LATE INTEGRATION IN AUDIO-VISUAL CONTINUOUS SPEECH RECOGNITION

Ashish Verma, Tanveer Faruquie

IBM Solutions Research Center
New Delhi, India 110016

Chalapathy Neti, Sankar Basu

IBM T. J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

Using visual information in speech recognition has been an area of interest because it can significantly improve the speech recognition efficiency in the conditions where audio only recognition suffers due to noisy environment. In this paper, we present a new approach to combine audio and video to improve the robustness of the speech recognition system in the noisy environments. We also compare the results of the new approach with the corresponding results of the approaches proposed earlier in the literature. Keywords: Late Integration, Viseme, PCA, LDA, etc.

1. INTRODUCTION

Increasing robustness of the speech recognition system against different kind of noises in the audio channel has become a focus area in the recent past. This is due to the fact that the performance of all the speech recognition systems suffers to a great extent in non-controlled environment like background noise, bad acoustic channel characteristic, crosstalk, etc. Video plays an important role in these conditions as it provides significant information about the speech which can compensate for the noise in the audio channel. Furthermore, it has been observed that some amount of orthogonality is present between the audio and the video channel which can be used to improve the recognition efficiency by combining the two channels [2, 3, 4].

Researchers have experimented with various features of audio and visual speech and different methods of combining the two information channels. One of the earliest audio-visual speech recognition system was implemented by Petajan [5, 6]. In his experiment, binary images were used to extract mouth parameters like height, width and area of the mouth of the speaker which were later used in the recognition system. The recognition system was an audio speech recognizer followed by a visual speech recognizer. Therefore, visual speech recognizer used to work only on a subset of all the possible candidates which were supplied to it by the audio speech recognizer. Later the system was modified to use the images themselves instead of the mouth parameters and the audio-visual integration strategy was changed to a rule based approach form the sequential integration.

Goldschen used a more elaborate scheme for the visual speech recognizer [7]. Goldschen analyzed a number of features of the binary images such as height, width, and perimeter, along with derivatives of these quantities, and used these features as the input to an HMM-based visual speech recognition system. Since then, several experiments have been performed by various researchers to improve upon these basic blocks of audio-visual speech recognition [2, 8, 9, 10, 11, 12, 15].

A central problem in audio-visual speech recognition is to combine the audio and visual streams in an intelligent way. While there is an enormous literature [1] on data fusion in general, in [14] we have described our early work on audio-visual recognition using one of the many possible techniques. In this paper, we investigate a new hierarchical strategy for fusion. We show that the new technique produces improved results.

In section 2 we describe some of the approaches used for the audio-visual speech recognition. Section 3 describes the system setup and the methods used in the experiments. In section 4, we discuss the results and their implications.

2. INTEGRATION APPROACHES

Generally speaking, there are following problems in combining audio with video for speech recognition

- Audio and Video features have different dynamic ranges.

- Audio and Video features have different number of distinguishable classes. In other words, there are different number of phonemes than the number of visemes.

- Due to complexities involved in articulatory phenomena there is a time offset between audio and video signals [9].

- Video signal is usually sampled at a slower rate than the audio, and therefore, needs to be interpolated.

2.1. Early Integration

In general, two different approaches to combine audio and visual information have been tried. In the first approach, called Early Integration or Feature Fusion, audio and visual features are computed from the acoustic and visual speech respectively and they are combined before the recognition experiment. Since the two set of features correspond to different feature spaces, they may differ in their characteristics as described above. Therefore, this approach requires an intelligent way to combine the audio-visual features. The recognition is performed with the combined features and the output of the recognizer is the final result. This approach has been described in [2, 8, 9, 14]. This approach can not handle different classifications in audio and video as it uses a common recognizer for both of them.

2.2. Late Integration

The other approach, called Late Integration or Decision Fusion, incorporates separate recognizers for audio and video channels and then combines the outputs of the two recognizers to get the final result. The final step of combining the two outputs is the most important step in this approach as it has to deal with the issues of orthogonality between the two channels and the reliability of the channels. This approach can easily handle the different classifications in audio and video channels as the recognizers for them are separate and the combination is at the output level. This approach has been described in [10, 11, 15, 12].

However, all of the previous approaches use a single phase experiment with fixed set of phonetic or visemic classes. In this paper, we investigate a two phase (hierarchical) combination strategy to combine audio and video. This approach is further elaborated in section 3.4.

3. SYSTEM DESCRIPTION

3.1. Audio Processing

We extract 24-dimensional mel-cepstral coefficient feature vectors from the audio signal using the standard techniques in speech recognition field. LDA (Linear Discriminant Analysis) has been used to capture the dynamics of the acoustic signal. A more elaborate description of audio processing is provided in [14].

