International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

A novel approach of speaker authentication by fusion of speech and image features using Artificial Neural Networks

[1]Jaweria Izhar, [2]Dr. Jitendra Seethlani

[1]Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, (India)

ABSTRACT

This paper presents a decision fusion technique for a bimodal biometric verification system that makes use of facial and speech biometrics. This report considers multimodal biometric systems and their applicability to access control, authentication and security applications. We have simulated three Artificial Neural Network (ANN) models: firstly, speaker identification by speech parameters, secondly person identification by image parameters and finally the person authentication by fusion of speech and image feature. All the three ANN models are trained by Back-propagation algorithm.

I INTRODUCTION

Decision of fusion technique for a bimodal biometric verification system that creates use of facial and speech biometry. Biometric identity authentication systems area unitsupported the biological characteristics of someone, like face, voice, fingerprint, iris, gait, hand pure mathematics or signature. Identity authentication using the face or the voice datamay be adifficultanalysisspacethat'spresentlyterribly active, mainlybecause of the natural and non-intrusive interaction with the authentication system. An identity authentication system musthandle2styles of events: either the person claiming a given identity is that the one who he claims to be known asconsumer or if it's not then it'sadeceiver. Moreover, the system mightusually take one decision: either settle for the consumer or reject him and choosehe'sadeceiver.

Depending upon the nature of the application, speaker identification or speaker verification systems could be modeled to operate either in text dependent or text-independent modes. For text dependent speaker authentication, the user is required to utter a specific password, while for text-independent ASR; there is no need for such a constraint. Success in both cases depends on the modeling of speech characteristics which distinguish one user from the other. Text-dependent SA is used for applications where the user is willing to cooperate by memorizing the phrase or password to be spoken which could be inconvenient to some users. Therefore, in this project, our focus is

on text-independent SAwhich is considered to be a more challenging problem. Face recognition research has also been about for over three decades with comprehensive surveys like Zhao et al. Classical face recognition research was based on matchingsingle pair's holistic facial images. Later, multiple independent images per individual were used to train a Linear Discriminate Analysis (LDA) classifier and recognition was performed on a single test image. These techniques did not cope well with changes in illumination, pose and facial expressions. Face recognition has become popular over the past few years because of the common availability of cameras and their ability to capture more information. Moreover, motion helps in the recognition of faces.

Multi-modal biometric research has recently gained popularity. Biometrics from independent methodology complements each other and increase the accuracy and robustness of the system. Speech and face are natural choices for multimodal biometric applications because they can be simultaneously acquired with camera and microphone. A review of audio-visual person. Identification and verification is given by Sanderson and Paliwal. A detailed book on the subject including fusion techniques is also available. The speaker identification module gets the input from the microphone the preprocessing like voice activity detection is used to detect the start and stop of the voice sample. The MFCC is calculated as the extracted feature and the decision is made using the Hidden Markov Model to calculate the likelihoods. Low resolution camera is used to capture image for face recognition module, the preprocessing algorithm are employed like filtering to remove high frequency noise. The geometric normalization is used to remove the variation between size, orientation and location of the face in the image. The feature extraction module uses principal component analysis (PCA) decomposition on the training set, which produces the Eigen vector and Eigen values. The classification module identifies the face in a face space. The critical parameter in this classification step is the subset of eigenvector used to represent the face. The nearest neighbor classifier is used as a main classifier which ranks the gallery image by similarity measure. For similarity measure the angle between feature vector and Mahalanobis Distances is used to provide the decision.

PHASES OF SPEAKER IDENTIFICATION SYSTEM

The process of speaker identification is divided into two main phases. The first phase is called the enrollment phase or learning phase. In this phase speech samples are collected from the speakers, and they are used to train their models. The collection of enrolled models is also called a speaker database. The processes of the enrollment phase [1] are represented in Figure. 1

