AUTOMATIC SPEECHREADING WITH APPLICATION TO SPEAKER VERIFICATION

C. C. Broun

X. Zhang, R. M. Mersereau, M. Clements

ABSTRACT

Speech not only conveys the linguistic information, but also characterizes the talker's identity and therefore can be used in personal authentication. While most of the speech information is contained in the acoustic channel, the lip movement during speech production also provides useful information. In this paper we investigate the effectiveness of visual speech features in a speaker verification task. We first present the visual front-end of the automatic speechreading system. We then develop a recognition engine to train and recognize sequences of visual parameters. The experimental results based on the XM2VTS database [1] demonstrate that visual information is highly effective in reducing both false acceptance and false rejection rates in speaker verification tasks.

1. INTRODUCTION

While speech primarily conveys the linguistic information, it also characterizes a talker's identity and, therefore, can be used for personal authentication. There is an increasing requirement for reliable personal authentication systems in areas of high security or secure access. The use of speech to identify a person has the advantage of requiring little custom hardware and is non-intrusive. However there are two significant problems in current generation speaker verification systems using speech. One is the difficulty in acquiring clean audio signals in an unconstrained environment. The other is that unimodal biometric models do not always work well with a significant percentage of the population. To combat these issues, systems incorporating the visual modality are being investigated to improve system robustness to environmental conditions, as well as to improve overall accuracy across the population. Face recognition has been an active research area during the past few years. However, face recognition is often based on static face images assuming a neutral facial expression and requires that the speaker not have significant appearance changes. Lip movement is a natural by-product of speech production, and it not only reflects speaker dependent static and dynamic features, but also provides "liveness" testing.

In this work, we incorporate visual speech features into a speaker verification system to assess the benefit of the visual modality. The paper is organized as follows. Section 2 presents our visual front end for the automatic speechreading system. The problem of speaker recognition using bimodal speech information is investigated in Section 3. Finally, Section 4 offers our conclusions.

2. VISUAL ANALYSIS

It is generally agreed that most visual speech information is contained in the lips. Thus, visual analysis focuses on lip feature extraction. The first stage of the visual analysis involves a region of interest (ROI) detection — locating the speaker's mouth in a video sequence.

In another study [2], we use hue and saturation thresholding to locate the lips. Although it works well for most cases, complications might occur when distracting red blobs exist in the clothing, or if the person has a ruddy complexion. To eliminate these non-lip red blobs, we developed strategies using gradient/geomeric/saturation constraints [3].

In the current paper we propose another solution that uses motion cues. Since a talking person's mouth is moving, we search for the moving lips in the image if an audio signal is present in the acoustic channel. To detect the moving object, we build difference images between subsequent frames and sum over a series of frames. The accumulated difference image (ADI) is defined as follows:

\[ ADI_0(x, y) = 0 \]
\[ ADI_k(x, y) = ADI_{k-1}(x, y) + \Delta R_k(x, y), k \in 1, \ldots, T \]

where the difference image \( \Delta R_k(x, y) \) is calculated by pixel-wise absolute subtraction between two adjacent frames \( \Delta R_k(x, y) = |R_k(x, y) - R_{k-1}(x, y)| \). Note we use the R component for our lip detection. \( T \) is set to 30 in this work. An example of an accumulated image is shown in Fig. 1(d). To separate the moving lips from the background, we require all pixels exceeding a certain value be assigned to 1, and 0 otherwise. We determine this threshold by using Otsu's method [4]. The resulting binary image is shown
in Fig. 1(e). Fig. 1(f) is obtained by using the AND operation to combine the binary image from the hue/saturation thresholding [2], as shown in Fig. 1(c), with the accumulated difference image. Based on the resulting image, we extract the lip region from its surroundings by finding the largest connected region. The identified lip area is shown as a white bounding box in Fig. 1(g). Subsequent processing is restricted to the identified lip region.

To derive the lip dimensions within the ROI, we make use of both color and edge information of an image. We developed an algorithm which uses a Markov random field framework to combine these two types of information and segment the lips from the background. Details of the MRF-based lip segmentation can be found in [5].

