AUDIO-VISUAL SPEECH RECOGNITION IN VEHICULAR NOISE USING A MULTI-CLASSIFIER APPROACH

H. Karabalkan, and H. Erdoğan

Faculty of Engineering and Natural Sciences
Sabancı University, Tuzla 34956 Istanbul, Turkey
karabalkan@su.sabanciuniv.edu, haerdogan@sabanciuniv.edu

ABSTRACT

Speech recognition accuracy can be increased and noise robustness can be improved by taking advantage of the visual speech information acquired from the lip region. To combine audio and visual information sources, efficient information fusion techniques are required. In this paper, we propose a novel SVM-HMM tandem hybrid feature extraction and combination method for an audio-visual speech recognition system. From each stream, multiple one-versus-rest support vector machine (SVM) binary classifiers are trained where each word is considered as a class in a limited vocabulary speech recognition scenario. The outputs of the binary classifiers are treated as a vector of features to be combined with the vector from the other stream and new combining binary classifiers are built. The outputs of the classifiers are used as observed features in hidden Markov models (HMM) representing words. The whole process can be considered as a nonlinear feature dimension reduction system which extracts highly discriminatory features from limited amounts of training data. To simulate the performance of the system in a real-world environment, we add vehicular noise at different SNRs to speech data and perform extensive experiments.

1. INTRODUCTION

Automatic Speech Recognition (ASR) systems make a considerable contribution to human-computer interaction. Although ASR systems utilizing only audio information attain a recognition rate converging to 100% on noiseless data, recognition rates decrease significantly in situations when there is environmental noise [1][2]. One possible solution to overcome the noise problem is to use the visual speech information. Considering that the visual information is not affected by the audio noise, ASR systems utilizing both the audio and visual information achieve increased recognition accuracy and improved noise robustness [3].

Most modern ASR systems employ Mel Frequency Cepstral Coefficients (MFCC) as the audio features and Hidden Markov Models (HMM) as the modeling tool for the data. MFCC's are good representatives of speech information, however they are not discriminative features.

To extract discriminative features from MFCC features, one can use a tandem approach where first the features are passed through a multi-class classifier (mostly neural networks) that emits posterior probabilities of classes. Then, these posterior probabilities are used as more discriminative observation vectors for HMMs [4]. Fousek and Hermansky used parallel binary classifier structure with N binary classifiers to recognize N words [5]. The authors declare that parallel discriminative binary classifier structure reduces the insertion error for out-of-vocabulary words which is desired for a closed-set vocabulary system. In this work, a similar approach with Fousek and Hermansky is proposed but the parallel binary classifier structure is not used to estimate posterior probabilities to train an artificial neural network.

Support Vector Machines as binary classifiers are used instead to obtain a nonlinear transform of the MFCC feature vector. The SVM output is not converted into posterior probabilities but directly used. The same approach is also applied to the visual feature vector.

This work also proposes an efficient information fusion technique for audio and visual information. This technique appears to be somewhere in between *Decision Fusion* [6][7][8][9][10] and *Early Fusion* [7][11][12]. We combine information from audio and video streams using classifiers before feeding the outputs into HMMs.

The paper is organized in six sections, first section being the introduction. In sections 2 and 3, audio feature extraction and visual feature extraction methods are introduced respectively. Section 4 explains the combination of audio and visual information. Experiments and Results are presented in Section 5. Finally the paper is concluded with Section 6.

2. AUDIO FEATURE EXTRACTION

2.1. Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) have almost become the standard audio features in speech recognition systems. In this paper, 12 dimensional cepstral coefficients and the energy in the window are extracted to obtain static MFCC coefficients. Then, the feature vector is extended to a 39 dimensional vector by taking the first and second time-differences of static coefficients. The usual approach is to use this feature vector consisting of MFCCs as the observations for building Hidden Markov Models (HMM). Nevertheless, in this work, we also pass the audio feature vector through multiple binary classifiers to obtain more discriminative and noise-tolerant features.

2.2. Support Vector Machines

Support Vector Machine (SVM) is used as the binary classifier where it classifies a given audio frame either as one that belongs to the referred class or not according to the 39 dimensional MFCC feature vector. Assuming there are N-1 classes representing each digit and an additional one representing the silence and short pause, N binary SVMs are trained. Then, the outputs of all binary SVMs are combined to form an N dimensional audio feature vector. The audio feature extractor scheme is seen in Figure 1. Note that, SVM outputs are not posterior probabilities but can take negative or positive values depending on the classification.

