ISCA Archive
http://www.isca-speech.org/archive

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

Using Lip Features for Multimodal Speaker Verification

Xiaozheng Zhang[†] *and Charles C. Broun*[‡]

[†]The Georgia Institute of Technology, Atlanta, Georgia , USA
[‡]Motorola Human Interface Lab, Tempe, Arizona, USA

Abstract

With the prevalence of the information age, privacy and personalization are forefront in today's society. As such, biometrics is viewed as an essential component of current and evolving technological systems. Consumers demand unobtrusive and non-invasive approaches. In our previous work, we have demonstrated a speaker verification system that meets these criteria. However, there are additional constraints for fielded systems. The required recognition transactions are often performed in adverse environments and across diverse populations, necessitating robust solutions.

We propose a multimodal approach that builds on our current state-of-the-art speaker verification technology. In order to maintain the transparent nature of the speech interface, we focus on optical sensing technology to provide the additional modality–giving us an audio-visual person recognition system. For the audio domain, we use our existing speaker verification system. For the visual domain, we focus on lip motion.

The visual processing method makes use of both color and edge information, combined within a Markov random field (MRF) framework, to localize the lips. Geometric features are extracted and input to a polynomial classifier for the person recognition process. A late integration approach, based on a probabilistic model, is employed to combine the two modalities. The system is tested on the XM2VTS database combined with additive white Gaussian noise (AWGN) (in the audio domain) over a range of signal-to-noise ratios.

1. Introduction

There are two significant problem areas in current generation speaker verification systems. The first is the difficulty in acquiring clean audio signals without encumbering the user with a head-mounted close-talking microphone. Second, unimodal biometric systems do not work with a significant percentage of the population. To combat these issues, multimodal techniques are being investigated to improve system robustness to environmental conditions, as well as improve overall accuracy across the population.

The use of multiple modalities to perform person recognition is not a new concept. However, work in multimodal automatic person recognition has recently gained a lot of momentum with the increasing processing power and storage available today. Two well-researched domains in person recognition are speaker and face recognition. However, face recognition does not provide the same dynamics as speech. In addition, the lip dynamics can aid speech recognition to provide liveness testing. Thus, lip tracking for person identification is gaining interest. A lip-tracking system must locate the lips in the video sequence and then perform the feature extraction. Subsequently, for a multimodal system, the two domains must be integrated, or fused.

There are several methods for lip localization [1]. *Deformable templates* use geometric shapes that are allowed to deform and move in order to minimize an energy function. *Template matching* traditionally employs correlation to locate facial features. *Knowledge based approaches*, seen in earlier systems, use pyramid images to detect faces, and employed edge detection and subjective rules to find facial features. *Visual motion analysis* techniques rely on the use of difference images after filtering and thresholding, and it is implicitly reliant upon intensity information.

There are also several types of features that can be employed for lip tracking [1]. With an *image-based approach*, the image containing the mouth is used directly. With *visual motion analysis* (e.g., optical flow), it is believed that the visual motion during speech production contains relevant speech information. Approaches that rely on *geometric features* assume relevant speech information is contained within certain measures of the mouth geometry (e.g., height and width of the mouth opening). A *model-based approach* uses parameterized models of the speech articulators.

The various methods of combining the modalities are as follows [2]. With the *direct identification* model, the classifier uses the multimodal data directly. With *separate identification*, or late integration, there is a separate classifier for each modality. The resulting outputs of each are fused. There are two forms of early integration. With *dominant recoding*, fusion of each modality precedes classification. With *motor recoding* each modality's inputs are projected into an amodal common space related to the characteristics of speech gestures. Fusion then occurs within this common domain.

The organization of the paper is as follows. In Section 2, feature extraction in the visual domain is discussed. The polynomial classifier and late integration approach is described in Section 3. The experiments with the XM2VTS database, along with the system performance, are presented in Section 4. Finally, Section 5 contains the conclusions.

2. Visual feature extraction

In visual processing, the basic visual features of the lip are measured and fed into a classifier. The main problem is how we extract these features from the video sequence.