AA,AH,AX	AE	A0
AW	AXR,ER	AY
CH	EH	EY
HH	IH,IX	IY
JH	L	OW
R	UH,UW	W
X,D$	B,BD,M,P,PD	S,Z
F,V	OY	D,DD,DX,G,GD,K, KD,N,NG,T,TD,Y
TS	TH,DH	SH,ZH

Table 1: Set of visemic units used

3.2. Video Processing

A pyramid based face detection approach has been used to extract the face from the video [13]. In this approach, an image pyramid over the permissible scales is used to search the image space for the possible face candidates. Every face candidate is given a score based on several features like skin tone and similarity to a training set of face images using Fisher Discriminant Analysis. Once the face has been found, 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, a mouth image of size 45x30 is extracted from the face image centered around the lips. Principal Component Analysis (PCA) has been used to get the first 100 modes of variations of the lip image. Furthermore, Linear Discriminant Analysis (LDA) has been used to obtain a 35 dimensional visual feature vector from the PCA modes which is used in the experiments.

3.3. Experiments

We have performed phonetic classification experiments over VVAV (Via Voice Audio-Visual) database which consists of about 700 sentences spoken by 6 speakers. To obtain the ground truth for the audio and video vectors, the input speech is aligned to the input text using the Viterbi Alignment by the standard Via-Voice speech engine. For the classification experiments, each phonetic (visemeic) class is represented by a mixture of 5 (respectively 10) Gaussian components. Since we are using Late Integration, there are separate Gaussian mixtures for audio and video. During the experiment, audio and video likelihoods are computed from the audio and video Gaussian mixtures respectively and then they are combined in a way particular to the methods described in the following section.

3.4. Integration Approaches

In the following, we describe the experiments with the hierarchical approach proposed in the present paper and compare its performance with that of the previously proposed approaches.

- **Method 1**

  This method uses phone based classification for both audio and video. We have used 52 phones for the English language. The combined likelihood for a given phone hypothesis is computed in the following way

  $$P_i = w_a * P_{a_i} + w_v * P_{v_i} \quad i = 1, 2, .., 52 \quad (1)$$

  where P_{a_i}, P_{v_i} and P_i are the audio, video and combined likelihoods for phone i respectively. w_a and w_v are weights given to audio and video hypothesis with $w_a + w_v = 1$.

- **Method 2**

  This method uses phone based classes for audio data and viseme based classes for the video data. We have come up with 27 visemes by merging those phones which look alike visually. Table 1 shows the list of visemic units used in the experiments. The equation for likelihood computation in this method is given as follows

  $$P_i = w_a * P_{a_i} + w_v * M_{ij} * P_{v_j} \quad (2)$$

  where P_{v_j} is the likelihood for viseme j given by video vector and M_{ij} is the conditional probability of phone i given viseme j. $M_{ij}s$ have been computed over 300 sentences from the same VVAV database. Here again $w_a + w_v = 1$.

- **Method 3**

  This method computes the combined likelihood for a phone in two phases. In the first phase, only 27 viseme based classes are used for both audio and video. At the end of phase one, we get the most likely viseme based class. Now, in the second phase, phone based models are used for both audio and video to get the most likely phone inside the viseme given by the first phase. In other words, in the second phase, we consider only those phones, which are embedded in the viseme given by the first phase. In cases, where the most likely viseme corresponds to only one phone, second phase is skipped as it is not required. Note that in the second phase, since only the most likely viseme is explored, the computational overhead involved in this phase is very

small. The corresponding equations for the likelihood computations are as follows.

1. Phase 1:

   $$P_i = w_a^1 * P_{a_i} + w_v^1 * P_{v_i}, \quad i = 1, 2.., 27 \quad (3)$$

2. Phase 2:

   $$P_j = w_a^2 * P_{a_j} + w_v^2 * P_{v_j}, j \in \{\text{viseme } k\} \quad (4)$$

   where viseme k is determined as the most likely viseme in the first phase.

Here, as before $w_a^\ell + w_v^\ell = 1$; $\ell = 1, 2$. Note that different weights w_a^1 and w_a^2 are used in the first and second phase.

In all the above experiments the relative weights for audio and video were adjusted manually to obtain the best classification accuracy.

4. RESULTS AND DISCUSSION

Results corresponding to all the methods are given in Tables 2, 3 and 4. Note that the results presented in this paper are only phonetic classification results. Usually, they correspond to much higher <u>recognition rates</u> for the speech recognition system. We are in the process of implementing the audio-visual speech recognition system based on this approach. For noisy audio signal, we have collected crosstalk (or cocktail noise) and used it at various SNR levels.

