USING LIKELIHOOD L-STATISTICS TO MEASURE CONFIDENCE IN AUDIO-VISUAL SPEECH RECOGNITION

Arpita Ghosh
Indian Institute of Technology
Mumbai , India
email: arpit7ue@ccs.iitb.ernet.in

Ashish Verma, Abhinanda Sarkar
IBM India Research Lab
New Delhi, India
email: {vashish,sabhinan}@in.ibm.com

Abstract - This paper describes recent work on decision fusion in audiovisual speech recognition. In this work, a novel approach is proposed to combine audio and video channel information in audiovisual speech recognition scenario. We have considered frame-level phonetic classification problem using two single-stream Gaussian Mixture Models. Audio and video streams are adaptively weighted using a cumulative mean of the sample confidence values over past frames in addition to the present sample confidence value. The confidence values for audio and video decisions are computed using an L-statistics (linear combination of order-statistics) of log-likelihoods against phone models. It is shown through various experiments, on a database of about 15000 sentences from large vocabulary continuous speech, that the proposed approach results in better classification accuracy as compared to other approaches.

INTRODUCTION

Improving robustness of speech recognition against various noise types in the audio channel has become a focus area in the recent past. Video can help to achieve this as it is not affected by same noise types as audio and there is some kind of orthogonality between the audio and video channels [6, 7].

Integration of audio and video channels can be broadly categorized into feature fusion (early integration) and decision fusion (late integration). Specific approaches related to feature fusion are described in [2, 3, 6], and for decision fusion in [1, 4, 8, 12]. In feature fusion, both audio and visual feature vectors are combined to form an audio-visual feature vector which is used in audio-visual speech recognition. In decision fusion, separate recognition is performed on audio and visual feature vectors and their decisions are combined to get the final decision. It is observed that decision fusion can produce better results in audio-visual speech recognition as the recognition block for each stream can be optimized for that particular stream [12]. In the present work, we propose a novel approach for decision fusion in audio-visual speech recognition.

EARLIER APPROACHES

In decision fusion, log-likelihoods are computed for audio and video separately using audio and visual phone models and they are linearly combined to get the final

log-likelihood [1, 4, 8]. However, not much effort has been put into using an optimal, possibly dynamic set of weights, in the linear combination. People have either simply used entropy based weights [4], or they propose to use fixed weights learned from the training data [8]. Some researchers have also proposed to use weights according to the signal-to-noise ratio (SNR) level, present in the audio channel [8]. Recently, we have started looking at this aspect with the simple experiments of phonetic classification with audio-visual phonetic models. In [11], this problem is addressed in general, with estimation of stream exponents being used to calculate a combined audio-visual score corresponding to the phone models. In this paper, we address the issue of audio corrupted by noise in more detail. We explore the possibility of incorporating another confidence measure to tackle the noise present in the audio channel.

PROPOSED APPROACH

We know that audio and video have different units of speech, viz., phoneme and viseme respectively. Visemes are fewer in number as compared to phonemes and hence video has lower discriminating power among various speech sounds as compared to audio in normal circumstances. Video helps audio in speech recognition only in two conditions. First, when audio has lower than normal discriminating power because of the presence of background noise and second, when the sound being recognized is acoustically very close to another speech sound, for example '/m/' and '/n/' in spoken English. In clean conditions, therefore, video should be added to audio only when audio is confused due to acoustically confusable sound being recognized. Otherwise, video can even degrade the audio only recognition performance. This can be easily observed by looking at the corresponding log-likelihoods of audio and video for a few frames where acousitcally distinct sounds are being processed for example '/p/' and '/m/'. Similarly, in noisy conditions, although video should help audio for all the sounds, but the amount of importance (or weight) given to video decision should be dependent upon the noise level present in the audio channel.

To determine the confusability of the sound being recognized *and* the amount of noise level present in the audio channel, we propose to use the separation of likelihoods against phone models as a confidence measure. Moreover, we propose to use the likelihood separation in two different ways to obtain the weight of the channel in the fusion. First of all, what we call a *L-statistic* of the log-likelihoods is calculated for every frame, to determine the acoustic confusability of the sound being processed in the current frame, which we call *sample confidence*. Secondly, a cumulative mean of this sample confidence values is calculated over all the previous frames, and then used as a measure of the noise level present in the audio channel. We call this *overall confidence* of the channel.

