APPLICATION OF AFFINE-INVARIANT FOURIER DESCRIPTORS TO LIPREADING FOR AUDIO-VISUAL SPEECH RECOGNITION

Sabri Gurbuz, Zekeriya Tufekci, Eric Patterson and John N. Gowdy

Department of Electrical and Computer Engineering
Clemson University
Clemson, SC 29634, USA
Email:{sabrig, ztufekc, epatter, jgowdy}@eng.clemson.edu

ABSTRACT

This work focuses on a novel affine-invariant lipreading method, and its optimal combination with an audio subsystem to implement an audio-visual automatic speech recognition (AV-ASR) system. The lipreading method is based on outer lip contour description which is transformed to the Fourier domain and normalized there to eliminate dependencies on the affine transformation (translation, rotation, scaling, and shear) and on the starting point.

The optimal combination algorithm incorporates a signal-to-noise ratio (SNR) based weight selection rule which leads to a more accurate global likelihood ratio test. Experimental results are presented for an isolated word recognition task for eight different noise types from the NOISEX data base for several SNR values.

1. INTRODUCTION

Using visual information in speech recognition has become an active research area because automatic speech recognition (ASR) performance degrades as acoustic background varies. Successful approaches for audio only ASR systems include representing the acoustic signal in ways that relate to the human auditory system, improved modeling of the acoustic phenomena and using of additional knowledge sources such as language modeling. These approaches have resulted in limited success in error reduction in audio only ASR systems depending on the task and the level and type of noise [1] [2]. It is well known that there is generally not enough information in the acoustic signal alone to determine the phonetic content of the message, especially as acoustic phenomenas varies from the ideal. The human audio-visual (HAV) system relies on additional knowledge sources to improve the recognition performance [3] [4] [5] [6].

Lipreading clearly meets at least two practicable criteria: It mimics human visual perception of speech recognition, and it contains information that is not always present in the acoustic signal. Petajan is one of the first researchers who built a lipreading system using oral-cavity features to improve the performance of an acoustic ASR system [7]. Silsbee et al. [1] utilized vector quantization (VQ) of acoustic and visual data for their HMM based audio and video subsystems. Teissier et al. [8] utilized 20 FFT based 1-bark wide channels between 0 and 5 Khz for acoustic features and inner lip horizontal width, inner lip vertical height and inner lip area for the visual features. Chiou et al. [9] utilized active contour modeling to extract visual features of geometric space, the Karhunen-Loève transform (KLT) to extract principal components in the color eigenspace, and HMMs to recognize the combined video only feature sequences. Potamianos et al. [3] used Fourier descriptor magnitudes for a number of Fourier coefficients, width, height, area, central moments, and normalized moments as contour features, and image transform features. It is worth noting here that the early visual feature extraction techniques are not affine (translation, rotation, scaling, and shear) invariant.

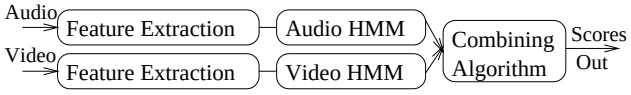

Fig. 1. Block diagram of the AV-ASR system.

Fig. 1 shows the block diagram of the AV-ASR system. We used affine-invariant Fourier descriptors (AI-FDs) to extract features of outer lip contour descriptions, and four affine-invariant oral cavity features (normalized outer lip width, normalized height, ratio of width to height, and the normalized inner area of the outer lip) for the video only ASR system. Invariance to affine transforms allows considerable robustness when applied to a sequence of speaker's lip images which may rotate and translate in the three dimensions while naturally speaking.

The proposed audio-visual automatic speech recognition (AV-ASR) system utilizes the lipreading system in conjunction with an existing audio only ASR system. This paper makes the following contributions:

1. It presents an outer lip contour detection algorithm, and a general affine invariant feature extraction method from outer lip contour images.
2. It presents an optimal decision combining algorithm for the audio and video HMM subsystems for a more accurate global decision.

This work is organized as follows. In section 2, we introduce the AV-ASR speech recognition system, and give brief overviews of the audio subsystem, video subsystem, the outer lip contour detection algorithm, and the affine-invariant Fourier descriptors extraction algorithm. Section 3 presents optimal decision combining algorithm for the audio and video HMM subsystems. In section 4, we describe the experimental setup and the results. Section 5 gives the concluding remarks and the proposed future work.