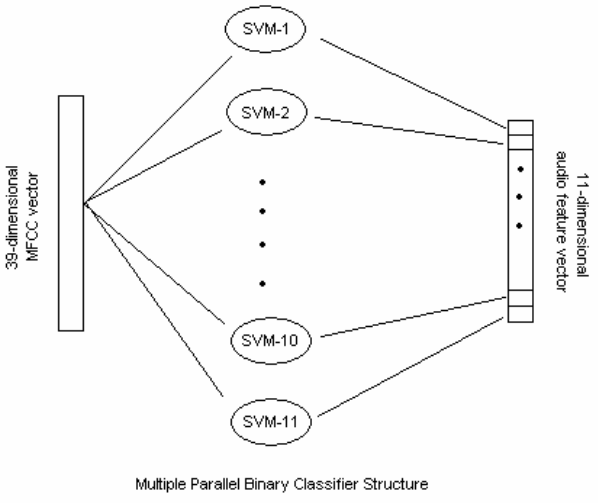

Figure 1: Audio Feature Extraction Scheme

One important point to mention in this process is that the training data has to be labeled to train an SVM. Class labels determined by audio-only speech recognition on noiseless data are considered as the true labels since a recognition rate higher than 97% can be achieved on noiseless data.

3. VISUAL FEATURE EXTRACTION

Despite the MFCCs as audio features, there is no such method for visual features that has found comprehensive acceptance. The visual feature extraction method proposed in this work can be analyzed in four main steps. First step is lip region extraction, second step is Principal Component Analysis (PCA), third step is the synchronization step and the last step is extraction of visual features as the outputs of SVMs.

3.1. Lip Region Extraction

In order to perform PCA to the lip, the lip region has to be extracted from the video. For this purpose, face detection is applied to video data and the initial lip region is extracted as the central 40% of the face horizontally and 30% of it above the bottom 5% vertically. After initializing the lip region to this area, we use correlation (as a means of tracking the region) between neighboring frames to enable continuous and smooth motion between frames (so that the derivative features make sense).

3.2. Obtaining Visual Features by Principal Component Analysis (PCA)

Now that the lip region is extracted, visual features can be obtained by means of PCA. PCA is a way of expressing data in such a way as to highlight their similarities and differences. In case of lips, PCA will return the eigen-lips and supply the information of how much of each eigen-lip is existent in a video frame. For it to perform well, PCA is not applied directly to the intensity values of frames, rather to the differences of the frames from the mean lip.

To find the mean lip and consequently the PCA matrix, i.e. matrix formed by concatenating eigenvectors/eigen-lips, a number of random frames are chosen from each video in the training database. The rest are not included due to computational efficiency.

It is obvious that as the number of samples increase, you get more representative eigen-lips. After PCA is applied to the difference images and the PCA matrix created, 30 most significant eigen-lips are computed. This reduces dimensionality as well as it reduces speaker independence. The visual features are extracted by multiplying the transpose of the matrix of eigen-lips with the frame in question. This multiplication operation returns a 30 dimensional vector for a frame with elements carrying information of how much of each eigen-lip is present at what amount.

3.3. Synchronization of Audio and Video

The visual feature vector obtained by PCA is at a frequency of 25 Hz (PAL video) whereas the audio feature vector is at a frequency of 100 Hz. The two data streams are synchronized by up-sampling the visual feature vector. A bandlimited interpolation is carried out by inserting zeros into the original feature vector and then applying a low pass filter. When the feature vectors from the two streams become synchronized, first derivatives of the visual features are calculated using the same approach as described in section (2.1). Finally, 60 dimensional visual feature vector at a frequency of 100Hz is created.

3.4. Visual Features by Support Vector Machines

Similar to audio feature extraction, the visual feature vectors are used to train binary SVMs for each class and the binary SVMs are combined to generate new forms of visual feature vectors which will be assigned as the observations for HMMs. The visual feature extraction scheme is represented in Figure 2.

