COMPARISON OF LOW- AND HIGH-LEVEL VISUAL FEATURES FOR AUDIO-VISUAL CONTINUOUS AUTOMATIC SPEECH RECOGNITION

Petar S. Aleksic and Aggelos K. Katsaggelos

Department of Electrical and Computer Engineering
Northwestern University
2145 North Sheridan Road, Evanston, IL 60208
Email: {apetar, aggk} @ece.northwestern.edu

ABSTRACT

In this paper, we compare two different groups of visual features that can be used in addition to audio to improve automatic speech recognition (ASR), high- and low-level visual features. Facial Animation Parameters (FAPs), supported by the MPEG-4 standard for the visual representation of speech are used as high-level visual features in this work. Principal component analysis (PCA) based projection weights of the intensity images of the mouth area were used as low-level visual features. PCA was also applied on the FAPs. We developed an audio-visual ASR (AV-ASR) system and compared its performance for two different visual feature groups, following two approaches. The first approach assumes the same dimensionality for both high- and low-level visual features, while in the second approach the percentage of statistical variance described by the visual features used was the same. Multi-stream Hidden Markov Models (HMMs) and a late integration approach were used to integrate audio and visual information and perform continuous AV-ASR experiments. Experiments were performed at various SNRs (0-30dB) with additive white Gaussian noise on a relatively large vocabulary (approximately 1000 words) database. Conclusions were drawn on the trade off between the dimensionality of the visual features and the amount of speechreading information contained in them and its influence on the AV-ASR performance.

1. INTRODUCTION

The use of visual information in addition to audio, improves speech understanding especially in noisy environments [1]. Improving ASR performance, by exploiting the visual information of the speaker's mouth region is the main objective of AV-ASR [2, 3]. Many researchers have reported results that demonstrate AV-ASR performance improvement over audio-only ASR systems [2-7]. The performance of an AV-ASR system depends strongly on the audio-visual integration approach used and the choice of visual features. Visual features used in AV-ASR can be divided into three groups, high-level (lip contour based), low-level (image transform based), and combined features. In the high-level approach a model is used to describe extracted lip contours. The parameters that control the model of the lips are used as high-level visual features. The low-level visual features are obtained as a result of image transformations applied on the intensity values of the mouth area image. Combined visual features are obtained when both high- and low-level features are used to generate joint visual features. The advantage of the low-level visual feature extraction algorithms is that they do not require sophisticated methods for extracting high-level visual features. In addition, low-level visual features contain speechreading information that lies inside the mouth area that cannot be captured by high-level features. The disadvantage of low-level approaches is that they are generally sensitive to lighting and rotation changes. Their dimensionality is also usually much higher than the dimensionality of the high-level visual features, which affects reliable training of the AV-ASR systems. Combined and low-level features were compared by means of AV-ASR performance in [8]. In this paper the performance of high– and low-level visual features is compared. Some researchers also compared AV-ASR results using high- and low-level visual features on clean speech and using only digit or letter strings [9]. We performed continuous AV-ASR on a relatively large vocabulary (1000 words) database at various SNRs (0-30dB). In our work the equal dimensionality or equal statistical variance constraints are imposed in choosing the visual feature dimensionality. This was done in order to secure fair comparison of the AV-ASR performance changes that occur as a result of the choice of different visual features.

MPEG-4 is an audiovisual object-based video representation standard supporting facial animation. MPEG-4 facial animation is controlled by the Facial Definition Parameters (FDPs) and FAPs, which describe the face shape, and movement, respectively [10]. The MPEG-4 standard defines 68 FAPs, divided into 10 groups. In this work, only group 8 FAPs, which describe the outer lip movement, are considered. FAPs contain important visual information that can be used in addition to audio information in ASR.

In this paper, we first describe the visual features used and the visual feature extraction algorithms (Section 2). Next the audio-visual integration model used and the AV-ASR system are described (Section 3). Finally, the ASR experiments are described, and the paper is summarized in Sections 4 and 5.