2. THE AV-ASR SPEECH RECOGNITION SYSTEM

This section describes the operation of the AV-ASR system. Note that the audio and the video HMM subsytems are entirely inde-

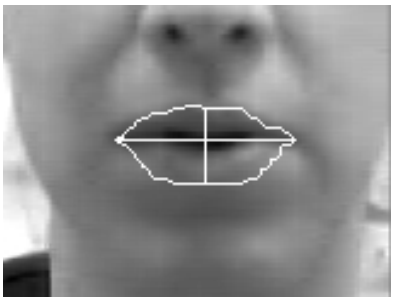

Fig. 2. Outer lip contour data with width and height line superimposed on the mouth image.

pendent until the decision combining algorithm is applied. Each subsystem described separately in the following sections.

2.1. The Audio Subsystem

The audio subsystem utilizes Mel-Frequency Discrete Wavelet Coefficients (MFDWC) as the audio features. The MFDWC are obtained by applying the Discrete Wavelet Transform (DWT) to the mel-scaled log filterbank energies of a speech frame. The complete description of the audio MFDWC feature extraction algorithm is described by Gowdy et al. in [2].

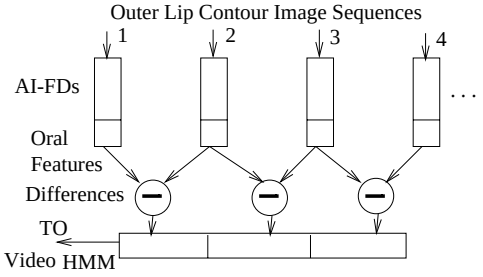

Fig. 3. Block diagram of the video feature extraction algorithm.

2.2. The Video Subsystem

Fig. 3 shows the block diagram of the video feature extraction algorithm. The algorithm extracts twelve AI-FDs of lip contour data as well as four affine-invariant oral cavity features which are width, height, ratio of width to height, and outer lip's inner area by normalizing next frame's corresponding oral cavity features.

Dynamic coefficients are obtained by differencing the consecutive image sequence features. Then these dynamic coefficients are utilized in the video HMM to generate log-likelihood scores.

2.2.1. The outer lip contour detection algorithm

The formulation of the outer lip contour detection algorithm is important for a successful video subsystem. The algorithm utilizes color images for lip contour extraction, and no prior labeling is required to obtain lip contour information. Here, our goal is to segment lip region and then detect the outer lip contour (edge). Steps are as follows.

- Divide every pixel value in the red plane by the corresponding green pixel value (if not zero), and threshold it with a proper red-to-green ratio threshold value for the lighting conditions [9].
- Smooth the binary candidate lip image using a smoothing filter, to obtain the smoothed image $\{S(x,y)\}$. The smoothing operation will taper the noise at the lip boundaries.
- Detect the edge image $\{E(x,y)\}$ in $\{S(x,y)\}$ and mask the smoothed image with the edge image which results gray level edges, $\{G(x,y)\} = \{E(x,y)\} \wedge \{S(x,y)$, and obtain the histogram of the gray level edge image and utilize the histogram for entropy based optimal threshold, T, selection.
- Threshold $\{S(x,y)\}$ with T to cope with edge curvature and noise, and segment the lip region (assumed to be the largest region) in the binary image using a recursive dynamic search algorithm, and then detect the ordered outer lip border pixel locations, $\{u[n], v[n]\}$, clockwise (or counter clockwise), where $n = 1 \ldots N$, and N is the number of lip border pixel locations in the video frame.

Fig. 2 shows the outer lip contour data superimposed on the mouth image. Here, $\mathbf{x[n]} = [u[n], v[n]]^t$ is a vector representation of a pixel location on the contour. The next section presents the extraction of the AI-FDs from the outer lip contour data, $\{x[n],\ n = 1, 2, \ldots, N\}$.

