ROBUST DETECTION OF VISUAL ROI FOR AUTOMATIC SPEECHREADING

G. Iyengar, G. Potamianos, C. Neti
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
USA
{giyengar,gpotam,cneti}@us.ibm.com

T. Faruquie, A. Verma
IBM India Solutions Research Center
New Delhi, India 110016
{ftanveer,vashish}@in.ibm.com

Abstract - In this paper we present our work on visual pruning in an audio-visual speech recognition scenario. Using visual information in speech recognition has been of interest because it can significantly improve performance in circumstances where audio-only recognition suffers (e.g. noisy environments). Tracking and extraction of region-of-interest (ROI) (e.g., speaker's mouth region) from video is an essential component of audio-visual (AV) speech recognition systems. It is important for the visual front-end in an AV speech recognition system to handle tracking errors that result in noisy visual data and hamper performance. In this paper, we present our robust visual front-end, investigate methods to prune visual noise and its effect on the performance of the AV speech recognition systems through phonetic and visemic classification experiments. We estimate the "goodness of ROI" using Gaussian mixture models and our experiments indicate that significant performance gains are achieved with pruned visual data.

INTRODUCTION

Many emerging multimedia and pervasive computing applications strive to model audio-visual events for the purposes of understanding, indexing and managing content. There is significant value in generating automatic transcripts and summarization of such audio-visual content. An example of such would be generating a textual transcription/keywords automatically from the speech portions of the audio-visual content. This holds the potential of enabling automatic text-based indexing and retrieval.

Automatic recognition of speech by using the video sequence of the speaker's lips, namely automatic lipreading, or speechreading, has recently attracted significant interest [1]-[12]. Much of this interest focuses on ways of combining the video channel information with its audio counterpart, in the quest for an AV automatic speech recognition (ASR) system that outperforms audio-only ASR. Such a performance improvement depends on both the audio-visual fusion architecture, as well as on the visual front end, namely, on the extraction of appropriate visual features that contain relevant information about the spoken word sequence. In this paper, we concentrate on the latter. We extract a visual region-of-interest (ROI) using a facial feature tracker. When tracking fails, the resulting visual features hamper the performance of ASR systems. In our approach, we post-process the resulting visual features for "goodness" and we discard sequences where poor tracking is detected. Experiments indicate that visual noise pruning results in significant improvements in the performance of automatic speechreading.

Various visual features have been proposed in the literature for speechreading that, in general, can be grouped into lip contour based and pixel based ones [3]. In the first approach, the speaker's lip contours are extracted from the image sequence. A parametric or statistical lip contour model is then obtained, and the model parameters are used as visual features. In the second approach, the entire image containing the speaker's mouth is considered as informative for lipreading, and appropriate transformations of its pixel values are used as visual features. In this paper, we follow the second approach. Briefly, our complete visual front-end consists of a facial feature tracker that extracts the speaker's mouth region. Once this ROI is extracted, it is verified for goodness with a classifier and appropriate visual features are extracted. For details on our novel visual front-end, please see [13]. In [14], we analyzed the effects of audio noise on the performance of AV speech recognition systems. In this paper, we focus on the effects of visual noise.

The paper is organized as follows. In Section , we detail our audio and video procesing. In Section , we present our mixture model for visual noise pruning followed by experiments in Section .

SYSTEM DESCRIPTION

We have a flexible architecture that permits extraction of different audio and video features. Similarly, our architecture permits different integration strategies for these two modalities. We will not focus on the integration aspects in this paper. For details on the various integration approaches used, see [14]. In the experiments presented in this paper, we report on visual phonetic and visemic classification performance as our focus is to isolate the effects of visual noise in the overall system performance. We now briefly describe the audio and video front-ends used in our system.

Audio Processing

We extract 24-dimensional mel-cepstral coefficient feature vectors from the audio signal using the standard techniques in speech recognition. To reduce dimensionality and capture dynamics, we use LDA (linear discriminant analysis). Specifically, in addition to the current frame, we take four previous and four succeeding audio frames and project them on to a 60 dimensional vector using LDA. A more elaborate description of our audio processing front-end is detailed in [11].

Video Processing

We use Fisher discriminant and eigenspace based face detection approach to extract the face and locate facial features from video [15]. 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 are used to extract and verify the locations of the important facial features, including the lip corners and centers. Subsequently, a size-normalized mouth image of size 45 × 30 pixels is extracted from the face image centered around the lips.

