), (B, BD, M, P, PD), (D, DD, DX, G, GD, K, KD, N, NG, T, TD, Y), (TS), (F, V), (S, Z), (SH, ZH), (TH, DH).

When visemes are used as classes, the video classification improves by about 37.5%, relative. However, improvement in noisy conditions is about the same for visemic classes.

In our very preliminary experiments with HUB4 broadcast news data, we get the following results. Audio-only phonetic classification accuracy is 33.98%. Video-only phonetic classification accuracy using 35 dimensional LDA features is 9.48%.

These results are relatively poor compared to VVAV data. First, the resolution of the mouth region for the HUB4 data is much less compared to VVAV data, with the possibility of providing very little discriminative information between phones. Secondly, the tracking of the lip region is a harder problem and hence may result in loss of crucial information for discrimination. We are investigating techniques to better track and represent the lower resolution images in the HUB4 data.

## DISCUSSION AND CONCLUSIONS

In this preliminary report we have undertaken phonetic/visemic classification experiments for large vocabulary continuous speech. Our goal is to proceed to HMM-based recognition techniques and compare the results of joint audio-video recognition with audio only recognition in the context of this specific environment.

In addition to speech recognition, the same problems of channel and environment dependence arise in speaker identification. Again, the problem can be alleviated by combining visual signatures of the speaker both in terms of characteristics of visual speech and other facial features to perform speaker identification. Combined use of audio and visual information is beginning to show improvements in such problems as well. One such example is [13] in which computer vision-based face recognition techniques have been shown to benefit significantly when augmented with speech-based authentication methods. See [14] for some application contexts for combined use of speech and vision.

## References

- [1] P. Cohen, S. Dharanipragada, J Gross, M. Monkowski, C. Neti, S.

- Roukos, T. Ward, Towards a universal speech recognizer for multiple languages, Proc. ICSLP, 1998.
- [2] D. Stork and M. Henecke, Speechreading by humans and machines, NATO ASI Series, Series F, Computer and System Sciences, vol.150, Springer Verlag, 1996.
  - [3] Q. Summerfeld, Use of visual information for phonetic perception, *Phonetica*(36), pp. 314-331 1979.
  - [4] A. J. Goldschien, O. N. Garcia and E. D. Petajan, Rationale for phoneme-viseme mapping and feature selection in visual speech recognition, In [2], pp. 505-515.
  - [5] Tsuhan Chen, Ram R. Rao, Audio visual integration in multimodal communication, *Proceedings of IEEE*, vol.86, no. 5, 837-852, May 1998.
  - [6] G. Potamianos and H. P. Graf, Discriminative training of HMM stream exponents for audio-visual speech recognition, *ICASSP*, 1998.
  - [7] C. Bregler and Y. Konig, Eigenlips for robust speech recognition, *ICSLP*, vol II, pp. 669-672, 1994.
  - [8] Iain Matthews, Features for audio visual speech recognition, Ph.D dissertation, School of Information systems University of East Angalia, January 1998.
  - [9] R. Stiefelhagen, U. Meier and J. Yang, Real-time lip-tracking for lipreading, preprint.
  - [10] David L. Hall, *Mathematical Techniques in multisensor data fusion*, Artech House, 1992.
  - [11] Linguistic Data Consortium, University of Pennsylvania.
  - [12] Andrew Senior, Face and feature finding for face recognition system, 2nd Int. Conf. on Audio-Video based Biometric Person Authentication, Washington DC, March 1999.
  - [13] Benoit Maison, Chalapathy Neti and Andrew Senior Audio Visual Speaker Recognition for Video Broadcast News: some fusion techniques *Proceedings of MMSP'99*, Denmark, September, 1999.
  - [14] Sankar Basu, E E Jan, Chalapathy V Neti and Mark Lucente, Beyond audio-based speech recognition for natural human computer interaction, NIST DARPA Invitational Workshop on Smart Spaces, August 1998.
  - [15] M. Turk and A. Pentland, Eigenfaces for Recognition. *Journal of Cognitive Neuro Science* Vol. 3 No. 1 pp. 71-86 1991.

Data Type	Dimension	Splice Param.	Reco. Rate
Audio Only (Training Data)	24	60 dim	53.66%
Video Only (Training Data)	100	35 dim	22.21%
Audio Only (Test Data)	24	60 dim	48.08%
Video Only (Test Data)	100	35 dim	20.15%
Audio-Video (Training Data)	24+50	35 dim	53.58%
Audio-Video (Test Data)	24+50	35 dim	48.71%

Table 1: Clean Data Experiments on VVAV data

Data Type	Dim	Splice dim	phonetic	visemic
Audio Only (Test)	24	60 dim	28.05%	40.40%
Video Only (Test)	100	35 dim	20.15%	27.76%
Audio-Video (Test)	24+50	35 dim	32.02%	44.81%

Table 2: Noise Data Experiments on VVAV data: Speech Noise at an average of 15 dB, ranging from 10-20 dB SNR. Phonetic accuracy is shown in the fourth column and visemic accuracy in the fifth column

Data Type	Dim	Splice Param.	Reco. Rate
Audio Only (Training Data)	24	(60 dim)	62.17%
Audio Only (Test Data)	24	(60 dim)	60.52%
Video Only (Training Data)	100	(35 dim)	28.14%
Video Only (Test Data)	100	(35 dim)	27.76%

Table 3: Viseme based video classification results