2. VISUAL FEATURES

This work utilizes speechreading material from the Bernstein database [11]. This audio-visual database includes a total of 954 sentences, of which 474 were uttered by a single female speaker,

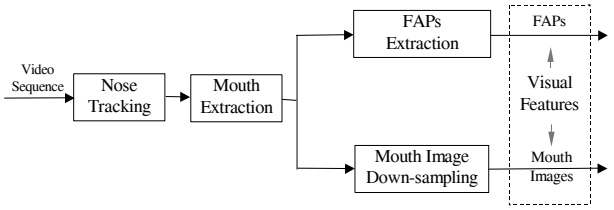


Figure 1. Visual feature extraction system.

(a) (b) (c)

Figure 2. (a) Neutral facial expression image; (b) Extracted nose template; (c) Extracted mouth image.

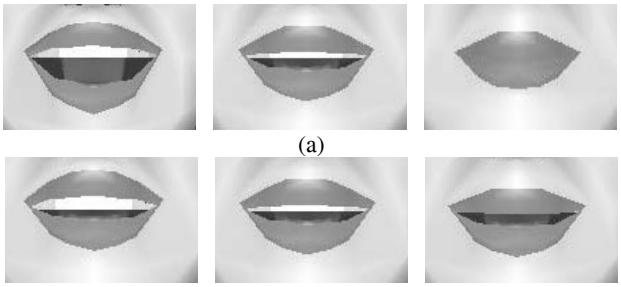
and the remaining 480 by a male speaker. For each of the sentences, the database contains a speech waveform, a word-level transcription, and a video sequence time synchronized with the speech waveform. The vocabulary size is approximately 1,000 words. The average utterance length is approximately 4 seconds. The video was sampled at a rate of 30 frames/sec (fps) with a spatial resolution of 320 x 240 pixels, 24 bits per pixel. The luminance information was used in the algorithms and the experiments. Audio was acquired at a rate of 16 kHz.

In ASR experiments we used both high level (FAPs) and low level (mouth images) visual features. In order to extract the visual features we developed an extraction system shown in Figure 1. An image of the nostrils (Figure 2b) was extracted from the neutral facial expression image (Figure 2a) to serve as a template for the template matching algorithm. The nostrils were chosen since they did not deform significantly during articulation. The template matching algorithm locates the nostrils in each video frame by searching the area centered at the neutral face nose location, for the best match. Once the nose location has been identified, a rectangular pixel region is extracted enclosing the mouth (Figure 2c). After the mouth image is extracted we combined active contour (snake) and parabola fitting algorithms in order to obtain the outer lip contour. The resulting lip contour was used to generate 10 Group 8 FAPs ($\boldsymbol{f}_t$) describing the outer lip movement. The feature extraction algorithm is described in detail in [6]. In order to decrease the dimensionality of the low-level video features the mouth image of each frame was down-sampled to obtain a final 19x33 mouth image (see Figure 1).

Through visual evaluation of the FAP extraction results we observed that the extracted parameters produced a natural movement of the MPEG-4 decoder [12] that synchronized well with the audio. Therefore, we concluded that the developed algorithm performed very well.

2.1. Visual features dimensionality reduction

In order to decrease the dimensionality of the visual feature vectors, Principal Component Analysis was performed on



(a)

(b)

Figure 3. The mean lip shape (middle column), and the lip shapes obtained by the variation of the first (a), and second (b) FAP eigenvector weights by +2 st. dev. (left column) and –2 st. dev. (right column) [12].

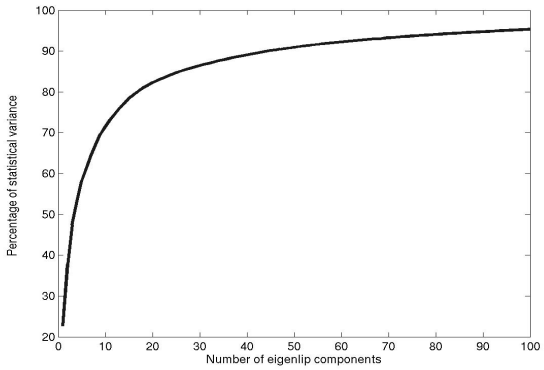


Figure 4. The additive statistical variance for a number of eigenlip components as a percent of the total variance.

both the 10-dimensional FAP vectors ($\boldsymbol{f}_t$), and mouth images. The FAP PCA training set consists of N FAP vectors, which are obtained from the training part of the visual data. The *10x10* covariance matrix $\boldsymbol{C}$ can be computed as

$$\boldsymbol{C} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{f}_n - \bar{\boldsymbol{f}})(\boldsymbol{f}_n - \bar{\boldsymbol{f}})^T , \qquad (1)$$

where $\bar{\boldsymbol{f}}$ denotes the mean FAP vector.