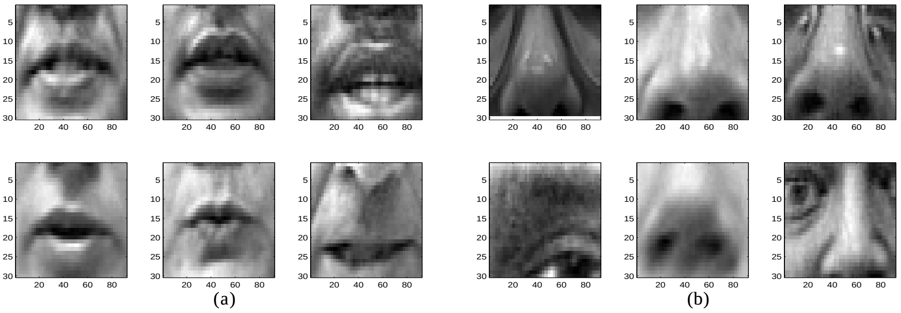
Once a suitable ROI image is extracted and verified (see Section for verification details), we use the novel pixel based visual front end for automatic speechreading, proposed in [13], that results in improved recognition performance of spoken words or phonemes. The algorithm is a cascade of three transforms applied to a 3-dimensional video region of interest that contains the speaker's mouth area. The first is a typical image compression transform aiming at a high "energy", reduced-dimensionality representation of the video data. In this work, principal component analysis (PCA) is used. The second is a linear discriminant analysis (LDA) based data projection applied to a concatenation of a small number of consecutive image transformed video data. The final data rotation is a maximum likelihood linear transform (MLLT) that optimizes the likelihood of the observed data given the traditional assumption of their class conditional Gaussian distribution with diagonal covariance [16].

All three steps, requires obtaining feature vector statistics from the ROI [13]. In the case of LDA and MLLT, knowledge of data class membership is required [13, 16].

PRUNING TRACKING NOISE AND ROBUST ROI DETECTION

We have observed that the face tracking system occasionally fails to track the face in the video sequence. This can be either due to mismatch between training and test conditions, or the candidate face is unlike any of the training examples, implying inability of the face model to generalize. In addition, the face tracking can also be poor, where the located face does not align accurately with the actual face in the video stream. In poor tracking, the visual processing results in geometry errors (e.g, nose tip marked as a lip) which gives rise to noise in the visual data. We note here that this noise is different from the signal noise (i.e, noise in video stream, per se). Figure 1, part (a) shows some successfully tracked regions-of-interest and part (b) shows some typical face tracker errors. Our experiments indicate that noisy visual data results in poor system performance. Therefore, verification techniques are needed to identify and prune potentially noisy tracking output.

Our approach to handling noisy visual data is to verify the output of the tracker and accept only those sequences that pass the verification stage. For verification, we use a classifier trained to differentiate between lips and non-lips. This classifier is a Gaussian mixture model (GMM) trained



(a) (b)

Figure 1: Successfully tracked ROI images in (a) and failures in (b)

on a small subset of PCA projections (typically 20-25 dimensions). As part of the pruning process, we classify the extracted ROI projections in a sequence and consider only those sequences that have a high percentage of “good” lips. In addition, we are currently integrating this classifier with the face tracker to identify and rectify geometry errors at the tracking stage.

Lip classification

Each mixture model is a semi-parametric density model (shown in Equation (1) below) of a particular observation class. Our lip classifier is composed of two density models, one for lips and one for non-lips. The likelihood of an observation vector $\mathbf{y}$ under the class θ is specified as

$$P(\mathbf{y}|\theta) = \sum_{i=1}^{N_c} \pi_i g(\mathbf{y}; \mu_i, \sigma_i) \tag{1}$$

where $\mathbf{y}$ is the observation vector (in our case, M-dimensional PCA projection), $g(\mathbf{y}; \mu_i, \sigma_i)$ is an M-dimensional Gaussian density with a mean vector μ_i and diagonal covariance σ_i, and the π_i are the mixture parameters of the components satisfying $\sum_i \pi_i = 1$. The GMM is completely specified by the class parameter vector $\theta = \{\pi_i, \mu_i, \sigma_i\}, i = 1, \ldots, N_c$. For each observation class (lips and non-lips), we estimate a parameter vector θ from the training examples using the Expectation-Maximization (EM) algorithm. Classification of a new observation vector $\mathbf{y}$ is taken as the class with the higher likelihood.