We first locate the lip region using hue color due to the following considerations [3]: i) hue color reduces intensity dependency, ii) hue color for the lip region is fairly uniform, iii) hue has high discriminative power, and iv) hue is relatively constant under varying conditions and different human skin

(a) (b) (c) (d)

Figure 1: General case of using color information to locate the lip region. (a): original image, (b) hue image, (c) resulting binary image after thresholding, (d): detected lip region.

color. In order to use hue, we require that saturation must exceed a certain preset value to eliminate the noise in the hue image [3]. To segment the lip, we use the following H and S constraints:

$$BW(x,y) = \begin{cases} 1, & H(x,y) > H_0, S(x,y) > S_0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where H_0=0.8, S_0=0.25 for $H/S \in [0,1]$. A typical case for using hue color to locate the lip is shown in *Figure 2*. Given a color image of a talking person in *Figure 2*(a), we first derive the hue image from color space conversion (*Figure 2*(b)). Using equation (1), the binary image is acquired as shown in *Figure 2*(c). Then the lip region can be easily detected [3], which is shown as a white bounding box in *Figure 2*(d).

In most of the cases this method works fine. However, complications can occur when a person has a very red face, or he is wearing a red shirt, scarf or tie. To eliminate distractions of other lip-colored blobs, we employ several strategies. First, we observe that mouth region is characterized by high edge content; we therefore require that the average gradient of the candidate region exceed a certain value. We then increase the saturation constraint and geometric constraint in order to eliminate the lip-colored area. A special case of a person having a very red face is shown in Figure 2(a). The hue image in Figure 2(b) shows a large red blob with lip and its surrounding area. By measuring the geometry of the blob combined with the gradient value, we conclude that the detected area includes a lip-neighboring region. To eliminate the non-lip area, we further increase the saturation threshold since the saturation of the face area is lower than that of the lip. The binary image after thresholding is shown in Figure 2(c). The lip region can subsequently be derived as shown in Figure 2(d).

Note that the procedures described above need only to be done once on the first image of a sequence. The lip region of the following frames is estimated from the segmented lip of the previous ones.

Besides color information, edges characterize object boundaries and provide additional useful information for describing the lip. The definition of hue color used here and the edge detection [4] are described in [3]. In order to extract lip features from the image of a lip, we use a Markov random field (MRF) framework to segment the lip. It has been shown to be suitable for the problem of spatial statistical modeling [5].

In the MRF, the state of a site is dependent only upon the state of its neighbors. It can be modeled by a Gibbs distribution.

$$\begin{aligned} P(\mathbf{x}) &= \frac{1}{Z}\exp\left[-\frac{1}{T}U(\mathbf{x})\right] \\ U(\mathbf{x}) &= \sum_{c\in C} V_c(\mathbf{x}) \\ V_c(i,j) &= \begin{cases} -\beta & \text{if } x_i = x_j \\ \beta & \text{otherwise} \end{cases} \end{aligned} \quad (2)$$

The normalizing constant Z is called the partition function. T is the temperature constant, and $U(\mathbf{x})$, the Gibbs potential, is the sum of potentials of each clique. C is the set of all cliques. $Vc(\mathbf{x})$ encodes *a priori* knowledge about spatial dependence of labels at neighboring sites. Spatial connectivity of the

(a) (b) (c) (d)

Figure 2: Special case of using hue color information to locate the lip region. (a): Original image, (b): Resulting binary image after thresholding, (c): binary image after increased saturation constraint, (d): detected lip region.

Figure 3: Lip segmentation.

segmentation is imposed by assigning the clique potential $Vc(\mathbf{x})$ as above, where β is a positive number. This potential assignment implies higher probability for pixel pairs with identical labels and lower probability for pairs with different labels, thus encouraging spatially connected regions.

We formulate the lip segmentation problem as a site-labeling problem. Each site is assigned to a label x_i from the set {lip, non-lip}, and b_i from {edge, non-edge}. The maximum *a posterior* (MAP) criterion is used to formulate what the best labeling should be, with

$$p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x})$$
$$p(\mathbf{y} \mid \mathbf{x}) \propto \exp\left[-\sum_i \frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^{\ 2}} \right] \tag{3}$$
$$p(\mathbf{x}) = \exp\left[-\frac{1}{T} \sum_{c \in C} V_c(\mathbf{x}) \right]$$

In equation (3), $p(\mathbf{y}|\mathbf{x})$ denotes the conditional pdf of the image given the segmentation, and $p(\mathbf{x})$ is the *a priori* pdf, modeled by a Gibbs distribution. The image is modeled by a uniform mean [6], where x_i is the label of site i, y_i is the observed image data, μ_{x_i} and σ_{x_i} are the mean and variance of all pixels in the image with the region label x_i.