2.3. Affine-Invariant Fourier Descriptors From Fourier Transform Coefficients

Let $\mathbf{x^o} = \{x^o[n],\ n = 1, 2, \ldots, N\}$ be the outer lip contour data for N points on the lip contour in the reference image, and similarly $\mathbf{x}$ be the outer lip contour data in the observation image, where reference and observation images represent the training and test images, respectively, in the video HMM subsystem. For the lipreading application, possible affine transformations on the lip contour data can be translation, scaling, rotation, and shear (uneven scaling of rotation matrix), alone or combined. The relationship between $\mathbf{x}$ and $\mathbf{x^o}$ can be written as,

$$\mathbf{x} = A\mathbf{x^o} + \mathbf{b}, \tag{1}$$

where A represents a 2×2 arbitrary matrix, $det(A) \neq 0$, that may have scaling, rotation, and shearing affect, and $\mathbf{b}$ represents a 2×1 arbitrary translation vector. Therefore, we have a total of seven parameters to remove, which are four elements of A, two elements of $\mathbf{b}$, and the starting point. Thus, we need to construct an algorithm to generate a description of outer lip contour which is independent of all these parameters. Application of the AI-FDs for feature extraction is a new approach to lipreading system. Our work was motivated by the work of Arbter et al. [10]. The Fourier transform is applied to the data sequence $\mathbf{x}$ and resulting in a matrix of following Fourier coefficients

$$X = \begin{bmatrix} \cdots & U_0 & U_1 & \cdots \\ \cdots & V_0 & V_1 & \cdots \end{bmatrix}. \tag{2}$$

From the basic Fourier transform theory, we know that $U_{-k} = U_k^*$ and $V_{-k} = V_k^*$, (where * represents the complex conjugate). Therefore, we can discard all the coefficients $\begin{bmatrix} U_k & V_k \end{bmatrix}$ for $k < 0$.

We discard the pair $\begin{bmatrix} U_0 & V_0 \end{bmatrix}$ since it only depends on translation and conveys no shape information. The remaining coefficients are shift invariant. Let X_k, and X_k^o represent the k^{th} Fourier

transform coefficient vector resulting from the observation and reference, respectively. So we have

$$X_k = AX_k^o,\ k \neq 0. \tag{3}$$

We choose another p^{th} coefficient, where $X_p \neq 0$, and form the following matrix equation.

$$\begin{bmatrix} X_k & X_p \end{bmatrix} = A \begin{bmatrix} X_k^o & X_p^o \end{bmatrix}. \tag{4}$$

Now taking determinants of both sides, we get

$$det \begin{bmatrix} X_k & X_p \end{bmatrix} = det(A)\ det \begin{bmatrix} X_k^o & X_p^o \end{bmatrix}. \tag{5}$$

Notice that $det(A)$ is a scalar constant. We can define $I_k = det \begin{bmatrix} X_k & X_p \end{bmatrix}$, $I_k^o = det \begin{bmatrix} X_k^o & X_p^o \end{bmatrix}$, $I_p = det \begin{bmatrix} X_p & X_p^* \end{bmatrix}$ and $\mu = det(A)$, where we are free to use either coefficient itself or its conjugate in Eqn.4. Now by rewriting Eqn.5, it becomes

$$I_k = \mu I_k^o,\ k = 1 \ldots m, \tag{6}$$

where m represents the number of Fourier coefficients. Similarly, we can define

$$I_p = \mu I_p^o. \tag{7}$$

We can eliminate the effect of μ by defining a new set of coefficients, Q_k as

$$Q_k = \frac{I_k}{I_p},\ k = 1 \ldots m \tag{8}$$

for arbitrary constant p such that $X_p \neq 0$ for any $p > 0$. In the presence of noise effect, considering Eqn. 8, it is desirable to choose a p value which makes I_p as large as possible.

So far, we have developed a method to generate a description of outer lip contour independent of A and $\mathbf{b}$ in Eqn. 1. We now consider the starting point problem in the contour data. Let $n^o = n + \tau$, where τ is an arbitrary shift value. The relationship between $\mathbf{x}[\mathbf{n}]$ and $\mathbf{x^o}[\mathbf{n^o}]$ in Eqn. 1 without the $\mathbf{b}$ parameter is

$$\mathbf{x}[\mathbf{n}] = A\mathbf{x^o}[\mathbf{n} + \tau]. \tag{9}$$

From the basic Fourier transform theory, the k^{th} coefficient of the observation image data is related to k^{th} coefficient of the reference image data by

$$X_k = A\, e^{j2\pi\tau k/N}\ X_k^o, \tag{10}$$

where N is the period of the outer lip contour data. Eqn. 5 now be rewritten as

$$det \begin{bmatrix} X_k & X_p \end{bmatrix} = det(A)\ e^{j2\pi\tau(k+p)/N}\ det \begin{bmatrix} X_k^o & X_p^o \end{bmatrix}. \tag{11}$$

Then, Eqn. 8 becomes

$$Q_k = e^{j2\pi\tau(k-p)/N}\ Q_k^o,\ k = 1 \ldots m. \tag{12}$$

Phase shift can be eliminated by simply taking the absolute value of the both sides in Eqn 12. That is

$$\left|Q_k\right| = \left|Q_k^o\right|,\ k = 1 \ldots m. \tag{13}$$

Thus, we have shown that the $\mathbf{x}$ and $\mathbf{x^o}$ which satisfy Eqn. 1 with an arbitrary A, $\mathbf{b}$, and different starting point have the same AIFDs.