In order to train the classifier, we accumulated a small set of lip and non-lip images (50 lip images and 36 non-lip images) for initial training. To simplify identification of good ground truth images, we adopted the following approach: We start with the bootstrapping set above and train a few different mixture models. Specifically, we trained with 20 and 25 PCA dimensions, and with one and two mixtures per class resulting in 4 different GMM classifiers. We iteratively ran these classifiers on a test set and identified classification errors (i.e., lip classified as non-lips and vice versa). These ROIs which were incorrectly classified were added to the appropriate training set and the classifiers were trained again. This procedure is akin to boosting in statistical learning literature. These iterations were stopped when the test set errors started rising again (indicating over-training). Our final classifier was trained on 220 lip images and 200 non-lip images. Table 1 shows the classifier performance on the training set for the various mixture models that were used. For final integration with the system, we selected the mixture model with 25 PCA dimensions and 2 mixtures per class.

Lip classifier evaluation

The performance of the lip classifier is presented in Table 2 below. The classifier is implemented as a 3 way classifier (i.e., if the likelihood of neither of the two classes is higher than a threshold, the classification is marked as unknown). In Table 2, column 2 shows the human evaluated (ground-truth) percentage of lips in the sequence and the next two columns show the percentage classifications of the various classes. We tested the performance on 3 video sequences (labeled Seq1-3), each approximately 8-11 seconds long (roughly translates to 800-1100 lip projections after interpolation

Seq	PCA	Gauss.	True class		Classif.	
			Lip	Non	Lip	Non
Lips	20	1	100	0	68.0	18.0
Non Lips	20	1	0	100	11.4	75.5
Lips	20	2	100	0	89.5	1
Non Lips	20	2	0	100	.5	92.9
Lips	25	1	100	0	70	17
Non Lips	25	1	0	100	11.4	82.6
Lips	25	2	100	0	92.5	1.5
Non Lips	25	2	0	100	1.1	95.6

Table 1: Lip classifier results for Training datasets

of video data from 30 Hz to 100 Hz to match the audio feature rate). For testing, we present the results only for the best classifier which in this case corresponds to 25 PCA dimensions and 2 gaussian mixtures per class.

Seq	True	Classification (%)	
	Lip%	Lip	Non Lip
Seq1	100.0	96.0	3.7
Seq2	68.9	66.4	33.4
Seq3	36.5	35.8	63.9

Table 2: Lip classifier results for Test datasets

We note here that in the context of this experiment, we are interested in an estimate of the visual noise. For this purpose, it is adequate to get a lip classification percentage that is close to the true percentage of lips in the data. It is not necessary to consider the false alarm and false reject numbers. We note here that in pruning the audio-visual data, we accept only those sequences that have a percentage of lips greater than a set threshold. In this context, classifying non-lips as unknown is acceptable and need not be considered as an error.

EXPERIMENTS

We have collected two multi-subject, continuous, large vocabulary, audio-visual databases, using ViaVoice$^{\text{TM}}$ training sentences. The first contains frontal video of the mouth region of 79 subjects and consists of about 10 hours of speech, whereas the second contains full frontal face of 162 subjects and consists of about 25 hours of speech. The video is captured at a resolution of 704 × 480 pixels (interleaved), a frame rate of 30 Hz, and is MPEG2 encoded in the YUV422 format.

In this version of the paper, we report visual-only phonetic and visemic classification experiments on a subset of the 162-subject dataset, containing 82 subjects and 6045 utterances, split into a training and test set of 5000 and 1045 sequences respectively. After applying the lip classifier on the "tracked" mouth region on both sets, we obtained 5 subsets of the data where the mouth region is tracked with decreasing accuracy, as listed in Table 3.

To have equal amount of data in all experiments, we consider 2165 training and 429 test sequences in all conditions, picked from their corresponding training and test supersets of Table 3 by random sampling.