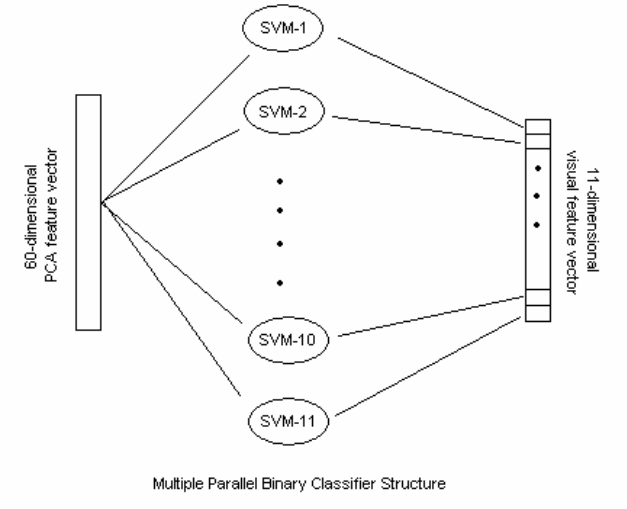

Figure 2: Visual Feature Extraction Scheme

4. COMBINING AUDIO AND VISUAL INFORMATION

There are two basic approaches to fuse audio and visual information. First approach, named *Early Fusion*, concatenates the features from the two streams, then implements classification. Second approach, named *Late Fusion*, implements classification for both streams separately and fuses the decisions of classifications. In this work, we propose a hybrid technique that merges information from two streams using classifiers but uses classifier outputs as observations to a HMM model. We compare our model to feature concatenation and LDA approaches.

4.1. Feature Concatenation

To investigate the performance of Early Fusion, 39 dimensional audio feature vector and 60 dimensional visual feature vector are concatenated and the resulting 99 dimensional bimodal feature vector is directly accepted as the observations of HMMs.

4.2. Linear Discriminant Analysis (LDA) Approach

Combination of audio and video information using the Linear Discriminant Analysis (LDA) approach is a well-known way to fuse two streams [13]. The purpose of LDA is feature selection. LDA achieves feature combination by selecting the best set of features from the two streams and it also provides linear dimensionality reduction through multiplication by a rectangular matrix. N dimensional feature vector is transformed into M dimensional feature space such that the class separability is maximum. This method is investigated to be compared with the proposed approach.

4.3. Support Vector Machine Approach

The outcomes of audio and visual feature extraction stages are N dimensional audio feature vector and N dimensional visual feature vector for the proposed approach. These two feature vectors are concatenated to form a 2N dimensional feature vector which is subsequently used to train another multiple parallel binary SVM classifier block as seen in Figure 3. The resulting N dimensional vector is the audio-visual feature vector and also an observation for HMM. So, the audio-visual feature vector automatically fuses the information from both streams.

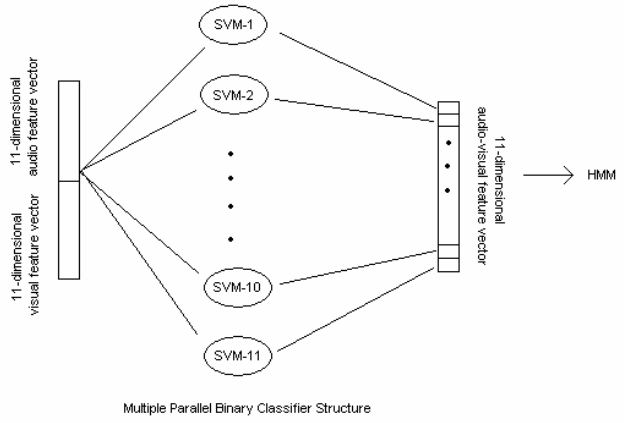

Figure 3: Feature Fusion Scheme

5. EXPERIMENTS AND RESULTS

5.1. Database

M2VTS database is used for experiments. This database consists of 185 videos from 37 different speakers. We show lip-regions extracted from 2 subjects' videos in Figure 4. Subjects count digits from zero to nine in order in French. We use a simple digit grammar of length 10 to recognize the sentences uttered. The database is extended by adding vehicular noise from NOISEX database at -10dB, -5dB, 0dB, 5dB, 10dB, 15dB and 20dB to test the system on noisy data.

Figure 4: Two example lip regions extracted from data.

5.2. Experimental Tools Used

Variety of tools is used for implementation. HMMs are built in Hidden Markov Model Tool Kit (HTK). Audio features are also extracted with HTK. The visual features are extracted in MATLAB and converted to HTK format. Lip region extraction stage is implemented in Visual C++. The tool *SVMLight* is used to implement SVMs.

5.3. Experiments

Extensive experiments are carried out to test the system. The results are presented in Table I.

SNR	**A**	**A + V** (concat)	**A + V** (LDA39)	**A + V** (LDA11)
-10dB	67.3	58.0	72.8	49.4
-5dB	81.1	59.0	87.0	61.4
0dB	87.0	62.7	89.7	80.0
5dB	94.9	64.2	92.5	85.7
10dB	96.5	68.7	94.2	88.9
15dB	96.8	69.0	94.2	90.3
20dB	97.0	69.7	94.4	89.7

Table I: Word accuracy rates (%) using various features. Visual-only accuracy is 36.8%.