After the covariance matrix was obtained and its eigenvalues determined, the FAPs, $\boldsymbol{f}_t$, were projected onto the eigenspace defined by the first K eigenvectors,

$$\boldsymbol{f}_t = \bar{\boldsymbol{f}} + \boldsymbol{E}\cdot\boldsymbol{o}_t^f , \qquad (2)$$

where, $\boldsymbol{E}=[\boldsymbol{e}_1\ \boldsymbol{e}_2 ... \boldsymbol{e}_K]$ is the matrix of K eigenvectors, which correspond to the K largest eigenvalues, and $\boldsymbol{o}_t^f$ the *Kx1* vector of corresponding projection weights. The first six, two and one eigenvectors represent 99.6%, 93%, and 81% of the total statistical variance, respectively. By varying the projection

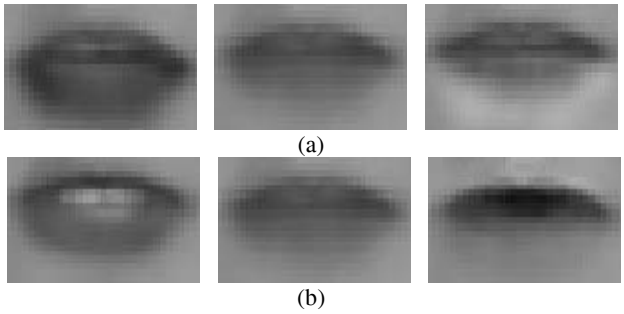


Figure 5. The mean mouth image (middle column), and mouth images obtained by the variation of the first (a), and second (b) eigenlip weights by +2 st. dev. (left column) and −2 st. dev. (right column).

weights by ±2 standard deviations (st. dev.), we concluded that the first and second eigenvector mostly describe the position of the lower and upper lip, respectively (Figure 3).

Similarly, PCA was performed on down-sampled mouth images. The dimensionalities of the mouth image meanvector, and the covariance matrix were 627x1 and 627x627, respectively. The distribution of the variance on the first 100 eigenvectors (i.e., eigenlips) is shown in Figure 4. The first 18, 12, nine, six and two eigenvectors represent 81%, 75%, 70%, 62%, and 38% of the total statistical variance, respectively. The mouth images obtained when the projection weights corresponding to the first two eigenlips were varied by ± 2 standard variations are shown in Figure 5.

3. AUDIO-VISUAL INTEGRATION

In this work, we utilized multi-stream HMMs and a late integration approach [13] for the combination of audio and visual speech information. The audio-visual system developed is shown in Figure 6. The audio feature vector ($\boldsymbol{o}_t^a$) consisted of 12 MFCC, an energy term, and the first and second order derivatives. The visual feature vector ($\boldsymbol{o}_t^v$) consisted of FAP projections weights ($\boldsymbol{o}_t^f$), or eigenlip coefficients and the first and second order derivatives. Since MFCCs were obtained at a rate of 90Hz, while visual features at a rate of 30Hz, visual features were interpolated in order to obtain synchronized data.

Audio and visual stream log-likelihoods are combined using the weights that capture their reliability. The audio-visual features were used to train a multi-stream HMM, with state emission probabilities given by

$$\boldsymbol{b}_j(\boldsymbol{o}_t^a, \boldsymbol{o}_t^v) = \prod_{s\in\{a,v\}} \left[\sum_{m=1}^{M_s} c_{jsm}\, \boldsymbol{N}(\boldsymbol{o}_t^s; \boldsymbol{\mu}_{jsm}, \boldsymbol{\Sigma}_{jsm}) \right]^{\gamma_s} \quad (3)$$

where subscript j denotes an HMM state, M_S denotes the number of mixtures in a stream, c_{jsm} denotes the weight of the m'th mixture of the stream s, and $\boldsymbol{N}$ is a multivariate Gaussian with mean vector $\boldsymbol{\mu}_{jsm}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_{jsm}$. The stream weights are denoted by γ_s, and they depend on the modality s. We assumed the weights satisfy $\gamma_a + \gamma_v = 1$.