Fig. 2 shows examples of segmentation results with different persons and different lip opening situations. We observe that the highlighted pixels fairly well match the true lip area. Based on the segmented lip image, it is easy to extract the key feature points on the lips and subsequently derive the geometric dimensions of the lips. The following features are used in our study: mouth width ($w_2$), upper/lower lip width ($h_1$, $h_2$), lip opening height/width ($h_3$, $w_1$), and the distance between horizontal lip line and upper lip ($h_4$). An illustration of the geometry is shown in Fig. 3. Besides the geometric dimensions of the lip, we also detect the visibility of the tongue and teeth [5]. The indicator of the tongue/teeth is set to 1 if they are detected, and 0 otherwise.

![Fig. 2. Segmented lip overlayed on original image.](image)

### 3. SPEAKER VERIFICATION

Speaker verification seeks to validate a claimed identity, i.e., either to accept or reject an identity claim.

For the speaker verification task, we use the polynomial-based approach developed in [6]. Polynomial-based classification requires low computation while maintaining high accuracy. Because of the Weierstrass approximation theorem, polynomials are universal approximators for a Bayes classifier [7].

The classifier consists of several parts as shown in Fig. 4. The extracted features $x_1, \ldots, x_N$ are introduced to the classifier. For each feature vector $x$, a score is produced by using the polynomial discriminant function $f(x) = w^T p(x)$, where $p(x)$ is the polynomial basis vector constructed from the input vector $x$, $p(x) = [1 \ x_1 \ x_2 \ x_1^2 \ x_2^2 \ x_1 x_2]^T$ for a two-dimensional feature vector $x = [x_1 \ x_2]^T$ and for polynomial order two, and $w$ is the class model. The final score is computed by averaging over all feature vectors.

$$
\text{Score} = \frac{1}{N} \sum_{i=1}^{N} w^T p(x_i). \quad (1)
$$

The accept/reject decision is performed by comparing the output score to a threshold. If $\text{Score} < T$, then reject the
claim, otherwise, accept the claim.

![Diagram](image)

**Fig. 4. Structure of a polynomial classifier.**

We perform the speaker verification test on the XM2VTS database [1]. This database includes four recordings of 295 subjects taken at one month intervals. (However we were able to use only 261 of the 295 speakers because of corrupted audio or video sequences.) Each sequence is approximately 5 seconds long and contains the subject speaking the sentence "Joe took father’s green shoe bench out." The database covers a large population from various ethnic origins and with various appearances. The same person might attend the four sessions with a different appearance, including hairstyles, with/without glasses, with/without beard, with/without lipstick.

To evaluate the performance of the person authentication systems on the XM2VTS database, we adopt the protocol defined in [8]. We choose configuration II due to the audio-visual data we are using. The protocol partitions the database into a training set, evaluation set and test set. In configuration II, data from the first two sessions are used to train the clients' models. The system threshold is set from evaluation data composed of the third session of the clients' data and all four sessions of the evaluation imposters' data. The final performance test uses data from the fourth session of the clients and from all four sessions of the test imposters. For this description, each subject appears only in one set. This ensures realistic evaluation of the impostor claims whose identity is unknown to the system.

The verification performance is characterized by two error rates computed during tests: the false acceptance rate (FAR) and the false rejection rate (FRR). The pooled equal error rate (EER) threshold is determined from the evaluation set and used against the test population to determine the system performance. Both FAR and FRR are reported for this EER operating point. The test results for a visual only speaker verification system are shown below:

<table>
<thead>
<tr>
<th>Features</th>
<th>Poly. Order</th>
<th>FRR %</th>
<th>FAR %</th>
</tr>
</thead>
<tbody>
<tr>
<td>all (8)</td>
<td>2</td>
<td>8.8</td>
<td>9.7</td>
</tr>
<tr>
<td>all + Δ (16)</td>
<td>2</td>
<td>6.1</td>
<td>9.3</td>
</tr>
<tr>
<td>all + Δ (16)</td>
<td>3</td>
<td>5.0</td>
<td>9.0</td>
</tr>
<tr>
<td>all + Δ (16)</td>
<td>3</td>
<td>4.4</td>
<td>8.2</td>
</tr>
<tr>
<td>all + time (9)</td>
<td>2</td>
<td>8.3</td>
<td>9.2</td>
</tr>
<tr>
<td>all + time (9)</td>
<td>3</td>
<td>4.8</td>
<td>8.5</td>
</tr>
</tbody>
</table>

