Modified Technique for Speaker Recognition using ANN

Amr Rashed, Waleed M. Bahgat
Deanship of Information technology, Taif university, Tai f, Saudi Arabia

Information technology Department, Faculty of computers and information sciences, Mansoura, Egypt

Abstract
Speaker recognition consists of three phases: pre-processing, feature extraction and classification. During the first phase, the computer records the voice pattern of the speaker and analyse it. By the end of the second phase, the main features of the voice pattern are extracted. In the third phase, many classification techniques are exist such as artificial neural network (ANN) , hidden Markov model (HMM) and vector quantization (VQ). Classifiers based on ANN are used in both text dependent and text independent speaker identification and speaker verification systems. Furthermore, it is extremely efficient at learning complex mappings between input and outputs. Unfortunately, ANN technique is complex and time consuming. In this paper, we use two different feature extraction techniques. These techniques are MFCC and PNCC. In addition, we use principle component analysis (PCA) as a feature reduction technique to enhance the classifier performance and speed. We apply ANN for both techniques with different training algorithms. The best results are achieved using PNCC as a feature extraction, the ANN as a classifier with sequential weight/bias training algorithm. Our proposed technique decreases the number of neurons that lead to have best performance and processing time.

Keywords
Speaker Identification, Speaker Recognition, Artificial Neural Network, MEL Cepstral coefficients , Pprincipal Component Analysis and PNCC.

1. INTRODUCTION

Speaker recognition is a branch of biometric authentication which refers to the automatic identity recognition of individuals using certain intrinsic characteristics of the person. Speaker recognition has many potential applications as a biometric tool since there are many tasks that can be performed remotely [1]. Biometric authentication has been an important technique for human-machine communication system in applications with security considerations [2]. not only the voice, there are many other physical and behavioural patterns examples such as eyes, faces, fingerprint, signature, etc., especially for biometric authentication. the development of speech processing technology has boosted many applications of speaker recognition. this technology or techniques makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialling, mobile banking, mobile shopping, database access services, information services, voice mail, and security control for confidential information areas, and remote access to computers [3].The areas that are using the speech processing can be listed as follows: (1) Access control to physical facilities, data networks and terminal, (2). Mobile purchases through electronic bank transaction and credit card. (3) Checking the available balance services and card activation. (4) Customer information retrieval services, e.g. customer information call. (5) Remote monitoring. (6) Forensic voice sample matching.

According to [3] speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. Compared the entire biometric authentication,[2]come out with a comparison of some biometric patterns to identity and classify the biometric patterns into the following classification.

1. Distinctiveness: the existence of wide differences in the pattern among the population.
2. Robustness: repeatable, not subject to large changes.
3. Accessibility: easily presented to sensor.
4. Acceptability: perceived as non-intrusive by the users.

The paper is structured as follows. In Section 2 we discuss Related Work section. In Section 3, 4 we discuss Difficulties 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.

2. 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

Manuscript received August 5, 2013
Manuscript revised August 20, 2013

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) [4].

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 [5].

3. DIFFICULTIES

There are a few special problems within the field of speaker recognition. These problems include: (a)human comprehension of speech compared to ASR. (b)body language.(c)noise.(d)spoken language is not the same with written language.(e) continuous speech [6] .(f)Channel mismatch is the most serious difficulty faced speaker recognition technology. Channel mismatch is the source of most errors in speaker recognition.(g) The signal variation problem, as stated earlier, is common to most biometrics. Some of these variations may be due to aging and time-lapse effects[5].(h) Another problem is the speaker variability which means that the voice is not only different between speakers; there are also wide variation within one specific speaker and this is due to different realization, speaking style, sex, anatomy of vocal tract, speed of speech, regional and social dialects[5].

4. 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, 2 shows an example of the voice signal in time domain and amplitude spectrum in frequency domain.

