Multimodal fusion of polynomial classifiers for automatic person recognition

Charles C. Broun[a] and Xiaozheng Zhang[b]

[a]Motorola Labs – Human Interface Lab, Phoenix, Arizona
[b]The Georgia Institute of Technology, Atlanta, Georgia

ABSTRACT

With the prevalence of the information age, privacy and personalization are forefront in today's society. As such, biometrics are viewed as essential components 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.

There are two significant problem areas in current generation speaker verification systems. The first is the difficulty in acquiring clean audio signals (in all environments) 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.

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. This is chosen, rather than static face or iris recognition, because it provides dynamic information about the individual. In addition, the lip dynamics can aid speech recognition to provide liveness testing.

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 AWGN (in the audio domain) over a range of signal-to-noise ratios.

Keywords: multimodal fusion, polynomial classifier, active shape model, Markov random field, lip tracking, speech recognition, speaker verification

1. INTRODUCTION

Biometrics is an emerging technology with outstanding potential in many modern authentication systems. Biometrics simplifies the interface to the human user by eliminating the need for passwords and PINs. They are cumbersome at best because of various practices currently in use. By their nature, passwords and PINs are difficult to remember, must be changed frequently, and are subject to “cracking”. Biometrics solves these problems by the use of various distinguishing characteristics of individuals. Authentication methods commonly used are voice, fingerprints, hand geometry, iris structure, facial characteristics, etc. Access is controlled through a verification process that determines whether a claimant’s characteristics match those of the claimed identity.

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. 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 both the audio and visual domains is discussed. The polynomial classifier and late integration approach in 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. FEATURE EXTRACTION

2.1. Audio Processing

Feature extraction for the audio modality is performed as shown in Figure 1. Discrete-time input speech, $x[n]$, is processed using 12th order LP (linear prediction) analysis every 10 ms using a frame size of 30 ms. The speech is pre-emphasized using the filter $H(z) = 1 - 0.97\, z^{-1}$. A Hamming window is applied to each frame before LP analysis. Twelve cepstral coefficients, c_n, are generated from the LP coefficients, a_n, using the following transform equation.

$$c_0 = 0, c_n = a_n + \sum_{k=0}^{n-1} \frac{k}{n} c_k a_{n-k} \tag{1}$$

Some robustness to noise and varying channel conditions is mitigated through cepstral mean subtraction (CMS).

The input speech utterance is split up in to speech and non-speech segments through a process called endpointing. A simple short-time energy scheme is used. For each input frame, the short-time energy

$$E_i = \sum_{i=0}^{N_s-1} \left| x[n] \right|^2 \tag{2}$$

is calculated, where N_s is the number of samples in the frame. The energy is then split into two distributions representing speech and non-speech signals. The average energy level is calculated for each of these distributions resulting in E_{speech} and $E_{\text{nonspeech}}$. The speech – non-speech threshold is then set at

$$\alpha \overline{E}_{\text{nonspeech}} + (1-\alpha)\overline{E}_{\text{speech}}\, , \tag{3}$$

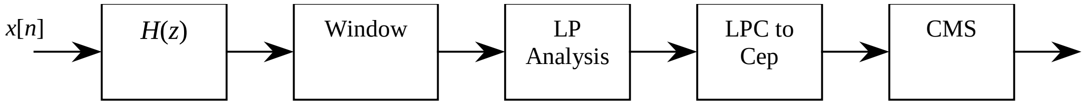

Figure 1: Audio feature extraction.

where α is typically 0.8. Every frame above this threshold is labeled as speech and subsequently processed. In addition to the cepstral coefficients, a normalized-time feature, i/N_{frames}, is appended for a total of 13 features.

2.2. Visual Processing

The main difficulty of lip feature extraction lies in the accuracy and reliability of the system. Recent research has shown that color is a powerful tool with regard to those two aspects. Unlike the gray-level approach, color image analysis increases the efficiency and robustness of locating the lips, and easily adapts to detect beards, teeth and tongue.