In our experiment, polynomial orders two and three are used. The visual features included are the eight parameters derived in the last section. Extra features are the corresponding delta features and the normalized time index $i/M$, where $i$ is the current frame index, and $M$ is the total number of frames. As can be seen, by incorporating extra features, a lower error rate is achieved. At the same time, increasing the polynomial order also contributes to improved verification results.

In the audio modality, each feature vector is composed of 12 cepstral coefficients and one normalized time index [3]. A third-order polynomial classifier is used. To fuse the two modalities, we use late integration. It is necessary that the classifier outputs represent class probabilities. We use an optimum Bayes approach. First we calculate the conditional probability $p(x_1, \ldots, x_N | \omega_j)$, which we abbreviate as $p(x_i^N | \omega_j)$, where $\omega_j$ is class $j$. By assuming independence, we obtain

$$p(x_i^N | \omega_j) = \prod_{k=1}^N p(x_k | \omega_j).$$

Using the relation

$$p(x_k | \omega_j) = \frac{p(\omega_j | x_k)p(x_k)}{p(\omega_j)}$$

and (2), we obtain the discriminant function

$$d'(x_i^N) = \prod_{k=1}^N \frac{p(\omega_j | x_k)}{p(\omega_j)}.$$

We have discarded the numerator term $\prod_{k=1}^N p(x_k)$ because it is independent of $\omega_j$. Two simplifications are performed. First we consider the logarithm of the discriminant function,

$$\log(d'(x_i^N)) = \sum_{k=1}^N \log \frac{p(\omega_j | x_k)}{p(\omega_j)}.$$

Using a Taylor series, a linear approximation of $\log(x)$ around $x = 1$ is $x - 1$. Thus, we can approximate $\log(d'(x_i^N))$ as

$$d(x_i^N) = \sum_{k=1}^N \frac{p(\omega_j | x_k)}{p(\omega_j)},$$

where we have dropped the $-1$, since a constant offset will be eliminated in a log likelihood ratio function. Thus, the scoring method is equivalent to computing a log probability. We can combine the classifier output from the audio and visual modalities by averaging the class scores, $s = \alpha s_A + (1 - \alpha)s_V$. For the following experiments, the audio and visual modalities are weighted equally (i.e., $\alpha = 0.5$).
Fig. 5. Performance of audio-visual speaker verification in noisy conditions.

The performance of the bimodal speaker verification system is shown in Fig. 5. As can be seen, the performance of the audio modality degrades as the noise level increases. This figure shows the False Rejection Rate (FRR) and the False Acceptance Rate (FAR) for each modality independently, as well as for the fused system. Both curves are of interest since the threshold is determined with an evaluation population separated from the test population. As illustrated in the figure, the audio-visual fusion is shown to outperform both modalities at high signal-to-noise ratios.

4. SUMMARY AND CONCLUSIONS

In this paper we have demonstrated that visual speech features can be successfully used as an aid for personal authentication. Experiments based on the XM2VTS database [1] with 261 speakers achieve an FRR of 4.4% and an FAR of 8.2% with polynomial order three and suggest that visual information is highly effective for reducing the false acceptance and false rejection rates in such tasks. The combined audio-visual system is shown to outperform both modalities at high signal-to-noise ratios. However, error rates over the range of signal-to-noise ratios (SNR) indicate that a dynamic fusion strategy, e.g., adjusting the weighting of the modalities as SNR degrades, may improve the overall system performance.

Acknowledgments

The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documents [1, 9].

5. REFERENCES

[1] “URL: ee.surrey.ac.uk/research/vssp/xm2vtsdb,”.