We can see that, since the visual features yield a very low accuracy rate, audio is highly dominant in this database. Visual-only accuracy rate is at 36.8% only. We attribute this to the fact that, in real life videos, we need to correct for shifts and rotations in the lip image to enable uniform feature extraction across subjects. Currently, we do not perform shift and/or rotation normalization. Thus, especially for high SNR values, it is beneficial to use only the audio information.

We also observe that the LDA method when projected to 11 dimensions do not help to improve the recognition rate. However, when projected dimension size is 39 (from 99), we see improvements in accuracy rate for SNR values below 0 dB. Currently, experiments are underway for the SVM-HMM tandem-hybrid method proposed in this paper. The method promises good accuracy rates throughout different SNR values. Even though it uses only 11 dimensions, we are expecting to achieve good results.

6. CONCLUSION AND FUTURE WORK

We have been experimenting with multiple binary classifier feature combinations for audio-visual speech recognition. We expect that the proposed approach yields good results across different SNR values. We plan to extend our work to other streams of information (such as geometric visual features, other audio features) to enable most discriminant use of information resources to enable better speech recognition accuracies in challenging environments. We will also look for better normalization methods for lip region images since otherwise shifts and rotations in the lip image yield noisy visual parameter estimates. We are in the process of running experiments with the data collected under the frameworks of the Drive-Safe and NEDO initiatives [14].

7. ACKNOWLEDGEMENT

This work is partially supported by the State Planning Organization of Turkey (DPT) under the umbrella initiative called Drive-Safe Consortium, the NEDO collaborative grant titled International Research Coordination of Driving Behavior Signal Processing Based on Large Scale Real World Database from Japan, and the European Commission under Grant FP6-2004-ACC-SSA-2 (SPICE). The authors would like to acknowledge advises and contributions from the industrial and academic partners of the Drive-Safe Consortium and the collaboration of researchers from the NEDO Collaborative Research Program.

8. REFERENCES

[1] P. L. Silsbee and A. C. Bovik, “Computer lipreading for improved accuracy in automatic speech recognition,” IEEE Transactions on Speech, and Audio Processing, vol. 4, no. 5, 1996.

[2] J. N. Gowdy and Z. Tufekci, “Mel-scaled discrete wavelet coefficients for speech recognition,” Proceedings of ICASSP, 2000.

[3] Y. Zhang, S. Levinson and T. Huang, “Speaker Independent Audio-Visual Speech Recognition,” IEEE International Conference on Multimedia and Expo, 2000

[4] Hynek Hermansky, Daniel P.W. Ellis, Sangita Sharma, “Tandem Connectionist Feature Extraction for Conventional HMM Systems,” Proceedings of ICASSP 2000, vol. 3, 2000

[5] P. Fousek, and H. Hermansky, “Towards ASR Based on Hierarchical Posterior-Based Keyword Recognition,” Proceedings of ICASSP 2006, vol. 1, 2006

[6] 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.

[7] 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.

[8] U. Meier, W. Hurst, and P. Duchnowski, “Adaptive bimodal sensor fusion for automatic speechreading”, Proceedings of ICASSP, Springer, Berlin, pp. 833- 836, 1996.

[9] A. Verma, T. Faruquie, C. Neti, S. Basu and A. Senior, “Late integration in audiovisual continuous speech recognition”, Proc. Workshop on Automatic Speech Recognition and Understanding, Colorado, 1999.

[10] A. Ghosh, A. Verma, and A. Sarkar, “Using Likelihood L-Statistics to Measure Confidence in Audio-Visual Speech Recognition,” IEEE 4th Workshop on Multimedia Signal Processing, 2001

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

[12] 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. 475-48 1, 1999.

[13] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior, “Recent Advances in the Automatic Recognition of Audiovisual Speech,” Proceedings Of The Ieee, Vol. 91, No. 9, September 2003.

[14] H. Abut, H. Erdogan, A. Ercil, B. C¨ur¨ukl¨u, H.C. Koman, F. Tas, A.O. Argunsah, S. Cosar, B. Akan, H. Karabalkan, E. Cokelek, R. Ficici, V. Sezer, S. Danis, M. Karaca, M. Abbak. M.G. Uzunbas, K. Ertimen, C. Kalaycioglu, M. Imamoglu, C. Karabat, and M. Peyic, “Data collection with “uyanik”: Too much pain; but gains are coming,” in Biennial on DSP for In-Vehicle and Mobile Systems, June 2007.