We start by examining various color spaces. RGB is the most widely used among many existing color spaces. However the triple [R,G,B] represents not only color but also brightness, which hinders the effectiveness of color in detection. Several studies have shown that even though different people have different colors in appearance, the major difference lies in intensity rather than color itself. To separate the chromatic and luminance components, various transformed color spaces can be employed, such as the normalized RGB space (we denote it as rgb in the following), YCbCr, and HSV [3].

To analyze the statistics of each color model, we build histograms of color components. We construct histograms for the entire image and for the extracted lip region bounded within the estimated boundary. From experiments on various video sequences taken under different test conditions and for different test subjects we have the following observations [4]: i) Color components (r,g,b), (Cb,Cr) and (H) exhibit peaks in their histograms. This indicates that the feature distribution of the lip region is narrow and implies that the color for the lip region is fairly uniform. ii) The color histogram of (r,g,b) and (Cb, Cr) of the lip region more or less overlaps with that of the whole image, while the hue component has the least similarity between the entire image and lip region only. This shows that hue has high discriminative power. iii) The distribution of (r,g,b) and (Cb,Cr) vary for different test subjects, while hue is relatively constant under varying conditions, such as lighting conditions, and for different talkers. We therefore conclude that hue is an appropriate model for our application. In order to use hue, we require that S must exceed a certain preset value. For segmenting the lip, we use the following H and S constraints:

$$BW(x,y) = \begin{cases} 1, & H(x,y) > H_0, S(x,y) > S_0 \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where H_0=0.8, S_0=0.25 for $H/S \in$ [0,1]. The accuracy of these values is not critical, and they prove to generalize well across talkers. A typical binary image is shown in Figure 2. A special case is shown in Figure 3. In this case, the use of increased saturation, geometric, and gradient constraints are required in order to eliminate distraction from other red areas. The lip region is extracted using morphological operations on the binary image [4].

Edge information is extracted using a Canny edge detector [5]. To combine the edge information with hue color information, we use the machinery of the Markov random field (MRF). The reason is twofold. First, extraction of lip features recovers the true image from the noisy observed image. It is, therefore, an inverse problem with many possible solutions and is ill posed [6]. This problem can be solved by the use of regularization methods employed in the MRF framework. Second, the MRF formulation allows us to embed many features of interest by simply adding appropriate terms in the energy function; therefore it provides an easy tool for fusing multiple low-level vision modules.

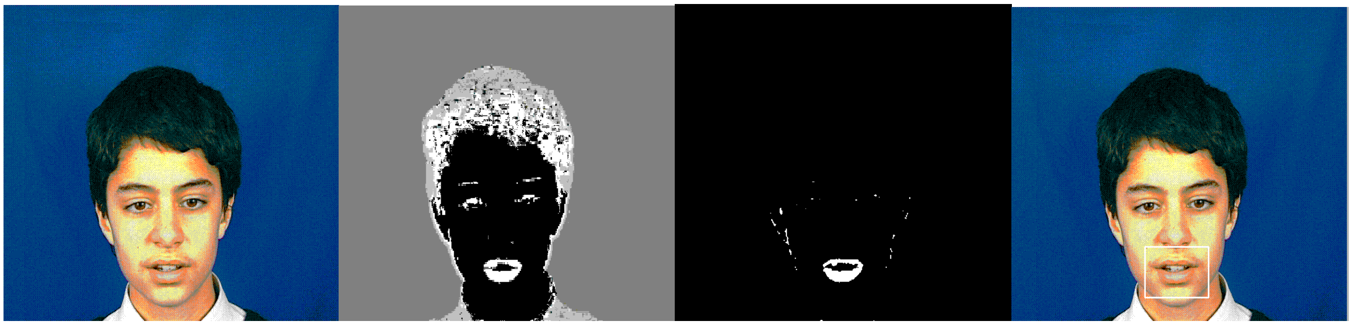

Figure 2: General case of using hue color information to locate the lip region.