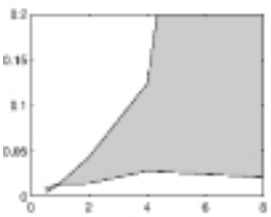

Fig. 4. λ (vertical axis) versus the linear SNR (i.e., -6dB = 0.5) for car noise.

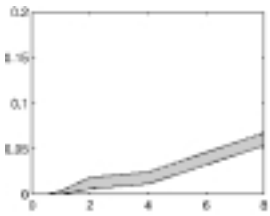

Fig. 5. λ (vertical axis) versus the linear SNR (i.e., +6 dB = 2) for speech noise.

3. COMBINING AUDIO AND VIDEO DECISIONS

Here we will briefly discuss how to combine the log-likelihood scores from the audio and video subsystems. For comparison purposes we adopt the same notation from Silsbee et al. [1]. Let $S_{ia} = log\, Pr(O_a \mid M_{ia})$, $S_{iv} = log\, Pr(O_v \mid M_{iv})$, and $S_i = log\, Pr(O_a, O_v \mid M_i)$. Here, i is the index of the word, $1 \leq i \leq W$ (W is the vocabulary size), M_i is the HMM for the i^{th} word, and O is the observation sequences of symbols. Thus, the combined likelihood for given log-likelihood scores from each subsystem is computed as

$$S_i = \lambda S_{ia} + (1 - \lambda) S_{iv}, \tag{14}$$

where λ is the influence factor. Clearly, increasing λ will increase the influence of S_{ia} while decreasing the influence of S_{iv}, and vice versa. Figs. 4 and 5 show λ versus linear SNR for car noise and speech noise from the NOISEX database, respectively. Any λ value within shaded area gives improved performance over the individual audio and video subsystems. By basing λ in Eqn.14 on the SNR annd the noise type, we obtain a combined optimal output score. The audio-visual recognition problem can then be regarded as that of computing

$$\arg\max_i \{log\, Pr(O_a, O_v \mid M_i)\},\ i = 1 \ldots W. \tag{15}$$

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Experimental Setup

The AV-ASR system is trained using the audio and the video features. Our current system is speaker-dependent and recognizes isolated words. When the speaker talks at a normal rate in the field of view (FOV) of the camera, a color AVI movie with audio is recorded. The speaker hits keys to start and stop recording for a word (i.e., isolated word recording). Then the lip contour detection algorithm extracts affine-invariant Fourier descriptors and affine-invariant oral-cavity features from video frames. In the audio subsystem, the MFDWC are extracted by applying the Discrete Wavelet Transform (DWT) to the mel-scaled log filterbank energies of a audio speech frame. Then, each HMM subsystem generates log-likelihood scores. Finally, a decision combining rule picks the optimal λ value based on the SNR and noise type for superior results.

The AV-ASR algorithm is tested for eight different noise types from the NOISEX data base by corrupting clean speech with noise signal (additive noise). Using the data sets of [9], we recorded the same ten words at 30 fps using an inexpensive PC camera, and collected the following ten isolated words from a single speaker: on, off, yes, no, up, down, forward, rewind, radio, and tape.

Table 1. Recognition Accuracy of audio and audio-visual (AV) system when used with proper weighting term λ (video only: 87.06%).

