

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 Ctr
New Delhi, India 110016
{ftanveer,vashish}@in.ibm.com

Abstract - In this paper we present our work on visual pruning in an audio-visual (AV) speech recognition scenario. Visual speech information has been successfully used 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 such systems. It is important for the visual front-end 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. Specifically, we estimate the "goodness of ROI" using Gaussian mixture models and our experiments indicate that significant performance gains are achieved with good quality visual data.

INTRODUCTION

Automatic recognition of speech using the video sequence of the speaker's lips, namely speechreading, has recently attracted significant interest [1]-[6]. Much of the work focuses on ways of combining the visual channel 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 the audio-visual fusion architecture, as well as on the visual front end, namely, on the extraction of appropriate visual features. In our approach, we detect and track the face and detect landmark facial features to extract a suitable ROI for speech reading. For more details on our complete visual front-end, please see [5]. So far, no attempt has been made to characterize the effects of visual noise due to tracking in the visual front-end. In this paper, we characterize the effect of visual noise on visual phonetic classification performance and describe a method based on lip region classification for visual noise pruning. The novelty of the work is that it demonstrates the importance of the quality of visual speech representations in AV speech recognition systems. To our knowledge, such a study has not been done before.

SYSTEM DESCRIPTION

Audio Processing

We extract 24-dimensional mel-cepstral coefficients from audio using standard techniques in speech recognition. To reduce dimensionality and capture dynamics, we use LDA (linear discriminant analysis). Specifically, in addition to the current

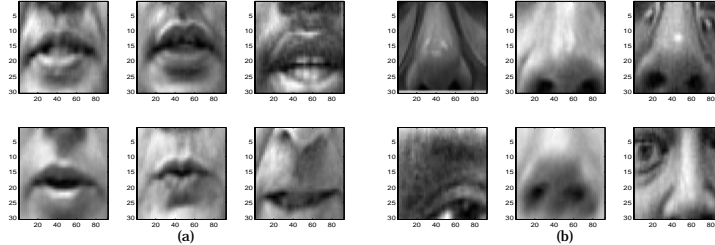


Figure 1: Successfully tracked ROI images in (a) and failures in (b)

frame, we take four previous and four succeeding audio frames and project them on to a 60 dimensional vector using LDA.

Video Processing

We use Fisher discriminant and eigenspace based face detector to extract the face and locate facial features from video [7]. An image pyramid over permissible scales is used to search the image space for possible face candidates. Once a face has been detected, an ensemble of feature detectors are used to extract the locations of important landmarks, including 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, we use the pixel based front end proposed in [5].

PRUNING TRACKING NOISE AND ROBUST ROI DETECTION

We observe that face tracking occasionally fails to track the face in the video. 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. These result 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 signal noise (i.e, noise in video stream, per se). Figure 1(a) shows some successfully tracked ROIs and Figure 1(b) shows some errors. Our experiments indicate that noisy visual data results in poor system performance.

Our approach 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 on a small subset of PCA projections (typically 20-25 dimensions). As part of the pruning process, we classify the ROI projections in a sequence and consider only those sequences that have a high percentage of “good” lips.

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, lips and non-lips. The likelihood of an observation vector \mathbf{y} under the class

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

θ is specified as

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

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 $\theta = \{\pi_i, \mu_i, \sigma_i\}, i = 1, \dots, N_c$. For each observation class (lips and non-lips), we estimate a parameter vector θ from the training examples using the Expectation-Maximization (EM) algorithm.

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 bootstrapping. To simplify identification of ground truth images, we adopted the following strategy: 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 development test set and identified classification errors. These incorrectly classified ROIs were added to the appropriate training set and the classifiers were retrained. 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.

The performance of the lip classifier on the test set 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 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(25 PCA, 2 GMM).

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 individually.