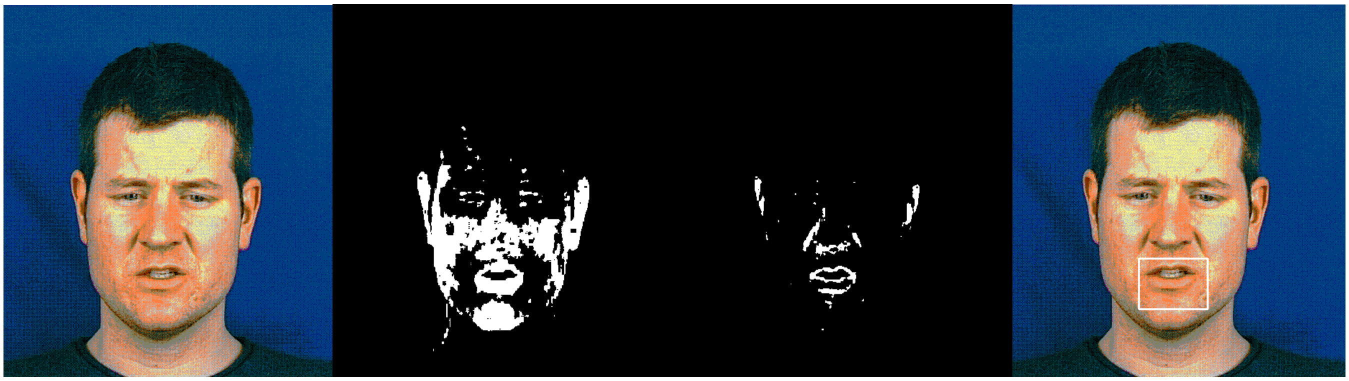

Figure 3: Special case of using hue color information to locate the lip region.

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.

$$\begin{aligned} P(\mathbf{x}) &= \frac{1}{Z}\exp\left[-\frac{1}{T}U(\mathbf{x})\right] \\ U(\mathbf{x}) &= \sum_{c\in C} V_c(\mathbf{x}) \\ V_c(i,j) &= \begin{cases} -\beta & \text{if } x_i = x_j \\ \beta & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

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.

$$\begin{aligned} p(\mathbf{x}\,|\,\mathbf{y}) &\propto p(\mathbf{y}\,|\,\mathbf{x})\,p(\mathbf{x}) \\ p(\mathbf{y}\,|\,\mathbf{x}) &\propto \exp\left[-\sum_i \frac{(y_i-\mu_{x_i})^2}{2\sigma_{x_i}{}^2}\right] \\ p(\mathbf{x}) &= \exp\left[-\frac{1}{T}\sum_{c\in C} V_c(\mathbf{x})\right] \end{aligned} \tag{6}$$

The HCF algorithm [7] allows one to reduce the estimation problem to the minimization of an energy function.

$$U(\mathbf{y}\,|\,\mathbf{x}) = \lambda\sum_i \frac{(y_i-\mu_{x_i})^2}{2\sigma_{x_i}{}^2} + \underset{c\in C}{} V_c(\mathbf{x}) \tag{7}$$

Figure 4 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 5).