The maximization of the *a posterior* probability in equation (3) is equivalent to the minimization of the energy function in equation (4), which consists of two parts: one associated with the difference between the predicted image and the actual observed data, the other describing the interaction potential between neighbors. To minimize the energy function, we use the Highest Confidence First (HCF) algorithm [7]. HCF is a deterministic iterative algorithm finds the lowest energy. The main ingredient of HCF is the order in which sites are visited. Instead of updating the pixels sequentially, as in other methods, HCF requires that the site that is visited next be the one that generates the largest energy reduction.

$$U(\mathbf{y} \mid \mathbf{x}) = \lambda \sum_i \frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^{\ 2}} + \sum_{c \in C} V_c(\mathbf{x}) \tag{4}$$

Figure 3 illustrates the results of segmentation. The geometric lip features are derived from the segmented image. Typical features are the height and width of the inner and outer lip, the height and width of the mouth opening, and the visibility of teeth and tongue (*Figure 4*).

Figure 4: Lip feature extraction.

3. Classification

3.1 Polynomial Classifier

Polynomial classifiers have been used for pattern classification for many years [8][9], and have excellent properties as classifiers. Because of the Weierstrass approximation theorem, polynomials are universal approximators for the Bayes classifier [8].

The basic structure of our classifier is shown in *Figure 5*. The feature vectors, $\mathbf{x}_1 \ldots \mathbf{x}_M$, produced from feature extraction, are introduced to the system. A discriminant function [9] is applied to each feature vector, $\mathbf{x}_k$, using a speaker model, $\mathbf{w}$, producing a scalar output, $d(\mathbf{x}_k, \mathbf{w})$. The final score, s, for the speaker model is then computed.

$$s = \frac{1}{M} \sum_{k=1}^{M} d(\mathbf{x}_k, \mathbf{w}) \tag{5}$$

Comparing the output score to a threshold, T, performs the accept/reject decision for the system. If $s < T$, then reject the claim; otherwise accept the claim.

Our pattern classifier uses a polynomial discriminant function,

$$d(\mathbf{x}, \mathbf{w}) = \mathbf{w}^t \mathbf{p}(\mathbf{x}). \tag{6}$$

The discriminant function is composed of two parts. The first part, $\mathbf{w}$, is the speaker model. The second part, $\mathbf{p}(\mathbf{x})$, is a polynomial basis vector constructed from input feature vector $\mathbf{x}$. This basis vector is the monomial terms up to degree K of the input features. For example, for a two dimensional feature vector, $\mathbf{x} = [x_1\ x_2]^t$, and $K = 2$, we have

$$\mathbf{p}(\mathbf{x}) = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_1 x_2 & x_2^2 \end{bmatrix}^t. \tag{7}$$

Thus, the discriminant function output is a linear combination of the polynomial basis elements. Since $\mathbf{w}$ does not depend on the frame index, scoring can be simplified as follows:

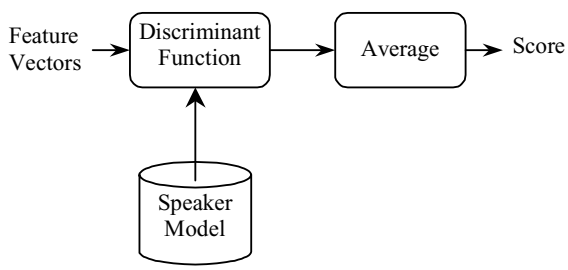

Figure 5: Classifier structure.

Figure 6: Examples of feature extraction results from the XM2VTS database.

$$s = \mathbf{w}^t \frac{1}{M} \sum_{k=1}^{M} \mathbf{p}(\mathbf{x}_k) = \mathbf{w}^t \overline{\mathbf{p}} \,. \tag{8}$$

Only a single vector is required to represent the input speech, and a single verification transaction equates to computing an inner product. The number of floating point operations is $2N_{\text{model}} - 1$, where N_{model} is the length of $\mathbf{w}$. Thus for 12 features and a 3rd order ($K = 3$) polynomial expansion, $\mathbf{w}$ is of length 455, resulting in only 909 flops per transaction, and a model size of 1820 bytes for a floating point representation.

An efficient method for training is given in [10].

3.2 Multimodal Fusion

A late integration approach is used to fuse the audio and visual modalities. Thus, it is necessary that the classifier outputs represent class probabilities. We use an optimum Bayes approach; we first calculate $p(\mathbf{x}_1,\ldots,\mathbf{x}_M \mid \omega_j)$. We abbreviate this as $p(\mathbf{x}_1^M \mid \omega_j)$.