Given the input video, we consider a 45 × 30 sub-sampled, monochrome ROI image centered at the subject's mouth and normalized for size. Visual features are extracted at 30 Hz using the three stage cascade algorithm described earlier [13], namely a 24-dimensional PCA projection, followed by LDA and MLLT. The PCA features are first interpolated to 100 Hz, so that they are aligned to the audio front end features, obtained from the database audio stream. In addition, cepstral mean subtraction (CMS) is applied element-wise to all features [5]. LDA is subsequently applied on the

Threshold	Total =	Train +	Test
90%	2594	2165	429
80%	3167	2651	516
70%	3618	3026	592
60%	4013	3353	660
50%	4307	3587	720

Table 3: Dataset at various levels of visual pruning. Threshold corresponds to greater than specified percentage of ROI classified as lips

Threshold	LDA	MLLT
	TR / TS	TR / TS
90%	22.71/21.64	23.63/22.78
80%	21.55/20.73	22.31/21.48
70%	21.20/21.23	21.86/21.73
60%	20.41/19.93	20.87/20.77
50%	19.25/18.75	19.66/19.04

Table 4: Phonetic classification results (% correct) for the 5-GMM system. TR corresponds to Training set performance and TS corresponds to Test set performance

vector consisting of 11 (or 15) consecutive 24-dim PCA-feature frames (at 100 Hz), projecting it onto a 41-dimensional space. Finally, a 41 × 41 size MLLT matrix is used to "rotate" the feature vector.

A phonetic alignment of the database frames into 52 phonetic classes is produced at 100 Hz using the audio stream and a suitable audio-only hidden Markov model (HMM). The training set sentence alignments are then used to train visual-only GMMs, based on the visual front end described above.

We use 52 mixture models, with 5, or 32, Gaussians each and the EM algorithm for training. Phonetic classification performance is computed by comparing the test set alignment labels based on the audio-only HMM to their classification based on the visual features and the corresponding visual-only GMMs.

Phonetic classification performance on the various sets are depicted in Tables 4 and 5, using 5-mixture, and 32-mixture per class GMMs and the LDA applied on 11 PCA-feature frames. Notice that, in general, the test set performance degrades, as the amount of "visual noise" increases. In addition, visual features obtained by means of MLLT outperform the ones obtained by using LDA only. Furthermore, our experiments indicate that using PCA only features (and their first and second temporal derivatives) without LDA or MLLT, results in lower performance, for example in the $> 90\%$ case, PCA-only results in 19.32 % phonetic classification accuracy (using 5 mixtures), as compared to the 22.78 % using MLLT. Therefore, we do not report PCA-only performance results in this paper. Notice also, that the 32-GMM system significantly outperforms the 5-GMM one.

Threshold	LDA	MLLT
	TR / TS	TR / TS
90%	27.19/23.29	28.54/24.17
80%	26.11/22.54	27.24/23.16
70%	24.42/21.43	26.28/22.72
60%	23.57/20.46	25.14/21.97
50%	22.83/19.50	24.59/21.20

Table 5: Phonetic classification results for the 32-GMM system

We compare phonetic classification accuracies for the five sets, using a 5-GMM system, but with

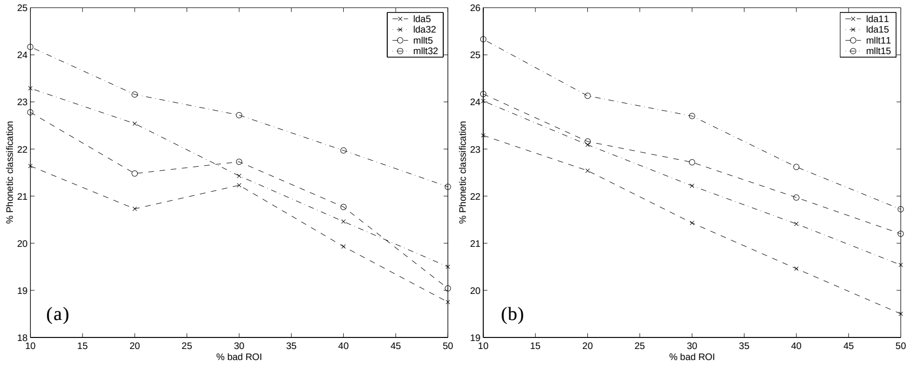


Figure 2: Phonetic classification performance of test set for 5-mixture GMMs vs 32-mixture GMMs, using 11 temporal frames in part (a) and Phonetic classification performance of test set for 11 temporal frames vs 15 temporal frames for 32-mixture GMMs in part (b)

the LDA applied on 15 consecutive PCA-feature frames, as opposed to the 11 frames considered in Tables 4 and 5. Figure 2 part (a) compares the relative phonetic classification performance of systems with increasing model complexity and Figure 2 part (b) compares the relative performance with increasing temporal window. It is clear that increasing the temporal window or model complexity does result in higher performance. However, the degradation of system performance with "visual noise" is consistent across all models. This underscores the need for identifying and compensating for visual noise in automatic speechreading systems. Similarly, the choice of visual front-end is important, as indicated by the superior performance of MLLT over LDA (and over PCA).