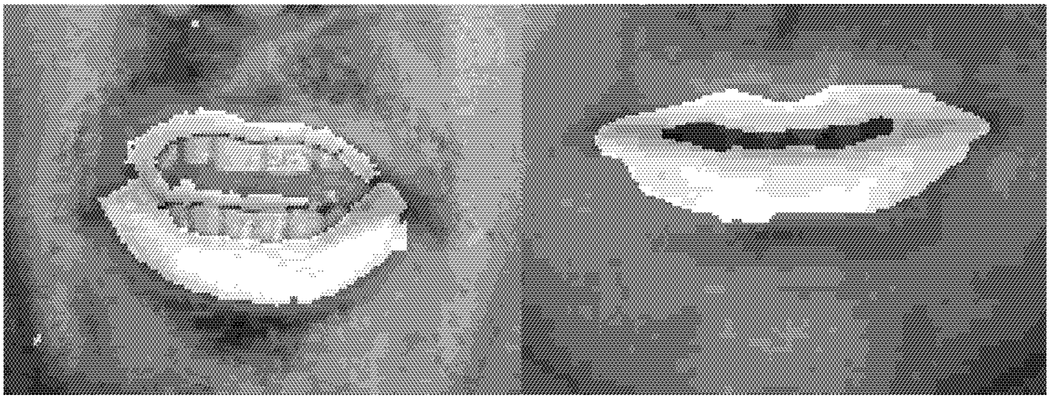

Figure 4: Lip segmentation.

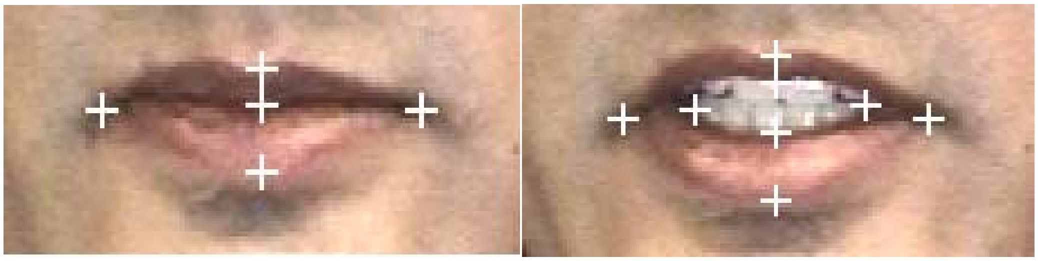

Figure 5: Lip feature extraction.

3. CLASSIFICATION

3.1. Polynomial Classifier

Many classification methods are currently being applied to the problem of speaker verification [8]. Traditionally, statistical methods are used to model the speaker's speech data from the feature extraction phase. Two of the most popular approaches are the Hidden Markov Model (HMM) [9] and the Gaussian Mixture Model (GMM) [10]. More recently, discriminative classification techniques employing artificial neural networks, such as neural tree networks (NTN) [11], have been applied to the problem. In order to provide the best performance for speaker verification systems, the latter methods include out-of-class data in the training phase. This technique produces robust speaker models by maximizing the separation between classes.

Polynomial classifiers have been used for pattern classification for many years [12][13], and have excellent properties as classifiers. Because of the Weierstrass approximation theorem, polynomials are universal approximators for the Bayes classifier [12].

The basic structure of our classifier is shown in Figure 6. The feature vectors, $\mathbf{x}_1 \ldots \mathbf{x}_M$, produced from feature extraction, are introduced to the system. A discriminant function [13] 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 for the speaker model is then computed

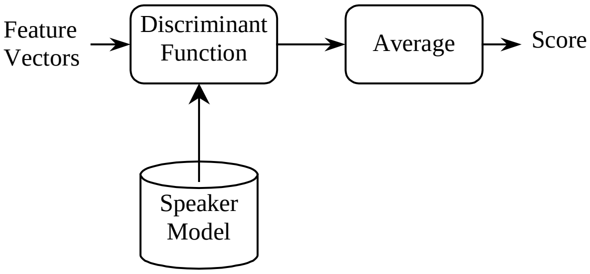

Figure 6: Classifier structure.