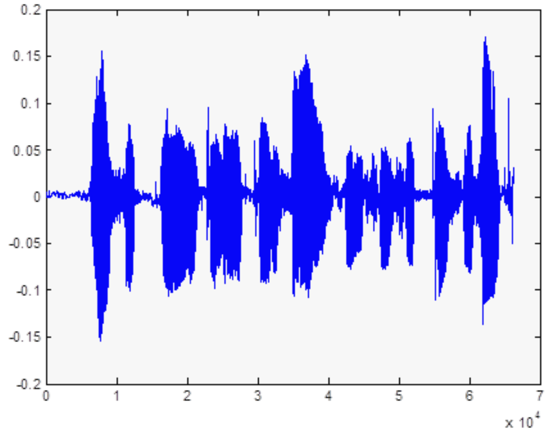

Fig 1.voice signal in time domain

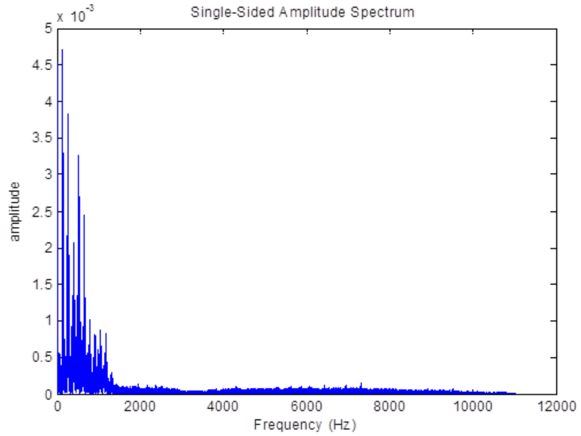

Fig 2.voice signal in frequency domain

5. PROPOSED TECHNIQUE

The algorithm has four steps: (A) noise reduction and silence removal (B) feature extraction using two different feature extraction techniques and compare between their results. (C) Principal component analysis. (D) Artificial neural network.

A. 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 [7]. Minimum Mean-Square-Error Short-Time Spectral Amplitude Estimator (MMSE-STSA) is used algorithm for noise reduction and speech enhancement. The single-sided amplitude spectrum after noise reduction is shown in figure 3.

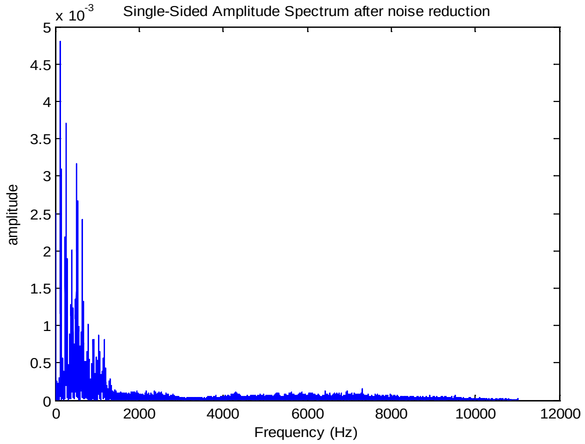

Fig 3.Single-sided amplitude spectrum after noise reduction

B. 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 [8]

The cepstral coefficient provide 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 *M*LP cepstral coefficient can easily be calculated from the *P*LP coefficients by [9]

$$C_0=\ln G^2_P$$

A new feature extraction algorithm called Power-Normalized Cepstral Coefficients (PNCC) that is based on auditory processing. Major new features of PNCC processing include the use of a power-law nonlinearity that replaces the traditional log nonlinearity used for MFCC coefficients, and a

novel algorithm that suppresses background excitation by estimating SNR based on the ratio of the arithmetic to geometric mean power, and subtracts the inferred background power. Experimental results demonstrate that the PNCC processing provides substantial improvements in recognition accuracy compared to MFCC and PLP processing for various types of additive noise. The computational cost of PNCC is only slightly greater than that of conventional MFCC processing [10].

Figure 4 shows a comparison between MFCC feature extraction and PNCC feature extraction [11,12].

As in figure MFCC ,PNCC are distinguished mainly in three steps founded in PNCC .these steps are gamma-tone frequency integration, medium-time power bias subtraction, and power function nonlinearity .these steps are met with triangular frequency integration and logarithmic nonlinearity in MFCC.