Calculating Sample Confidence

For a given frame index k, L_{aki} denotes log-likelihood of audio frame corresponding to phone class i. Details for calculating the log-likelihood are standard in this field

and described in [11], where a similar approach is used. The index i is such that L_{aki}'s form order statistics, i.e.,

$$L_{ak1} \geq L_{ak2} \geq \ldots \geq L_{akN} \tag{1}$$

where N is the number of phone classes. It should be noticed that the index i does not represent any particular phone. It is 1 for most likely phone class and N for least likely phone class for a particular frame. Sample confidence for audio frame k, C_{ak}, is calculated as

$$C_{ak} = \sum_{i=1}^{n} \alpha_i * L_{aki} \tag{2}$$

where α_i's are the coefficients to choose a particular L-statistic. A prescription for the choice of the α_i's is described in Section .

Calculating Overall Confidence

It is evident from the discussion earlier in this section that in clean speech, the sample confidence should be low for acoustically confusable sounds and high for relatively easily distinguishable sounds. However, in noisy speech, there is another factor, noise, due to which the sample confidence value will vary and consequently be hard to measure reliably. It will be high for low noise level and low for high noise level. Hence, just by looking at a single frame or a few frames, one can not determine that low (or high) sample confidence value is due to acoustically confusable (or distinguishable) sound being recognized in the current frame, or due to the noise present in the audio channel. Therefore, we propose to use a cumulative mean of the sample confidence, over a large number of frames, to determine the noise level present in the audio channel. This overall confidence value, H_k, for the audio channel is computed as follows

$$H_k = (H_{k-1} * (k-1) + C_{ak})/k \tag{3}$$

We found the value of H_k to be well separated for various SNR levels. This approach to measure the noise level is easier compared to other approaches which fail in case of cafeteria (or cocktail) noise.

Weight estimation from sample and overall confidence

The weight given to the audio decision, at frame k, is calculated as a function of the two confidence values, namely, the sample confidence value and the overall confidence value, in the following way

$$w_{ak} = f_1(C_{ak}) + f_2(H_k) \tag{4}$$

We have chosen a simple threshold function as $f_1()$ and an exponential function for $f_2()$, which are described in the experiment section.

At each frame, the combined log-likelihood is computed for every phone class from the corresponding audio and video log-likelihoods (L_{aki} and L_{vki}) and their respective weights.

$$CL_{ki} = w_{ak} * L_{aki} + (1 - w_{ak}) * L_{vki}, \quad \forall i = 1, .., N \tag{5}$$

The phone class with highest CL_{ki} is chosen as the correct phone class for the frame.

EXPERIMENTS

We have performed phonetic classification experiments over VVAV database. Details about this database and system setup used to generate audio and video feature vectors can be found in [10]. In all the following experiments, we have used $\alpha_1 = 1$, $\alpha_2 = -1$ and $\alpha_i = 0$ for $i \neq 1, 2$ for the L-statistic. This results in taking the statistical separation of the first and second most likely phone classes which is also intuitively a measure of confidence. We tried to extend this and use the third most likely phone class also in the L-statistic which did not result in significant improvement over the first approach. For noisy audio, we have recorded cafeteria noise and added this to clean audio speech at different SNR levels.

In the first set of experiments, we used only the sample confidence value to assign weights. Whenever, the sample confidence was greater than a threshold, video was not used in the fusion ($w_a = 1$). If it was less than the threshold value, video was given an experimentally determined weight ($w_v = 0.29$). This threshold was learned from the training data.

In the next set of experiments, the cumulative mean of the audio sample confidence values was used to measure the noise level present in the audio channel and hence the weight assignment was made dependent upon this overall confidence value as well. The following exponential function was used to calculate the audio weight from the overall confidence value as described in Section .

$$f_2(H_k) = \beta/(1 + \exp(-b * (H_k - \gamma))) \tag{6}$$

where β, b and γ were learned from the training data once and they work well across all noise levels.