	Speech Noise			Lynx Noise			Operation Room Noise			Machine Gun Noise		
	λ	audio %	AV %	λ	audio %	AV %	λ	audio %	AV %	λ	audio %	AV %
Clean	0.0244	100.0	100.0	0.0244	100.0	100.0	0.0244	100.0	100.0	0.0244	100.0	100.0
18 dB	0.0525	94.7	99.4	0.0391	96.5	99.4	0.0397	98.8	100.0	0.0227	99.4	100.0
12 dB	0.0098	65.3	95.3	0.0097	71.8	95.9	0.0458	88.8	97.7	0.0279	97.1	99.4
6 dB	0.0189	36.5	92.4	0.0108	36.5	91.2	0.0159	58.8	94.1	0.0208	79.4	97.1
0 dB	0.0013	13.5	88.8	0.0018	15.9	88.2	0.0100	21.8	90.6	0.0087	64.7	94.7
-6 dB	0.0015	12.4	88.8	0.0015	12.9	88.2	0.0093	14.7	88.8	0.0096	45.3	91.8

Table 2. Recognition Accuracy of audio and audio-visual (AV) system when used with proper weighting term λ (video only: 87.06%).

	STITEL Noise			F16 Noise			Factory Noise			Car Noise		
	λ	audio %	AV %	λ	audio %	AV %	λ	audio %	AV %	λ	audio %	AV %
Clean	0.0244	100.0	100.0	0.0244	100.0	100.0	0.0244	100.0	100.0	0.0244	100.0	100.0
18 dB	0.0757	90.0	98.8	0.0388	97.1	99.4	0.0463	98.2	99.4	0.0204	100.0	100.0
12 dB	0.0165	62.4	95.3	0.0133	79.4	97.1	0.0715	88.2	98.2	0.0269	97.7	99.4
6 dB	0.0013	30.0	88.2	0.0105	40.6	91.8	0.0335	54.7	92.9	0.0262	88.2	97.7
0 dB	0.0000	13.5	87.1	0.0104	26.5	90.0	0.0260	24.7	91.2	0.0105	58.2	94.1
-6 dB	0.0000	11.2	87.1	0.0161	21.2	90.6	0.0316	23.5	89.4	0.0063	30.6	91.2

Each word has 17 examples (audio-video), each video sequence has about 25 to 60 color video frames (total 4070 frames), and each image frame contains R, G, B components of size 160×120.

As a validation procedure of the AV-ASR algorithm, we leave a set out as a test set and use all the other 16 sets for training of audio and video HMM subsystems. The audio signal in the test set is corrupted with additive noise from the NOISEX data base. Recognition is performed with unmatched training data. This process is repeated until every set in the data base is a test set.

4.2. Experimental Results

We are able to achieve 87.06% recognition accuracy with the visual information alone for ten isolated words of 170 test size with video subjected to arbitrary affine.

The optimal decision combining algorithm incorporates a novel noise type and SNR based weight selection rule which leads to a more accurate global likelihood ratio test. Table 1 and 2 show λ values (see Figs. 4 and 5), the audio subsystem's recognition accuracies, and the AV-ASR system's recognition accuracies for the various SNR values and noise types from NOISEX database.

5. CONCLUDING REMARKS AND FUTURE WORK

Invariance to affine transforms allows considerable robustness to a lipreading system for natural speaking. By basing the weighting term λ based on the SNR value and the noise type, the combined audio-visual system performs better than either individual subsystem under all acoustical conditions tested.

Future work will include adding a noise type and SNR detector algorithm to the AV-ASR system for continuous audio-visual speech recognition.

6. ACKNOWLEDGEMENTS

We would like to thank Sarah M. Murrell for data collection and analysis.

7. 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," in *Proceedings of ICASSP*, 2000.

[3] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for hmm-based automatic lipreading," in *Proceedings of ICIP*, 1998.

[4] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proceedings of ICASSP*, 1993.

[5] S. Basu, C. Neti, N. Rajput, A. Senior, L. Subramaniam, and A. Verma, "Audio-visual large vocabulary continuous speech recognition in the broadcast domain," in *Proceedings of IEEE Multimedia Signal Processing Conference (MMSP99), Denmark*, 1999.

[6] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an hmm-based asr," in *Speech Reading by Humans and Machines, D. G. Stork and M. E. Hennecke Eds. Springer, Berlin*, 1996.

[7] E. Petajan, B. Bischoff, D. Bodoff, and N. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *ACM SIGGHI, pp. 19-25*, 1988.

[8] P. Teissier, J. Robert-Ribes, J. Schwartz, and A. Guérin-Dugué, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Transactions on Speech, and Audio Processing*, vol. 7, no. 6, 1999.

[9] G. I. Chiou and J. Hwang, "Lipreading from color video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, 1997.

[10] K. Arbter, W. E. Snyder, H. Burkhardt, and G. Hirzinger, "Application of affine-invariant fourier descriptors to recognition of 3-d objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, 1990.