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

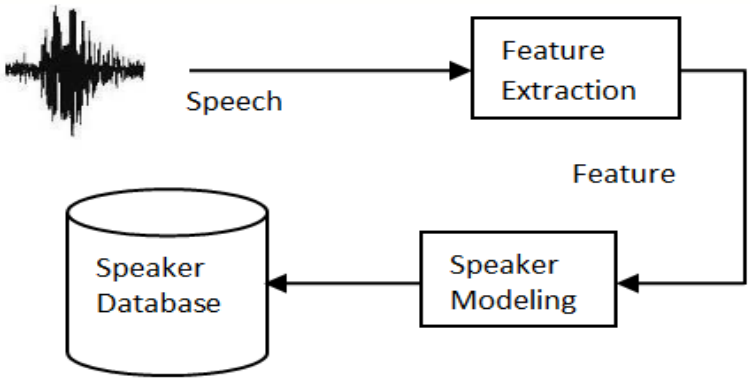
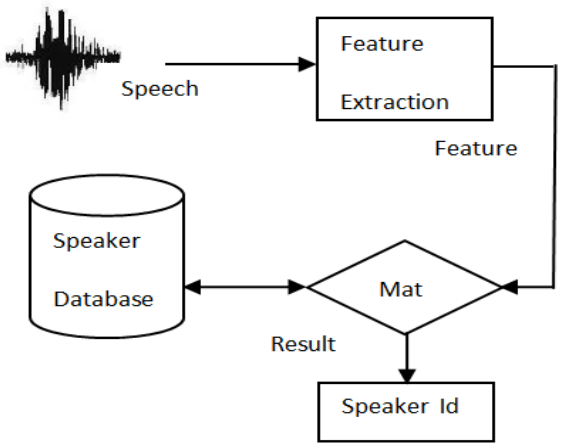


Figure 1. Enrollment phase

The second phase is called the identification phase, in which a test sample from an unknown speaker is compared against the speaker database. Both phases include the same initial step, feature extraction, which is used to extract speaker dependent characteristics from speech. The main purpose of the features extraction step is to reduce the amount of test data while retaining speaker discriminative information. The processes of the identification phase [1] are represented in Figure. 2

However, these two phases are closely related, and so the identification algorithm usually depends on the modeling algorithm used in the enrollment phase



. Figure.2 Identification Phase

THE SPEECH SIGNAL

A signal is defined as any physical quantity that varies with time, space, or any other independent variable or variables. Speech signals are examples of information bearing signal that evolve as functions of a single

independent variable namely, time [24]. A speech signal is a complex signal, can be represented as

$$s(n) = h(n) * u(n)$$

Where, the speech signal s(n) is the convolution of a filter h(n) and some signal u(n). h(n) is also called the impulse response of the system. In our system (human body) h(n) is related with teeth, nasal cavity, lips etc. u(n) is approximately a periodic impulse train referred to as the pitch of speech, where pitch is synonymous with frequency.

IMAGE FEATURES

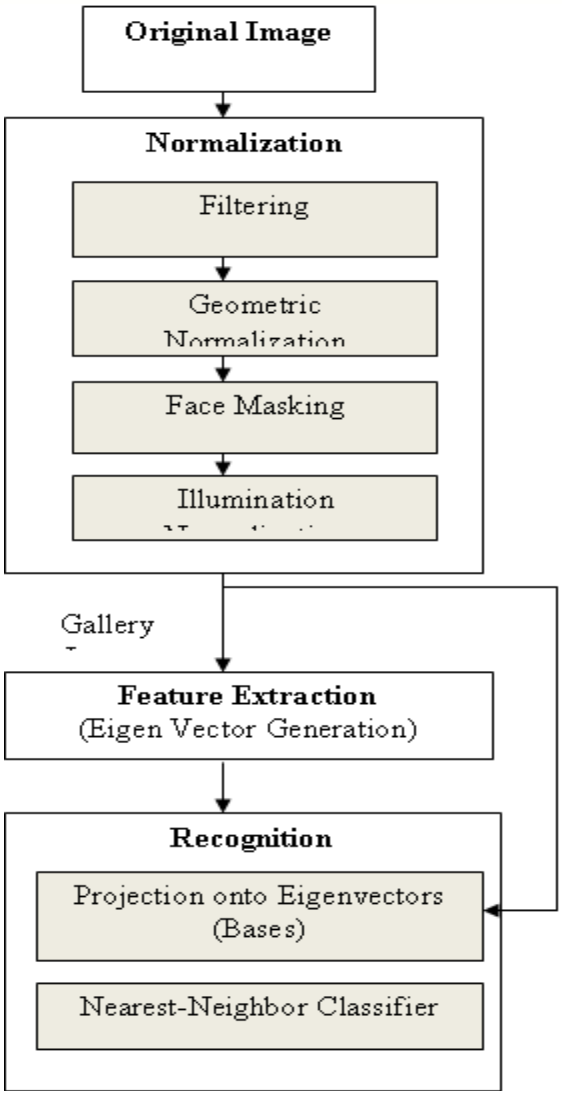
PCA is a statistical dimensionality reduction method whichproduces optimal linear least squares decomposition of a training set Kirby and Sirovich (1990) applied PCA to representing faces and Truk and Pentland (1991) extended PCA to recognizing faces. In PCA based face recognition algorithm, the input is training set t1…tN of N facial images such that ensemble mean of the training set is zero. In computing the PCA representation, each image is interpreted as a point in IRN×M, where each image is N×M pixel.PCA finds the optimal linear least square representation in N-1 dimensional space, with the representation preserving variance .the PCA representation is characterized by a set of N-1 Eigen vectors (e1,…eN-1) and Eigen values (λ1…. λ N-1).in the face recognition literature, Eigen vectors can be refereed as a Eigen faces.