By assuming independence, we obtain

$$p(\mathbf{x}_1^M \mid \omega_j) = \prod_{k=1}^{M} p(\mathbf{x}_k \mid \omega_j) \,. \tag{9}$$

Using the relation

$$p(\mathbf{x}_k \mid \omega_j) = \frac{p(\omega_j \mid \mathbf{x}_k) p(\mathbf{x}_k)}{p(\omega_j)} \tag{10}$$

and (9), we obtain the discriminant function

$$d'(\mathbf{x}_1^M) = \prod_{k=1}^{M} \frac{p(\omega_j \mid \mathbf{x}_k)}{p(\omega_j)} \,. \tag{11}$$

We have discarded the numerator term $\prod_{k=1}^{M} p(\mathbf{x}_k)$ because it is independent of ω_j.

Two simplifications are now performed. First, we consider the logarithm of the discriminant function,

$$\log(d'(\mathbf{x}_1^M)) = \sum_{k=1}^{M} \log\left(\frac{p(\omega_j \mid \mathbf{x}_k)}{p(\omega_j)} \right) \tag{12}$$

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

$$d(\mathbf{x}_1^M) = \sum_{k=1}^{M} \left(\frac{p(\omega_j \mid \mathbf{x}_k)}{p(\omega_j)} \right), \tag{13}$$

where we have dropped the -1, since a constant offset will be eliminated in a log likelihood ratio function. (See [11] for additional details.)

We now see that our scoring method is equivalent to computing a log probability. Thus, we can combine the classifier output from the audio and visual modalities by averaging the class scores.

4. Experiments

4.1 XM2VTS Database

The XM2VTS database [12] is a large multimodal database created for automatic person recognition. In total, the database is composed of audio-only speech recordings, audio-visual speech recordings, and frontal and profile views (for face and mug shot authentication). For our task, only the audio-visual speech portion of the database is of interest. There are 295 participants who each speak three sentences two times over four different sessions. Unfortunately, the distribution set of the audio-visual recording only contains the third sentence and only the first repetition from each of the four sessions.

Our final system is only able to use 261 of the 295 speakers due to either incorrectly labeled data, or corrupt audio or video sequences. The spoken phrase is "*Joe took fathers green shoe bench out*." The audio sequences are recorded at a sampling rate of 32 kHz with a resolution of 16 bits. The video is captured at a color sampling resolution of 4:2:0, and it is compressed at the fixed ratio of 5:1 in the DV format.

The evaluation protocol for the XM2VTS database is given in [13]. There are two preferred configurations for training the system, determining parameters, and testing the performance. Configuration I provides for good *expert* training, but poor *fusion* training. Configuration II provides for good *fusion* training at the expense of poor *expert* training. Since only the first sentence of each session is available on the audio-visual distribution of the database, we only consider Configuration II, and we are limited to only half of the data. Thus, training of the *expert* classifiers is expected to be difficult. In this configuration, data from the first two sessions is used to train

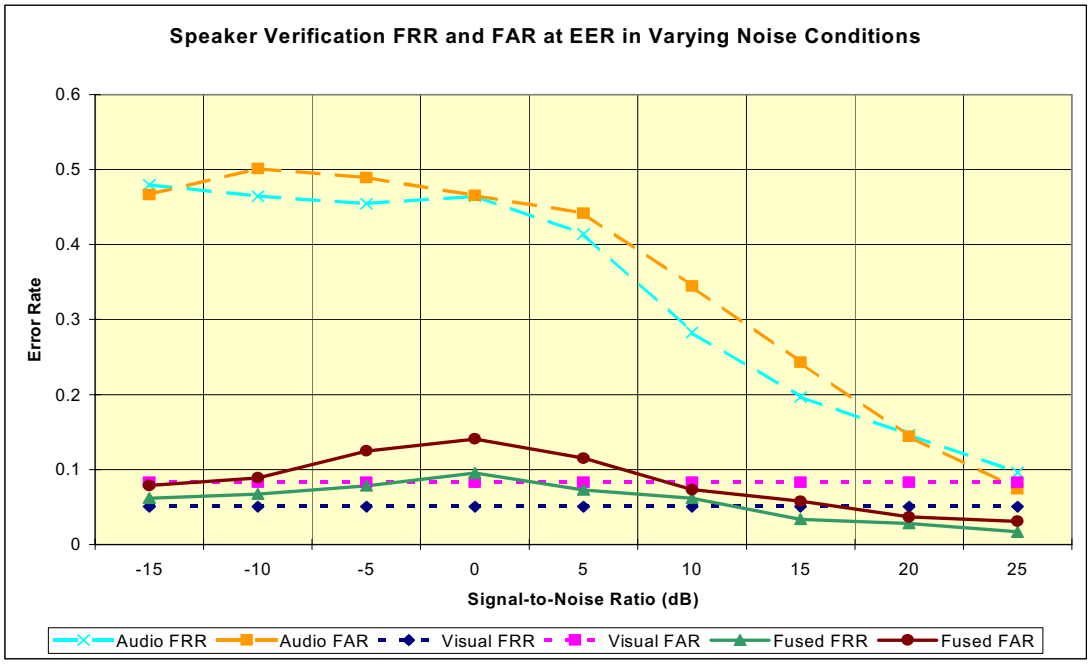

Figure 7: Performance of audio-visual speaker verification in noisy conditions.