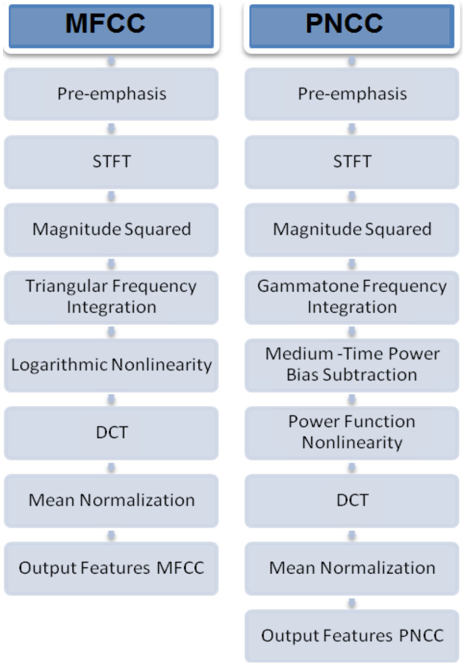

Figure 4 shows a comparison between MFCC feature extraction and PNCC feature Extraction

C. Principal Component Analysis

Principal components analysis (PCA) is originated in work by Pearson (1901).Geometrically, principal components analysis can be thought of as a rotation of the axes of the original coordinate system to a new set of orthogonal axes that are ordered in terms of the amount of variation of the original data they account for [8].

PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centring (and normalizing or using Z-scores) the data matrix for each attribute[13].

In some situations, the dimension of the input vector is large, but the components of the vectors are highly correlated (redundant). It is useful in this situation to reduce the dimension of the input vectors. An effective procedure for performing this operation is principal component analysis. This technique has three effects:

1. It orthogonalizes the components of the input vectors (so that they are uncorrelated with each other).
2. It orders the resulting orthogonal components (principal components) so that those with the largest variation come first.
3. It eliminates those components that contribute the least to the variation in the data set [8].

D. Artificial neural network

An ANN is an information processing system that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a large Number of highly interconnected processing elements (Neurons) working with each other to solve specific Problems. Each processing element (neuron) is basically a summing element followed by an activation function. The output of each neuron (after applying the weight parameter associated with the connection) is fed as the input to all of the neurons in the next layer. The learning process is essentially an optimization process in which the parameters of the best set of connection coefficients (weighs) for solving a problem are found and includes the following basic steps:

- Present the neural network with a number of inputs (Vectors each representing a pattern)
- Check how closely the actual output generated for a Specific input matches the desired output.
- Change the neural network parameters (weights) to better approximate the outputs [14, 15].

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 [9].

6. RESULTS

Using MFCC features with multilayer perceptron neural network which contains three layers feed forward with 194 input neuron, 192 hidden neuron, and one neuron at the output layer. The used training algorithm is sequential weight/bias rule with 801 iteration, and the neural network Performance was 0.000994 as shown in figure 5, 6.

Figure 5 shows Relation between MSE and number of Epochs with MFCC feature extraction, and figure 6 shows regression curve (relation between target and output) for training, testing and validation data .

When using PNCC features with ANN as classifier .we use multilayer perceptron with three layer feed forward with Number of input neurons=13 neuron, Number of hidden neurons=12 and one neuron at the output. the Training algorithm used is sequential weight/bias rule with Number of iteration used equal to 5476 iteration, and the output Performance of the network was 0.000999 as shown in figure 7,8.

As in figures 5,7 . Figure 7 shows Relation between MSE and number of Epochs with PNCC 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, table.1 shows comparison between training algorithm and accuracy rate achieved.

