FAST ALGORITHM FOR NOISY SPEAKER RECOGNITION USING ANN

Amr Rashed
Deanship of Higher Studies and Research, Taif university, Taif, Saudi Arabia

ABSTRACT

Speaker Recognition is a multi-disciplinary technology which uses the vocal characteristics of speakers to deduce information about their identities. It is a branch of biometrics that may be used for identification, verification, and classification of individual speakers, with the capability of tracking, detection, and segmentation by extension. This paper proposed a fast algorithm for speaker recognition. This algorithm first records voice patterns of speakers via noisy channel and use some of noise removal techniques. The feature is extracted by Mel Frequency Cepstral Coefficient (MFCC). Then the feature is reduced by Principal component analysis technique. Then the result vector is fed to ANN classifier. Experimental results indicates that using ANN with weight/bias training algorithm have better performance. The result shows that the proposed algorithm achieved on average about 99% accuracy rate and higher speed rate in comparison with other methods.

Keywords: Speaker Identification, Speaker Recognition, Artificial Neural Network (ANN), Frequency Mel Cepstral Coefficients (MFCC), Principal Component Analysis (PCA).

I. INTRODUCTION

The speech signal conveys many levels of information to the listener. At the primary level, speech conveys a message via words. But at other levels speech conveys information about the language being spoken and the emotion, gender and, generally, the identity of the speaker. While speech recognition aims at recognizing the word spoken in speech, the goal of automatic speaker recognition systems is to extract, characterize and recognize the information in the speech signal conveying speaker identity.

The general area of speaker recognition encompasses two more fundamental tasks. Speaker identification is the task of determining who is talking from a set of known voices or speakers. The unknown person makes no identity claim and so the system must perform a 1:N classification. Generally it is assumed the unknown voice must come from a fixed set of known speakers, thus the task is often referred to as closed-set identification. Speaker verification (also known as speaker
authentication or detection) is the task of determining whether a person is who he/she claims to be (a yes/no decision). Since it is generally assumed that imposters (those falsely claiming to be a valid user) are not known to the system, this is referred to as an open-set task. By adding a “none-of-the-above” option to closed-set identification task one can merge the two tasks for what is called open-set identification. [1]

The paper is structured as follows. In Section 2 we discuss Related Work section. In Section 3, 4 we discuss Application and Constrains that faces our proposed algorithm. In section 5 we explain our proposed technique. In section 6 we discuss the results. The paper is ended with a conclusion and future work.

II. RELATED WORK

One of the most challenging contemporary problems is that recognition accuracy degrades significantly if the test environment is different from the training environment and/or if the acoustical environment includes disturbances such as additive noise, channel distortion, speaker differences, reverberation, and so on. Over the years dozens if not hundreds of algorithms have been introduced to address this problem. Many of these conventional noise compensation algorithms have provided substantial improvement in accuracy for recognizing speech in the presence of quasi-stationary noise. Unfortunately these same algorithms frequently do not provide significant improvements in more difficult environments with transitory disturbances such as a single interfering speaker or background music. Virtually all of the current systems developed for automatic speech recognition, speaker identification, and related tasks are based on variants of some features extraction techniques such as MEL frequency cepstral coefficients (MFCC) [2].

Some use the Linear Predictive, also known as Auto-Regressive (AR) features by themselves: Linear Predictive Coefficients (LPC), Partial Correlation (PARCOR) – also known as reflection coefficients, or log area ratios. However, mostly the LPCs are converted to cepstral coefficients using autocorrelation techniques. These are called Linear Predictive Cepstral Coefficients (LPCCs). There are also the Perceptual Linear Predictive (PLP) features. PLP works by warping the frequency and spectral magnitudes of the speech signal based on auditory perception tests. The domain is changed from magnitudes and frequencies to loudness and pitch [3].