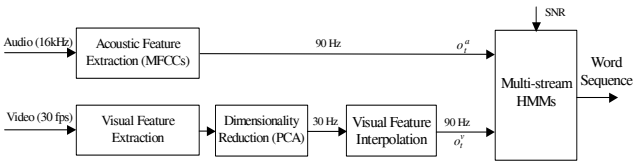


Figure 6. Block diagram of the AV-ASR system.

4. SPEECH RECOGNITION EXPERIMENTS

The baseline ASR system was developed using the HTK toolkit version 3.2 [14]. The experiments used the portion of the Bernstein database with the female speaker. Context dependent phoneme models, triphones, were used as speech units. Approximately 80% of the data was used for training, 18% for testing, and 2% as a development set for obtaining roughly optimized stream weights, word insertion penalty and the grammar scale factor. The bi-gram language model, used for decoding, was created based on the transcriptions of the training data set, and its perplexity was approximately 40. The same training and testing procedures were used for both audio-only and audio-visual experiments. To test the algorithm over a wide range of SNRs (0-30 dB), white Gaussian noise was added to the audio signals. All results were obtained using HMMs trained in matched conditions, by corrupting the training data with the same level of noise, as used for corrupting the testing data. This approach was used in order to accurately measure the influence of visual data on ASR performance. Audio-only ASR results are summarized in Figure 7, where WER represents the word error rate. It can be observed that the ASR performance is severely affected by additive noise.

4.1. Audio-visual speech recognition experiments

The AV-ASR experiments were performed using both high- and low-level visual features. When choosing the dimensionality of the visual features to be used for AV-ASR one should have in mind the trade-off between the number of HMM parameters that have to be estimated and the amount of the speechreading information contained in the visual features. In order to compare fairly the performance of the AV-ASR systems when low- or high-level visual features were used, based on the statistical variance distribution results obtained in Section 2.1, we performed the following experiments:

(i) In the first experiment the dimensionality (K) of both low-level and high-level visual features was chosen such that the variance corresponding to the first K eigenvectors is 81%. In this case the FAP visual feature size was one, and the low-level visual feature size was 18.

(ii) In the second experiment the dimensionality of both low- and high-level features was the same, six. The variance described by the first six eigenlips was 62%, and by the first six FAP eigenvectors 99.6%.

We also performed experiments in which the dimensionality of the low-level visual features (mouth image PCA projection weights) used were nine and twelve, and the dimensionalities of the high-level visual features (FAP PCA projection weights) used were two. Experiments were

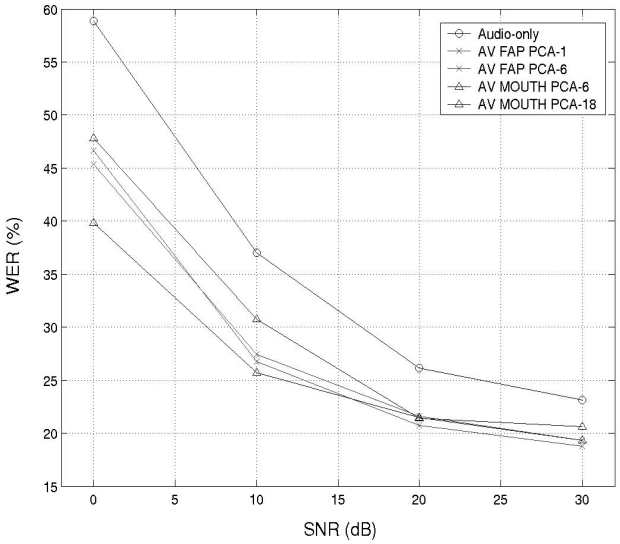


Figure 7. Audio-only and audio-visual system WERs vs. SNR.