Seq	True Lip%	Classification (%)	
		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

EXPERIMENTS

We have collected two multi-subject, continuous, large vocabulary, audio-visual databases, using ViaVoiceTM 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 210 subjects and consists of about 60 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 this version of the paper, we report visual-only phonetic classification experiments on a subset of the 210-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.

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

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 normalized, monochrome ROI. Visual features are extracted using the three stage cascade algorithm described earlier [5]. The PCA features are first interpolated to 100 Hz, so that they are aligned to the audio features. In addition, cepstral mean subtraction (CMS) is applied element-wise to all features. LDA is subsequently applied on the 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 aximum-likelihood linear transform (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.

Threshold	5-GMM		32-GMM	
	LDA	MLLT	LDA	MLLT
	TR / TS	TR / TS	TR / TS	TR / TS
90%	22.71/21.64	23.63/22.78	27.19/23.29	28.54/24.17
80%	21.55/20.73	22.31/21.48	26.11/22.54	27.24/23.16
70%	21.20/21.23	21.86/21.73	24.42/21.43	26.28/22.72
60%	20.41/19.93	20.87/20.77	23.57/20.46	25.14/21.97
50%	19.25/18.75	19.66/19.04	22.83/19.50	24.59/21.20

Table 4: Phonetic classification results (% correct) for the 5-GMM and 32-GMM systems. TR corresponds to Training set performance and TS corresponds to Test set performance

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 Table 4 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 LDA followed by 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 LDA+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.

We compare phonetic classification accuracies for the five sets, using a 5-GMM system, but with the LDA applied on 15 consecutive PCA-feature frames, as opposed to the 11 frames considered in Table 4. Figure 2(a) compares the relative phonetic classification performance of systems with increasing model complexity and Figure 2(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 LDA+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

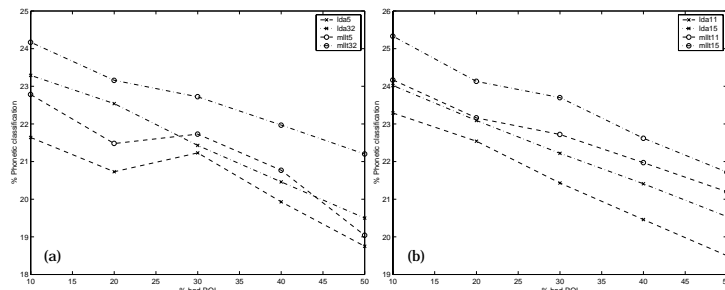


Figure 2: Phonetic classification performance of test set for 5-mixture GMMs vs 32-mixture GMMs, using 11 temporal frames in part (a) and Phonetic classification performance of test set for 11 temporal frames vs 15 temporal frames for 32-mixture GMMs in part (b)

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 a judicious choice of visual front-end, a systematic treatment of visual noise is an important component of a robust speechreading system. More importantly, our experiments indicate the importance of acquiring a good estimate of the visual features and underscore the importance of robust mouth region extraction for visual speech representations.

References

- [1] T. Chen and R. R. Rao, “Audio-Visual Integration in Multimodal Communication”, Proc. IEEE, vol. 86, pp. 837-852, 1998.
- [2] 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.
- [3] G.I. Chiou and J.-N. Hwang, “Lipreading from color video”, *IEEE Trans. Image Process.*, vol. 6, pp. 1192-1195, 1997.
- [4] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, “Audio-visual speech recognition,” *Final Workshop 2000 Report, CLSP, Johns-Hopkins, Baltimore, 2000.*
- [5] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, “A cascade image transform for speaker independent automatic speechreading,” *Proc. ICME*, vol. II, pp 1097–1100, New York, 2000.
- [6] A. Verma, T. Faruque, C. Neti, S. Basu, and A. W. Senior, “Late Integration in Audio-Visual Continuous Speech Recognition”, *Proc. Work. ASRU, Keystone, 1999.*
- [7] A. W. Senior, “Face and feature finding for face recognition system”, *2nd Int. Conf. AVBPA, Washington, pp. 154-159, 1999.*