III. APPLICATION

Speaker recognition. Speaker recognition has been applied most often as a security device to control access to buildings or information. One of the best known examples is the Texas Instruments corporate computer center security system. Security Pacific has employed speaker verification as a security mechanism on telephone-initiated transfers of large sums of money. In addition to adding security, verification is advantageous because it reduces the turnaround time on these banking transactions. Bellcore uses speaker verification to limit remote access of training information to authorized field personnel. Speaker recognition also provides a mechanism to limit the remote access of a personal workstation to its owner or a set of registered users.[4]

In addition to its use as a security device, speaker recognition could be used to trigger specialized services based on a user’s identity. For example, you could configure an answering machine to deliver personalized messages to a small set of frequent callers.

In May 2013 it was announced that Barclays Wealth was to use speaker recognition to verify the identity of telephone customers within 30 seconds of normal conversation. The system used had been developed by voice specialists Nuance, the company behind Apple’s Siri technology. A verified
voiceprint was to be used to identify callers to the system and the system would in the future be rolled out across the company.

The private banking division of Barclays was the first financial services firm to deploy voice biometrics as the primary means to authenticate customers to their call centres. 93% of customer users had rated the system at "9 out of 10" for speed, ease of use and security.[5]

IV. CONSTRAINTS AND DATA COLLECTION

Due to the limited time interval; a set of constraints have been placed on the system to make the algorithm more manageable. These constraints are: A 58 speech files are received by a sensitive wide band microphone and saved as samples of voice signals to be analysed (wave files). These files are pronounced by (11) different persons, (47) different word used for text independent speaker identification algorithm and (11) files recording the same word for text dependent speaker identification algorithm.

Afterwards, data is stored in wave format and analysed using Matlab software, figure.1 shows an example of the voice signal in time domain.

![Voice signal in time domain](image)

V. PROPOSED TECHNIQUE

The algorithm has four steps: (A) Noisy (WGN) white Gaussian Noise Channel (B) noise reduction and silence removal (C) feature extraction using MFCC feature extraction techniques. (D) Principal component analysis (PCA) Technique for feature reduction. (E) Artificial Neural Network (ANN).

A. NOISY (WGN) CHANNEL

After data is stored in wave format, a White Gaussian Noise(WGN) is added to each file using Matlab function (awgn) with different signal to noise ratio(SNR). power of the original signal are measured using Matlab.

Figure 3 shows a sample of voice signal in time domain after adding white Gaussian noise with SNR=-10DB.
B. NOISE REDUCTION AND SILENCE REMOVAL

Speech enhancement aims to improve speech quality by using various algorithms. The objective of enhancement is improvement in intelligibility and/or overall perceptual quality of degraded speech signal using audio signal processing techniques. Enhancing of speech degraded by noise, or noise reduction, is the most important field of speech enhancement, and used for many applications such as mobile phones, VoIP, teleconferencing systems, speech recognition, and hearing aids [6]. Minimum Mean-Square-Error Short-Time Spectral Amplitude Estimator (MMSE-STSA) is used algorithm for noise reduction and speech enhancement. One voice signal after noise reduction is shown in figure 3.

C. FEATURE EXTRACTION

Feature extraction is the transformation of the original data (using all variables) to a data set with a reduced number of variables. In the problem of feature selection, the aim is to select those variables that contain the most discriminatory information. Alternatively, we may wish to limit the number of measurements we make, perhaps on grounds of cost, or we may want to remove redundant or irrelevant information to obtain a less complex classifier.

In feature extraction, all variables are used and the data are transformed (using a linear or nonlinear transformation) to a reduced dimension space. Thus, the aim is to replace the original variables by a smaller set of underlying variables. There are several reasons for performing feature
extraction: (i) to reduce the bandwidth of the input data (with the resulting improvements in speed and reductions in data requirements) ;(ii) to provide a relevant set of features for a classifier, resulting in improved performance, particularly from simple classifiers; (iii) to reduce redundancy;(v ) to recover new meaningful underlying variables or features that the data may easily be viewed and relationships and structure in the data identified [7]

The cepstral coefficient provides a better alternative to the LP coefficient for speech and speaker recognition. The cepstral coefficient can be derived either through LP analysis or MEL filter bank analysis. The former method generates features which are more commonly known as the LP cepstral coefficient .the MLP cepstral coefficient can easily be calculated from the PLP coefficients by [8].