performed at various SNRs (0, 10, 20, 30dB). The results obtained in experiments described by scenarios (i) and (ii) are shown in Figure 7. Results obtained by the additional experiments performed are shown in Table 1. As can be clearly seen, the AV-ASR systems perform considerably better than the audio-only ASR system, for all SNR values using both high- and low-level visual features. In scenario (i) the high dimensionality of the low-level visual features affected the reliable training of the HMMs and therefore the performance of the AV-ASR system. The performance of the high-level AV-ASR system was better by one to four percent for various SNRs. In scenario (ii) the low-level visual features provided much better results than high-level visual features for SNR of 0dB, and approximately the same results for SNRs of 22dB and 30 dB. The high-level visual features system performance was approximately the same in both scenarios. The fact that the system has similar performance for different dimension of visual features is due to the trade-off between the number of HMM parameters that have to be estimated, and the amount of the speechreading information contained in the visual features. How the AV-ASR performance is affected by this trade-off can be seen in Table 1.

5. SUMMARY

In this paper, an AV-ASR system that uses two different groups of visual features is described. Two testing scenarios for comparison of the quality of the speechreading information contained in visual features in terms of AV-ASR performance are defined. The system was evaluated for various SNRs (0-30dB) on a relatively large audio-visual database, for different values of the dimensionality of the visual feature vectors. Considerable improvement in ASR performance was obtained, for all noise levels tested, and for both visual feature groups. Conclusions were drawn on the trade off between the dimensionality of the visual features and the amount of speechreading information contained in them and its influence on the AV-ASR performance.

Visual Feature Group	Number of PCA Components	SNR			
		0	10	22	30
Mouth Images	6	60.17	74.26	78.52	80.65
	9	59.46	72.55	80.51	81.22
	12	57.61	71.89	80.09	81.37
	18	52.20	69.28	78.58	79.37
FAPs	1	53.34	73.26	79.23	81.22
	2	53.91	72.26	79.52	81.08
	6	54.62	72.55	78.38	80.65
Audio-only		41.96	64.02	75.68	77.24

Table 1. ASR systems performance.

6. REFERENCES

[1] R. Lippman, "Speech recognition by machines and humans," *Speech Communication*, vol. 22(1), pp. 1-15, July 1997.

[2] D. G. Stork and M. E. Hennecke, editors, Speechreading by Man and Machine, Springer-Verlag New York Inc., 1996.

[3] C. Neti et al., "Audio-visual speech recognition," Tech. Rep., Johns Hopkins University, Baltimore, 2000.

[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. of the IEEE*, vol. 91, no. 9, September 2003.

[5] S. Dupont, J. Luettin, "Audio-visual speech modeling for continuous speech recognition, "*IEEE Trans. on Mult.*, vol. 2(3), pp. 141-151, 2000.

[6] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features", *EURASIP Journal on Applied Signal Processing,* pp.1213-1227, 2002.

[7] P. S. Aleksic, J. J. Williams, Z. Wu, A. K. Katsaggelos, "Audio-visual continuous speech recognition using MPEG-4 compliant visual feature," *Proc. Int. Conf. on Image Processing (ICIP)*, pp. 960-963, Rochester, NY, September 2002.

[8] I. Matthews, G. Potamianos, C. Neti and J. Luettin, "A Comparison of model and transform-based visual features for audio-visual LVCSR," *Proc. Int. Conf. on Multimedia and Expo (ICME),* Tokyo, 2001.

[9] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," *Proc. Int. Conf. on Image Process*sing, vol. III, pp. 173-177, 1998.

[10] Text for ISO/IEC FDIS 14496-2 Visual, ISO/IEC JTC1/SC29/WG11 N2502, November 1998.

[11] L. E. Bernstein, Lipreading Corpus V-VI: Disc 3., Gallaudet University, Washington, D.C., 1991.

[12] G. A. Abrantes, FACE-Facial Animation System, version 3.3.1, Instituto Superior Tecnico, (c) 1997-98.

[13] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77(2), pp. 257-286, February 1989.

[14] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," Entropic Ltd., Cambridge, 2002.