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

Comparing the output score to a threshold 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 [13]

$$d(\mathbf{x}, \mathbf{w}) = \mathbf{w}^t p(\mathbf{x}) \tag{9}$$

The discriminant function is composed of two parts. The first part, $\mathbf{w}$, is the speaker model. The second part, $p(\mathbf{x})$, 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

$$p(\mathbf{x}) = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_1x_2 & x_2^2 \end{bmatrix}^t \tag{10}$$

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:

$$s = \mathbf{w}^t \frac{1}{M}\sum_{k=1}^{M} p(\mathbf{x}_k) = \mathbf{w}^t \overline{p} \tag{11}$$

Thus, only a single vector represents the input speech and a single transaction equates to computing an inner product. The number of floating point operations (FLOPS) is

$$2N_{\text{model}} - 1, \tag{12}$$

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 [14].

3.2. Multimodal Fusion

A late integration approach is used to fuse the audio and visual modalities. It is necessary that the classifier outputs represent class probabilities. As demonstrated in [15], the polynomial classifier discriminant function can be expressed as

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

where ω_j is class j. Two simplifications are performed. First, we consider the logarithm of the discriminant function,

$$\log(d'(\mathbf{x}_1^M)) = \sum_{k=1}^{M} \log\left(\frac{p(\omega_j \mid \mathbf{x}_k)}{p(\omega_j)}\right) \tag{14}$$

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

$$d(\mathbf{x}_1^M) = \sum_{k=1}^{M} \left(\frac{p(\omega_j \mid \mathbf{x}_k)}{p(\omega_j)}\right), \tag{15}$$

where we have dropped the –1 since a constant offset will be eliminated in a log likelihood ratio function. We now see that our scoring method is equivalent to computing a log probability. Thus, combining the classifier output from the audio and visual modalities is a simple matter of adding the class scores.

4. EXPERIMENTS

4.1. XM2VTS Database

The XM2VTS database [16] 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 mugshot authentication). For our interests, only the audio-visual speech portion of the database is of interest. There are 295 participants who each spoke three sentences two times each over four different sessions. (For authentication problems, it is critical that a database is multi-session.) 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*." However, the recognition system does not make use of this *a priori* knowledge. 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 [17]. 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, on the other hand, 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 are forced to 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 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 [17].

4.2. Results

Two classifiers are designed, one for the audio modality and one for the visual modality. Both classifiers are trained as a 3rd order system [14]. 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 pooled equal error rate (EER) threshold is determined from the evaluation set and used against the test population to determine the system performance. Both the false reject rate (FRR) and the false accept rate (FAR) are reported for this EER operating point. 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. In addition, the audio-visual fusion is shown to outperform both modalities at high signal-to-noise ratios. This type of performance is typical for audio-visual speech recognition systems.

5. CONCLUSION

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. We have also shown the probabilistic interpretation of the polynomial classifier output, and the subsequent fusion of this two-classifier system. The resulting performance of the multimodal system is shown to outperform either modality in isolation in quiet conditions. Additionally, the fused system exhibits good behavior when the output of the audio mode classifier becomes unusable.

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 [16].
CVSSP URL: http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb.