Figure 4 shows MFCC features for all input signals.

Figure 4. Shows MFCC feature for all input signals

D. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convertset of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables.

Depending on the field of application, it is also named the discrete Karhunen–Loève transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, proper orthogonal decomposition (POD) in mechanical engineering, singular value decomposition (SVD) of X, eigen value decomposition (EVD) of XTX in linear algebra, factor analysis, Eckart–Young theorem , or Schmidt–Mirsky theorem in psychometrics, empirical orthogonal functions (EOF) in meteorological science, empirical Eigen function decomposition ,empirical component analysis , quasiharmonic modes , spectral decomposition in noise and vibration, and empirical modal analysis in structural dynamics.[9]

PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualised as a set of coordinates in a high-
dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection or "shadow" of this object when viewed from its (in some sense; see below) most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

PCA is also related to canonical correlation analysis (CCA). CCA defines coordinate systems that optimally describe the cross-covariance between two datasets while PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset. [10, 11]

Figure 5 shows principal components curves for all input signals.

E. ARTIFICIAL NEURAL NETWORK

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable [12, 13, and 14].

Classifiers based on neural networks (N,N) are used in both text dependent and text independent speaker identification and speaker verification system. The NN is extremely efficient at learning complex mappings between inputs and outputs and is able to approximate posterior probabilities for the trained classes. The neural networks are able to approximate nonlinear decision surfaces and exhibit a high level of parallelism[15].

Figure 6. Multilayer perceptron(MLP) used here
VI. RESULTS

Using MFCC features with multilayer perceptron neural network which contains three layers feed forward with 11 input neurons, 10 hidden neurons, and one neuron at the output layer. The used training algorithm is sequential weight/bias rule with 6043 iterations, gives us neural network performance 0.000999 as shown in figure 7, 8.

Figure 7 shows Relation between MSE and number of Epochs with MFCC feature extraction, and figure 8 shows regression curve (relation between target and output) for training, testing and validation data.

Sequential weight/bias rule training algorithm gives us fast training and best accuracy than any other training algorithms, figure 9 shows a sample Matlab script of the algorithm used in this paper.

![Figure 7. Relation between MSE and number of Epochs with MFCC feature extraction](image1)

![Figure 8. Regression curve (relation between target and output) for training, testing and validation data](image2)
Figure 9. Sample Matlab script used for this algorithm
VII. FUTURE WORK

Modern FPGA platforms provide the hardware and software infrastructure for building a bus-based system on chip (SoC) that meet the applications requirements. The designer can customize the hardware by selecting from a large number of pre-defined peripherals and fixed IP functions and by providing new hardware, typically expressed using RTL. In order to accelerate the system we can implement ANN classifier using FPGA with parallel processing instead of using Matlab. We expect that we can achieve an overall LPR system speed up. Also web based authentication system can be made using automatic speaker recognition algorithm.

VIII. CONCLUSIONS

In this paper a comparative study is made between different neural network learning algorithms to choose the best one to apply for automatic speaker recognition system. A feed forward back propagation (MLP) multilayer perceptron neural network with sequential weight/bias rule training algorithm gives us fast and accurate result better than any other training function like Levenberg-Marquardt Back Propagation. ANN classifier used is contained of 11 input neurons, 10 hidden neurons and 1 output neuron. We achieve accuracy about 99.9% with performance 0.0000999 for MFCC features taken from input data.

When adding white Gaussian noise to all input patterns with different signal/noise ratio (SNR). It did not affect classifier performance in our test. Sequential weight/bias training algorithm gives best results when using in speaker or speech recognition and it is fast compared with any other training algorithms.

REFERENCES