We normalize the Eigen vectors so that they are orthonormal. The Eigen vectors are ordered so that $\lambda 1 \geq \lambda i+1$. The λiare equal to the variance of the projection of the training set on to the ith Eigen vector. Thus the lower order Eigen vectors encode the larger variations in the training set. The lower order refers to the index of Eigen vectors and Eigen values. The higher order Eigen vectors encode smaller variations, it is commonly assumed that they represent noise in the training set because of this assumption and empirical results, higher order Eigen vectors are excluded from the representation. Faces are represented by their projection on to a subset of $M \leq N-1$ Eigen vectors, which we call face space .thus a facial image is represented as a point in an M dimensional face space.

The first step is normalization of the input image. The goal of the normalization step is to transform the facial image into a standard format that removes or attenuates variations that can affect recognition performance. This step consists of four sub steps. The first sub step low-pass filters or compresses the original image. Images are filtered to remove high-frequency noise. An image is compressed to save storage space and reduce transmission time. The second sub step places the face in a standard geometric position by rotating, scaling, and translating the center of eyes to standard locations. The goal of this sub step is to remove variations in size, orientation, and location of the face in an image.

Flow diagram of Face identification Module

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354



II FEATURE EXTRACTION

We can think about speech signal as a sequence of features that characterize both the speaker as well as the speech. The amount of data, generated during the speech production, is quite large while the essential characteristics of the speech process change relatively slowly and therefore, they require less data. Hence we can say that feature extraction is a process of reducing data while retaining speaker discriminative information. A variety of choices for this task can be applied. Some commonly used methods for speaker identification is linear prediction and mel-cepstrum.

The cep-strum coefficients are the result of a cosine transformation of the real logarithm of the short time energy spectrum expressed on a Mel-frequency scale. This is a more robust, reliable feature set for speech recognition than

the LPC coefficients. The sensitivity of the low order cep-strum coefficients to overall spectral slope, and the sensitivity of the high-order cepstrum coefficients to noise, has made it a standard technique. It weights the cepstrum coefficients by a tapered window so as to minimize these sensitivities, frame and these are used as the feature vector. In MFCC‟s, the main advantage is that it uses mel frequency scaling which is very approximate to the human auditory system. The coefficients generated by algorithm are fine representation of signal spectra with great data compression.

III DESIGN AND IMPLEMENTATION

The design and implementation of the text independent speaker identification system can be subdivided into two main parts: Speech signal processing and artificial neural network. The Speech signal processing contains speech signal acquisition and feature extraction. The neural network part consists of two main subparts: Learning and Identification.