CONCLUSION AND FUTURE WORK

Using visual information for speech recognition is becoming an important topic in multimedia content analysis and coding. In this paper, we presented some sources of visual noise, its effect on system performance and approaches to prune this noise. It is clear that a systematic treatment of visual noise is an important requirement for robust system performance. Our experiments also indicate that choice of visual features is an important component of overall system performance – notice that MLLT performs the best and also tends to be more robust to visual noise compared to LDA. Our experiments indicate that while performance of automatic speechreading systems can be boosted by judicious choice of visual front-end, a systematic treatment of visual noise is an important component of a robust speechreading system.

We are currently integrating this pruning process into the visual feature extraction stage by making verification a part of the feature tracker. This we believe will result in better overall system performance.

References

[1] T. Chen and R. R. Rao, "Audio-Visual Integration in Multimodal Communication", Proceedings of IEEE, vol. 86, pp. 837-852, 1998.

[2] C. Bregler and Y. Konig, "Eigenlips" for Robust Speech Recognition", Proc. Int. Conf. Acoust. Speech Signal Process., Adelaide, 1994.

[3] M.E. Hennecke, D.G. Stork, and K.V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems", in Speechreading by Humans and Machines, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 331-349, 1996.

[4] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel, "Toward movement-invariant automatic lip-reading and speech recognition", Proc. Int. Conf. Acoust. Speech Signal Process., Detroit, vol. 1, pp. 109-112, 1995.

[5] G. Potamianos, H.P. Graf, and E. Cosatto, “An image transform approach for HMM based automatic lipreading”, Proc. Int. Conf. Image Process., Chicago, vol. III, pp. 173-177, 1998.

[6] I. Matthews, J.A. Bangham, and S. Cox, “Audio-visual speech recognition using multiscale nonlinear image decomposition”, Proc. Int. Conf. Speech Lang. Process., Philadelphia, vol. 1, pp. 38-41, 1996.

[7] N.M. Brooke, “Talking heads and speech recognizers that can see: The computer processing of visual speech signals”, in Speechreading by Humans and Machines, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 351-371, 1996.

[8] M.S. Gray, J.R. Movellan, and T.J. Sejnowski, “Dynamic features for visual speech-reading: A systematic comparison”, in Advances in Neural Information Processing Systems 9, M.C. Mozer, M.I. Jordan, and T. Petsche eds., MIT Press, Cambridge, pp. 751-757, 1997.

[9] G.I. Chiou and J.-N. Hwang, “Lipreading from color video”, IEEE Trans. Image Process., vol. 6, pp. 1192-1195, 1997.

[10] J. Luettin, “Towards speaker independent continuous speechreading,” Proc. Eurospeech, Rhodes, pp. 1991-1994, 1997.

[11] S. Basu, C. Neti, A. W. Senior, N. Rajput, L. Subramanium, and A. Verma, “Audio-Visual Large Vocabulary Continuous Speech Recognition in the Broadcast Domain”, Proc. Work. Multimedia Signal Process., Copenhagen, pp. 475-481, 1999.

[12] G. Potamianos and H.P. Graf, “Linear discriminant analysis for speechreading,” Proc. Works. Multimedia Signal Process., Los Angeles, pp. 221-226, 1998.

[13] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, “A cascade image transform for speaker independent automatic speechreading,” Submitted IEEE Intern. Conf. Multimedia Expo., New York, 2000.

[14] A. Verma, T. Faruquie, C. Neti, S. Basu, and A. W. Senior, “Late Integration in Audio-Visual Continuous Speech Recognition”, Proc. Works. Autom. Speech Reco. Underst., Keystone, 1999.

[15] A. W. Senior, “Face and feature finding for face recognition system”, 2nd Int. Conf.Audio-Video based Biometric Person Authen., Washington, p. 154-159, March 1999.

[16] R. A. Gopinath, “Maximum likelihood modeling with Gaussian distributions for classification”, Proc. Int. Conf. Acoust. Speech Signal Process., Seattle, vol. 2, pp. 661-664, 1998.