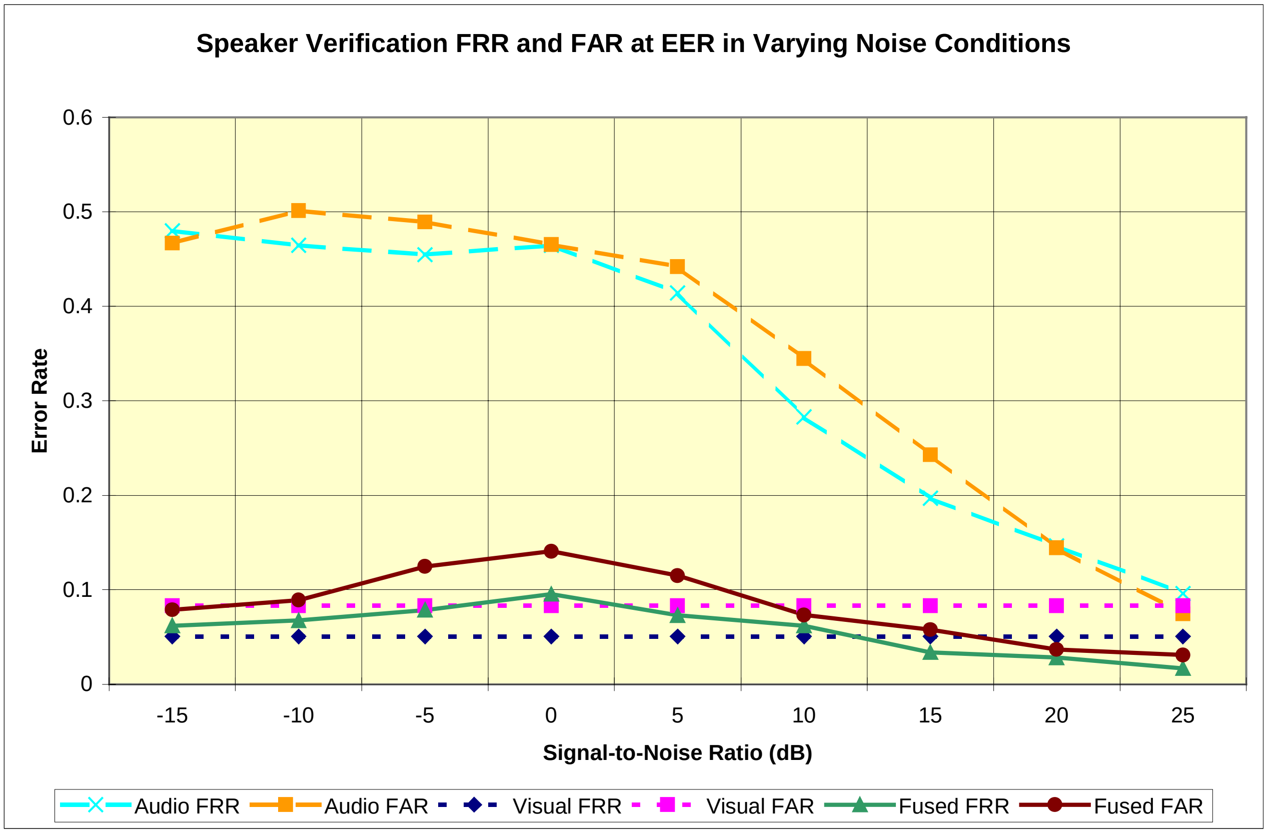

Figure 7: Performance of audio-visual speaker verification in noisy conditions.

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. K. Jack, "Video Demystified – A Handbook for the Digital Engineer," 1996.
4. X. Zhang, R. M. Mersereau, "Lip Feature Extraction Towards an Automatic Speechreading System," in *Proceedings of the International Conference on Image Processing*, 2000.
5. J. F. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679-698, 1986.
6. T. Poggio, V. Torre, C. Kock, "Computational Vision and Regularization Theory," *Nature*, vol. 317 (26), pp. 314-319, September 1985.
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. P. Campbell, Jr., "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437-1462, Sept. 1997.
9. A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 599-602, 1992.
10. D. A. Reynolds, "Automatic speaker recognition using gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173-192, 1995.
11. K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. on Speech and Audio Processing*, vol., pp. 194-205, Jan. 1994.
12. J. Schürmann, *Pattern Classification*. John Wiley and Sons, Inc., 1996.
13. K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
14. W. M. Campbell and K. T. Assaleh, "Polynomial classifier techniques for speaker verification," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 321-324, 1999.

15. W. M. Campbell and C. C. Broun, “A Computationally Scalable Speaker Recognition System,” in *Proceedings of EUSIPCO*, 2000
16. 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.
17. J. Luettin and G. Maitre, “Evaluation Protocol for the XM2VTS Database,” IDIAP-Com 98-05, October 1998.