The first part of text independent speaker identification i.e. speech signal processing consists of several steps. Firstly, we collected the speech signal using microphone and used low pass filter to remove noise from the signal. Then we detected the start and end point of the signal using the energy theory. Finally, we extracted the exact and effective features applying the feature extraction procedure. These extracted features were then fed into the neural network.

The second part of text independent speaker identification is the Artificial Neural Network (ANN). The first subpart of the artificial neural network is the learning and the second is the identification. For learning or training we applied the back propagation learning algorithm. We adjusted the weight and threshold in learning phase, and saved into the database. In the identification phase, we used the database from learning algorithm to match the unknown speech signals.

IV RESULTS AND PERFORMANCE ANALYSIS

We performed our experiment considering different issues. We took different error rate (5 times) and completed the execution of the experiment again and again. In this case, we noticed how the error rate affects the identification of the speakers. We considered the identification capability of the network against the static speech signals and the instant speech recorded signal. The speaker database consisted of 16 speech samples from 8 speakers. Speakers were asked to read a given text in normal speed, under normal laboratory conditions. The same microphone was used for all recordings. For each speaker, two files were recorded; one for training and one for testing. Training and testing samples were recorded about 2 seconds long. The number of tests in each was 16. The experimented results are shown in the Table I.

Table 1
RESULTS WITH VARYING ERROR RATE

Error Rate	Successfully identify	Error result shown
0.1	**4**	**12**
0.05	**6**	**10**
0.03	**8**	**8**
0.01	**12**	**4**
0.005	**14**	**2**

V CONCLUSION

This research considers multimodal biometric systems and their applicability to access management, authentication and security applications. We've got simulated three ANN models: first, identification by speech parameters, second person identification by image parameters and at last the person authentication by fusion of speech and image feature. The entire three artificial neural network models area unit trained by Back-propagation formula.

REFERENCE

1. Furui, S.: An Overview of Speaker Recognition Technology. In: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification (1994).
2. Pawlewski, M., Jones, J.: Speaker Verification: Part 1. BiometricTechnology Today 14(6), 9–11 (2006).
3. Reynolds, D.: A Gaussian Mixture Modeling Approach to TextindependentSpeaker Identification. PhD Thesis, Georgia Institute ofTechnology (1992).
4. McLachlan, G.: Mixture Models, vol. Wright, J. and Yang, A. andGanesh, A. and Sastri, S, S. and Ma, Y. Marcel Dekker, New York (1988).
5. Tishby, N.: On the Application of Mixture AR Hidden Markov Models toText independent Speaker Recognition. IEEE Trans. on Signal Proc. 39,563–570 (1991).
6. Poritz, A.: Linear Predictive Hidden Markov Models and the SpeechSignal. In: Proceedings of IEEE ICASSP, pp. 1291–1294 (1982).
7. Rosenberg, A.: Sub-word Talker Verification using Hidden MarkovModels. In: Proceeding of IEEE ICASSP, pp. 269–272 (1990).
8. Levinson, D.: A Perspective on Speech Recognition. CommunicationMagazine 28 (1990).
9. Kohata, M.: Interpolation of LSP Coefficients using Recurrent NeuralNetworks. Electronics Letters 32 (1996).

10. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition:A Literature Survey. ACM Computing Survey 35(4), 399–458 (2003)
11. Turk, M., Pentland, A.: Eigenfaces for Recognition. JournalofCognitive Neuroscience 3, 71–86 (1991).
12. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces:Recognition Using Class Specific Linear Projection. IEEE Trans. on PAMI19, 711–720 (1997).
13. Sanderson, C., Paliwal, K.: Identity Verification Using Speech and FaceInformation. Digital Signal Processing 14(5), 449–480 (2004).
14. Sanderson, C.: Biometric Person Recognition: Face, Speech and Fusion.VDM Verlag (2008).
15. Turk, Matthew A. and Pentland, Alex P., Face Recognition UsingEigenfaces. Proc. IEEE Conference on Computer Vision and PatternRecognition, Maui, Hawaii, 1991.
16. Etemad, K. and Chellappa, R., Discriminant Analysis for Recognition ofHuman Face Image. *Journal of Optical Society of America A*, Vol. 14, pp.1724-1733, 1997.