	Audio (phonetic)	Visual (phonetic)	Combined (Phonetic)
clean	47.67%	21.12%	51.15%
20db	36.26%	21.12%	43.40%
15db	27.90%	21.12%	32.23%
10db	21.77%	21.12%	26.33%

Table 2: Results for Method 1

As shown in Table 2, the simplest form of integration gives about 20.94% relative improvement (10db SNR) in the phonetic classification experiments when we combine video and audio. The second method which uses separate sets of classes for audio and video, the relative improvement is about 17.76%. Note that the video only classification rates shown in Table 3 are for viseme based classes.

Table 4 gives the results for method 3 which emerges as the most important method. The rows designated

	Audio (phonetic)	Visual (Visemic)	Combined (Phonetic)
clean	47.67%	29.19%	49.51%
20db	36.26%	29.19%	38.10%
15db	27.90%	29.19%	30.36%
10db	21.77%	29.19%	25.64%

Table 3: Results for Method 2

	Audio	Visual	Combined
clean(visemic)	61.25%	29.21%	63.23%
clean(phonetic)	47.67%	21.12%	50.94%
20db(visemic)	57.65%	29.21%	60.69%
20db(phonetic)	36.26%	21.12%	45.18%
15db(visemic)	42.53%	29.21%	53.26%
15db(phonetic)	27.90%	21.12%	38.15%
10db(visemic)	35.96%	29.21%	49.57%
10db(phonetic)	21.77%	21.12%	34.34%

Table 4: Results for Method 3

as "visemic" represent results for the first phase where the classification is done based on 27 viseme classes. The "phonetic" rows show the overall result for phonetic classification after the second phase. We see a significant improvement in the viseme classification in the first phase. The overall improvement in the second phase outperforms all the other methods in which we get upto 57% relative improvement for the phonetic classification in 10 db SNR case.

5. CONCLUSION AND FUTURE WORK

From the above results, it can be concluded that multiple phase experiment performs better as compared to single phase experiments. We are working further to explore the optimal set of weights and number of viseme classes for the audio-visual speech recognition system.

6. REFERENCES

[1] David L. Hall, "Mathematical Techniques in multisensor data fusion", Artech House, 1992.

[2] Tsuhan Chen and Ram R. Rao, "Audio-Visual Integration in Multimodal Communication", Proceedings of IEEE, vol. 86, May 1998

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices", Nature, pp. 746-748, Dec. 1976

[4] K. Green, "The use of auditory and visual information in phonetic perception", Speechreading by Humans and Machines, D. Stork and M. Hennecke, Eds Berlin, Germany

[5] E. D. Petajan, "Automatic lipreading to enhance speech recognition", Proc. IEEE Global Telecommunication Conf., Atlanta, 1984

[6] E.D. Petajan, B. Bischoff, D. Bodoff and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition", Proc. CHI'88 pp. 19-25

[7] A. J. Goldschen, "Continuous automatic speech recognition by lipreading", Ph.D. dissertation, George Washington University, Washington, Sep, 1993

[8] Gerasimos Potamianos and Hans Peter Graf, "Discriminative Training of HMM Stream Exponents for Audio-Visual Speech Recognition", ICASSP'98

[9] Christopher Bregler and Yochai Konig, " "Eigenlips" for Robust Speech Recognition", ICASSP'94

[10] Christoph Bregler, Stefan Manke, Hermann Hild, Alex Waibel, "Bimodal Sensor Integration on the Example of "Speech Reading"", IEEE International Conference on Neural Networks, 1993

[11] Uwe Meier, Wolfgang Hürst, and Paul Duchnowiski, "Adaptive Bimodal Sensor Fusion for Automatic Speechreading", ICASSP'96

[12] C. Bregler, H. Manke, A. Waibel, "Improved Connected Letter Recognition by Lipreading", ICASSP'93

[13] Andrew Senior, "Face and feature finding for face recognition system", 2nd Int. Conference on Audio-Video based Biometric Person Authentication, Washingtion DC, March 1999

[14] S. Basu, C. Neti, A. Senior, N. Rajput, L. Subramanium, A. Verma, "Audio-Visual Large Vocabulary Continuous Speech Recognition in the Broadcast Domain", IEEE Workshop on Multimedia Signal Processing, Sep 13-15, Copenhagen 1999

[15] Mamoun Alissali, Paul Deleglise and Alexandrina Rogozan, "Asynchronous Integration of Visual Information in An Automatic Speech Recognition System", ICSLP'96