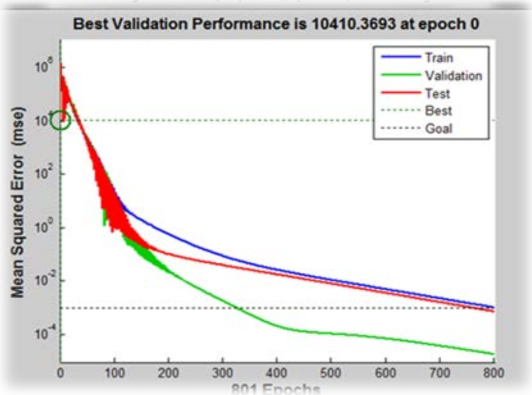

Fig 5.Relation between MSE and number of Epochs with MFCC feature extraction

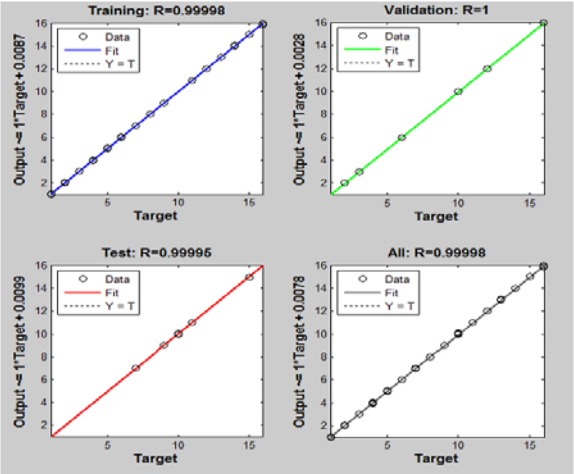

Fig.6 regression curve (relation between target and output) for training, testing and validation data.

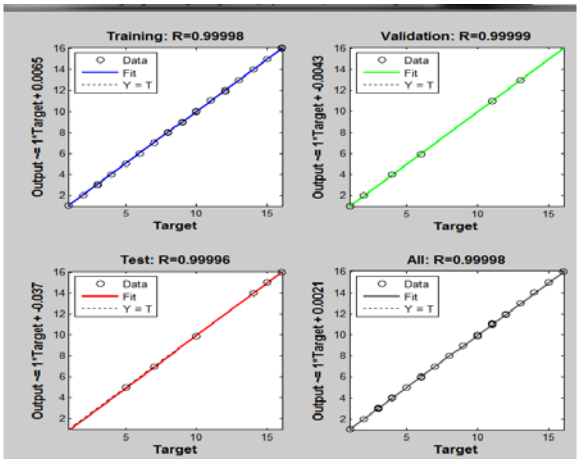

Fig .7.Relation between MSE and number of Epochs with PNCC feature extraction

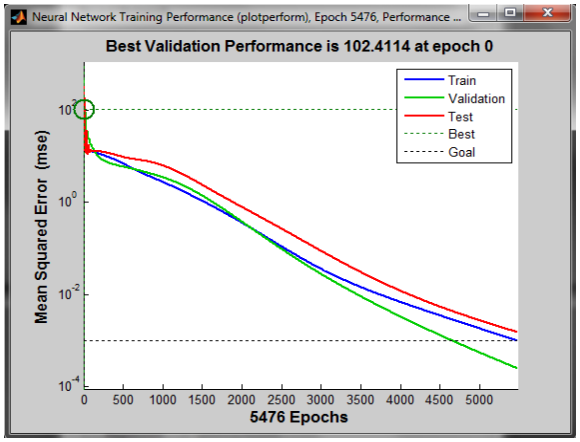

Fig .8.Regression curve for training, testing and validation signals with PNCC feature extraction

Table 1.Comparison between achieved accuracy rate with different neural network training algorithms.

Neural Network Training algorithms	***Achieved accuracy rate***
Gradient Descent with Adaptive Lr Back Propagation	*11%*
Conjugate Gradient Back Propagation with Fletcher-Reeves Updates	*61%*
Conjugate Gradient Back Propagation	*51%*
One-Step Secant Back Propagation	*0%*
Conjugate Gradient Back Propagation with Polak-Ribiere Updates	*49%*
BFGS Quasi-Newton Back Propagation	*42%*
Bayesian Regulation Back Propagation	*0%*
Levenberg-Marquardt Back Propagation	*40%*
Cyclical Order Weight/Bias Training	*9%*
Gradient Descent with Momentum and Adaptive Learning Rate Back Propagation	*4%*
Resilient Back Propagation	*22%*
Sequential Order Weight/Bias Training	*99.9%*

7. Conclusions

In this paper we compare between different learning algorithms for speaker recognition and we conclude that when using three layer (MLP) multilayer perceptron feed forward back propagation neural network with trains (sequential weight/bias rule) training algorithm. It contains of 194 input neurons, 192 hidden neurons and 1 output neuron. We achieve accuracy about 99.9% with performance 0.000994 for MFCC features taken from input data.