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 impostors' data. The final performance test uses data from the fourth session of the clients and from all four sessions of the test impostors. Our experiments use the same client, evaluation impostor, and test impostor populations as defined in [13].

4.2 Results

This database provides more than a thousand video sequences, which cover a large amount of population among male/female, young/old, and with various skin color. In addition, the same person might attend four sessions with a different appearance, including hairstyles, with/without glasses, with/without beard, and with/without lipstick. Our scheme proves to be robust to all these variations. Examples of visual feature extraction results from the database are shown in *Figure 6*.

We use two classifiers, one per modality, each trained as a 3rd order system [10]. For the audio modality, each feature vector is composed of 12 cepstral coefficients and one normalized-time index, for a total vector length of 13. The visual feature vectors are of length 9, and consist of inner and outer lip height and width, mouth opening height and width, presence of teeth and tongue, and a normalized-time index. The normalized-time index makes implicit use of the knowledge that the verification phrase is constant. It is computed as i/M, where i is the current frame index, and M is the total number of frames. For text-prompted or text-independent applications, this feature is not used.

The pooled equal error rate (EER) threshold is determined from the evaluation set and used against the test population to determine the system performance. In addition, the audio modality is subjected to additive white gaussian noise (AWGN) at various signal-to-noise ratios (SNR). As is illustrated in *Figure 7*, the performance of the audio modality degrades as the relative noise level increases. This figure shows the False Reject Rate (FRR) and the False Accept 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 separate from the test population. It is reasonable that the FRR and FAR curves for the test population do not follow each other exactly, but that they are close.

As indicated by the error rates of the different systems (single modal and fused), the performance of the fused system is degraded when the performance of one of the modalities is very poor. This type of behavior in integrated systems is seen quite often, leading to the conclusion that multimodal solutions should incorporate confidence measures for each of the modalities to control the integration. A simple scheme might incorporate a weighting function, which has not been examined for the system presented in this paper.

5. Conclusions

We have demonstrated an audio-visual multimodal system for person recognition. The audio-processing portion builds from our previous work in speaker recognition. The visual domain uses recent developments in lip feature extraction techniques using color information. The resulting performance of the multimodal system is shown to perform well in all conditions.

6. Acknowledgements

The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in

[12]. http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb

7. References

[1] J. Luettin, *Visual Speech and Speaker Recognition*, PhD thesis, University of Sheffield, 1997.

[2] P. Teissier, J. Robert-Ribes, J.-L. Schwartz, and A. Guerin-Dugue, "Comparing Models for Audiovisual Fusion in a Noisy-Vowel Recognition Task," *IEEE Transactions on Speech and Audio Processing*, vol. 7, issue 6, pp. 629-642, November 1999.

[3] X. Zhang, R. M. Mersereau, "Lip Feature Extraction Towards an Automatic Speechreading System," *ICIP*, 2000.

[4] J. F. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679-698, 1986.

[5] S.Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol PAMI-6, pp721-741, Nov. 1984

[6] H.Derin, H.Elliott, "Modeling and segmentation of noisy and textured images using Gibbs random field," ," *IEEE Trans. Pattern Anal. Machine Intell*. 9, (1987)39-55

[7] P. Chou, C. Brown, and R. Raman, "A Confidence-Based Approach to the Labeling Problem", in *Proceedings of the IEEE Workshop on Computer Vision*, pp. 51-56, Miami Beach, Florida, 1987.

[8] J. Schürmann, *Pattern Classification*. John Wiley and Sons, Inc., 1996.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[10] W. M. Campbell and K. T. Assaleh, "Polynomial classifier techniques for speaker verification," ICASSP, 1999.

[11] W. M. Campbell and C. C. Broun, "A Computationally Scalable Speaker Recognition System," *EUSIPCO*, 2000

[12] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," in *Proceedings 2nd Conference on Audio and Video-Based Biometric Personal Verification* (AVBPA99), 1999.

[13] J. Luettin and G. Maitre, "Evaluation Protocol for the XM2VTS Database," IDIAP-Com 98-05, October 1998.