RESULTS AND DISCUSSION

Results for the experiments are presented in Table 1. The first and second rows of the table contain the phonetic classification results for audio only video only classification. The third row contains the results for the experiments where only sample confidence values were used. The fourth row has the results for the experiments where overall confidence of the audio channel was also used. The last row shows the results which were obtained by manually adjusting audio and video static weights for the best

Noise Level	Clean	15 DB	10dB
Only Audio	50.43%	39.68%	14.68%
Only Video	27.40%	27.40%	27.40%
Only sample Conf.	54.90%	44.73%	23.85%
Sample and Overall Conf.	54.90%	44.84%	29.84%
Manual Optimum	53.77%	43.51%	28.10%

Table 1: Results for Phonetic Classification

possible results. Here static means that the weights do not change during the experiment. We are using these results as a benchmark to compare our approach with the other approaches proposed in the literature which use static weights.

As we see from the third row, in case of clean speech, by just using the sample confidence values we get significant improvement in the results. If we compare these results with the results presented in [1, 4], we see a considerable improvement in our results. For example in [1], the improvement achieved in clean speech is about 5.51% while our results show about 8.86% improvement. Note the significant improvement in our results is because video is added to audio, *only* when audio is confused (has low confidence), and video is not (has high confidence). Otherwise, video is not used. We don't see much improvement by adding the overall confidence values in this case as there is no noise present.

However, in case of noisy speech, we get better results when we incorporate the overall confidence values in weight assignment process. This is due to the fact that on the detection of noise, more weight is given to video across all the phones as compared to clean speech. This improvement is very much evident in case of 10 dB SNR. In the 15 dB SNR case, we do not get significant improvement by using the overall confidence values. This is because at this particular noise level, the discrimination capability of audio is still better than video and hence video can only help audio because of the orthogonality present between the channels. Finally, note that we improve upon the use of static optimal (manual) weights in all cases.

CONCLUSION

In the present work, we show that likelihood separation can be used as a confidence measure to weigh audio and video streams in audio-visual speech recognition. Further, a mean value of the likelihood separation over large number of frames can be used to determine the noise level present in the audio channel and hence the weight of the audio stream can be adjusted automatically to get the optimal results.

ACKNOWLEDGEMENTS

The authors would like to acknowledge contributions to this work by Gerasimos Potamianos of IBM T. J. Watson Research Center, New York, for providing the visual

features used in the paper and his valuable comments about the work.

References

[1] M. Alissali, P. Deleglise and A. Rogozan, "Asynchronous integration of visual information in an automatic speech recognition system", *Proc. Int. Conf. Spoken Lang. Process.*, pp. 34-37, 1996.

[2] S. Basu, C. Neti, A. Senior, N. Rajput, L. Subramaniam, A. Verma, "Audio-Visual large vocabulary continuous speech recognition in the broadcast domain", *IEEE Workshop on Multimedia Signal Processing*, Copenhagen, pp. 475-481, 1999.

[3] C. Bregler and Y. Konig, " "Eigenlips" for robust speech recognition", *Proc. Int. Conf. Acoust. Speech Signal Process.*, Adelaide, pp. 669-672, 1994.

[4] C. Bregler, S. Manke, H. Hild and A. Waibel, "Bimodal sensor integration on the example of "Speech Reading"", *IEEE International Conference on Neural Networks*, 1993.

[5] C. Bregler, H. Manke and A. Waibel, "Improving connected letter recognition by lipreading", *Proc. Int. Conf. Acoust. Speech Signal Process.*, 1993.

[6] T. Chen and R. R. Rao, "Audio-Visual integration in multimodal communication", *Proceedings of IEEE*, pp. 837-852, vol. 86, 1998.

[7] K.P. Green, "The use of auditory and visual information in phonetic perception", *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke, pp.55-77, Eds Berlin, Germany.

[8] U. Meier, W. Hürst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading", *Proc. Int. Conf. Acoust. Speech Signal Process.*, Springer, Berlin, pp. 833-836, 1996.

[9] G. Potamianos and H.P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition", *Proc. Int. Conf. Acoust. Speech Signal Process.*, Seattle, pp. 3733-3736, 1998.

[10] G. Potamianos, A. Verma, C. Neti, G. Iyengar and S. Basu, "A cascade image transform for speaker independent automatic speechreading", *IEEE Conference on Multimedia and Expo*, New York, pp. 1097-1100, 2000.

[11] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition", *Proc. Int. Conf. Spoken Lang. Process.*, Beijing, vol. 3, pp. 746-749, Beijing.

[12] A. Verma, T. Faruquie, C. Neti, S. Basu and A. Senior, "Late integration in audio-visual continuous speech recognition", *Proc. Workshop on Automatic Speech Recognition and Understanding*, Colorado, 1999.