When using PNCC features with ANN as classifier using MLP (multilayer perceptron) with three layer feed forward back propagation neural network, with 13 input neuron, 12 hidden neuron, and one neuron at the output, with the same training algorithm used for MFCC features, Number of iteration used= 5476 iteration, Performance=0.000999.

References

[1] Hosseinzadeh D., Krishnan S. (2007), "Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs", Proc. Int. Conf. on Signal Processing, PP. 365–368.

[2] Nengheng Zheng, Tan Lee and P.C. Ching, "Comparative analysis of discrimination power of the vocal source and vocal tract features for speaker verification," in Proc. 8th

National Conference on Man-Machine Speech Communication, Beijing, China, 2005, pp.210-213

[3] http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/

[4] Mohd Syahrizad Bin Elias, "Speaker Recognition Using Enhanced MFCC", university of UTARA, Malaysia,2009.

[5] Homayoon,Beigi,"SpeakerRecognition",http://www.intechopen.com/books/biometrics/speaker-recognition,USA, 2011 .

[6] http://home.arcor.de/hertlein/navIntro.html

[7] J. Benesty, S. Makino, J. Chen (ed)." Speech Enhancement". Pp.1-8. Springer, 2005.

[8] Andrew R. Webb, QinetiQ Ltd., Malvern, Statistical Pattern Recognition, John Wiley & Sons Ltd.2002.

[9] "Text independent speaker recognition using source based features", master thesis,2001.

[10] Chanwoo Kim2,Richard M Stren1,2,"Feature Extraction for Robust Speech Recognition using A Power-Law Nonlinearity and Power-Bias Subtraction",1Departement of Electrical and Computer Engineering and 2 language technologies institute Carnegie Mellon university, Pittsburgh, INTERSPEECH-2009, pp. 28-31, Sept. 2009.

[11] Gellért Sárosi1, Mihály Mozsáry1, Péter Mihajlik1,2, and Tibor Fegyó1,3,"Comparison of Feature Extraction Methods for Speech Recognition in Noise-Free and in Traffic Noise Environment", 1 Dept. of Telecommunications and Media Informatics Budapest University of Technology and Economics Budapest, Hungary,2THINKTech Research Center Nonprofit LLC,3 Aitia International Inc.,IEEE 2011.

[12] Chanwoo Kim ,"Signal Processing for Robust Speech Recognition Motivated by Auditory Processing ",Language Technologies Institute School of Computer Science Carnegie Mellon University, phd thesis 2010.

[13] Abdi. H., & Williams, L.J."Principal component analysis." Wiley Interdisciplinary Reviews: Computational Statistics, 2: 433–459,2010.

[14] S. Y. Kung, Digital Neural Networks, Prentice Hall, Englewood Cliffs, NJ, 1993.

[15] D. Hush and B. Home, "Progress in Supervised Neural Networks: What's New since Lippmann?" IEEE Signal Processing Magazine, pp. 8-39, January 1993.

AMR RASHED received the B.Sc. and M.Sc. degrees in Electronic and Communication Engineering from Faculty of Engineering at Mansoura university(FEMU),in 2005 he is currently working on biomedical image processing ,computer vision ,MATLAB programming language.in 2007 he is working on embedded systems including FPGA and VHDL ,License Plate Recognition(LPR).in 2010 he is working on signal processing, speaker and voice recognition, traffic signals recognition. His current research is on image, signal processing, computer vision, FPGA software/hardware, and neural network .he has been lecture of electronic circuits at the Misr Engineering and Technology (MET).