International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:1, No:12, 2007

Speaker Identification using Neural Networks

R.V Pawar, P.P.Kajave, and S.N.Mali

***Abstract*—**The speech signal conveys information about the identity of the speaker. The area of speaker identification is concerned with extracting the identity of the person speaking the utterance. As speech interaction with computers becomes more pervasive in activities such as the telephone, financial transactions and information retrieval from speech databases, the utility of automatically identifying a speaker is based solely on vocal characteristic. This paper emphasizes on text dependent speaker identification, which deals with detecting a particular speaker from a known population. The system prompts the user to provide speech utterance. System identifies the user by comparing the codebook of speech utterance with those of the stored in the database and lists, which contain the most likely speakers, could have given that speech utterance. The speech signal is recorded for N speakers further the features are extracted. Feature extraction is done by means of LPC coefficients, calculating AMDF, and DFT. The neural network is trained by applying these features as input parameters. The features are stored in templates for further comparison. The features for the speaker who has to be identified are extracted and compared with the stored templates using Back Propogation Algorithm. Here, the trained network corresponds to the output; the input is the extracted features of the speaker to be identified. The network does the weight adjustment and the best match is found to identify the speaker. The number of epochs required to get the target decides the network performance.

***Keywords*—**Average Mean Distance function, Backpropogation, Linear Predictive Coding, Multilayered Perceptron,

I. Introduction

MOST signal processing involves processing a signal without concern for the quality or information content of that signal. In speech processing, speech is processed on a frame by-frame basis usually only with the concern that the frame is either speech or silence The usable speech frames can be defined as frames of speech that contain higher information content compared to unusable frames with reference to a particular application. We have been investigating a speaker identification system to identify usable speech frames. We then determine a method for identifying those frames as usable using a different approach. However, knowing how reliable the information is in a frame of speech can be very important and useful. This is where usable speech detection and extraction can play a very important role.

The usable speech frames can be defined as frames of speech that contain higher information content compared to unusable frames with reference to a particular application. We have been investigating a speaker identification system to identify usable speech frames .We then determine a method for identifying those frames as usable using a different approach.

II. Paradigms of Speech Recognition

1. Speaker Recognition - Recognize which of the population of subjects spoke a given utterance.

2. Speaker verification -Verify that a given speaker is one who he claims to be. System prompts the user who claims to be the speaker to provide ID. System verifies user by comparing codebook of given speech utterance with that given by user. If it matches the set threshold then the identity claim of the user is accepted otherwise rejected.

3. Speaker identification - detects a particular speaker from a known population. The system prompts the user to provide speech utterance. System identifies the user by comparing the codebook of speech utterance with those of the stored in the database and lists, which contain the most likely speakers, could have given that speech utterance.

There are two types of speaker identification

- Text dependent – speakers speech corresponds to known text,cooperative user, 'PIN' type applications
- Text independent – no constraints on what the speakers speaks,potentially uncooperative user

III. Approaches to Speech Recognition

1. The Acoustic Phonetic approach
2. The Pattern Recognition approach
3. The Artificial Intelligence approach

A. The Acoustic Phonetic Approach

The acoustic phonetic approach is based upon the theory of acoustic phonetics that postulate that there exist a set of finite, distinctive phonetic units in spoken language and that the phonetic units are broadly characterized by a set of properties that can be seen in the speech signal, or its spectrum, over time. Even though the acoustic properties of phonetic units are highly variable, both with the speaker and with the neighboring phonetic units, it is assumed that the rules

R.V Pawar (e-mail: rvspawar@rediffmail.com, fax: 020-24280926, phone: 9890024066).
P. P. Kajave (e-mail: kprakash@mahindrabt.com, phone: 9822876330).
S. N. Mali (e-mail: hodcomp@vit.edu, phone: 9890009182).

governing the variability are straightforward and can readily be learned and applied in practical situations.

Hence the first step in this approach is called segmentation and labeling phase. It involves segmenting the speech signal into discrete (in Time) regions where the acoustic properties of the signal are representatives of one of the several phonetic units or classes and then attaching one or more phonetic labels to each segmented region according to acoustic properties.

For speech recognition, a second step is required. This second step attempts to determine a valid word (or a string of words) from the sequence of phonetic labels produced in the first step, which is consistent with the constraints of the speech recognition task.

B. The Pattern Recognition Approach

The Pattern Recognition approach to speech is basically one in which the speech patterns are used directly without explicit feature determination (in the acoustic – phonetic sense) and segmentation. As in most pattern recognition approaches, the method has two steps – namely, training of speech patterns, and recognition of patterns via pattern comparison. Speech is brought into a system via a training procedure The concept is that if enough versions of a pattern to be recognized (be it sound a word, a phrase etc) are included in the training set provided to the algorithm, the training procedure should be able to adequately characterize the acoustic properties of the pattern (with no regard for or knowledge of any other pattern presented to the training procedure)

This type of characterization of speech via training is called as pattern classification. Here the machine learns which acoustic properties of the speech class are reliable and repeatable across all training tokens of the pattern. The utility of this method is the pattern comparison stage with each possible pattern learned in the training phase and classifying the unknown speech according to the accuracy of the match of the patterns

Advantages of Pattern Recognition Approach

- Simplicity of use. The method is relatively easy to understand. It is rich in mathematical and communication theory justification for individual procedures used in training and decoding. It is widely used and best understood.
- Robustness and invariance to different speech vocabularies, user, features sets pattern comparison algorithms and decision rules. This property makes the algorithm appropriate for wide range of speech units, word vocabularies, speaker populations, background environments, transmission conditions etc.
- Proven high performance. The pattern recognition approach to speech recognition consistently provides a high performance on any task that is reasonable for technology and provides a clear path for extending the technology in a wide range of directions.

C. The Artificial Intelligence Approach

The artificial intelligence approach to speech is a hybrid of acoustic phonetic approach and the pattern recognition approach in which it exploits ideas and concepts of both methods.

The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies intelligence in visualizing, analyzing and finally making a decision on the measured acoustic features. In particular, among the techniques used within the class of methods are the use of an expert system for segmentation and labeling.

The use of neural networks could represent a separate structural approach to speech recognition or could be regarded as an implementational architecture that may be incorporated in any of the above classical approaches.

IV. Principle of Speaker Identification

The task of automatic speaker identification consists of labeling an unknown voice as one of a set of known voices. The task can be carried out using several approaches, either with text dependent recognition or with text independent recognition. The choice of the recognition situation determines the architecture to be used. In the case of text dependent situations a time alignment the dynamic time warping (DTW) of the utterance with the test can be enough. In the case of text independent situations a probabilistic approach might be more adequate. The focus of this project is to limit to the close-set situation, where the problem consists in identifying a speaker from a group of N-known speakers, in a text dependent situation.

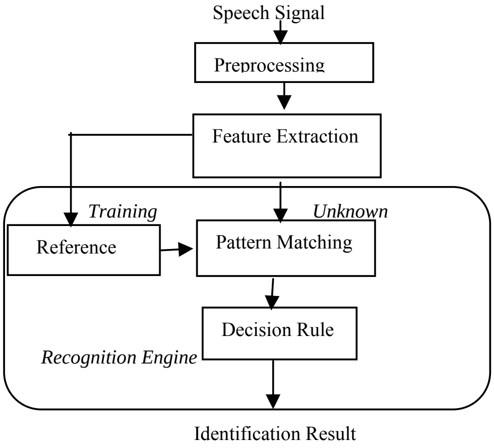

Fig. 1 Speaker Identification Model

Preprocessing

This step is to adjust an input speech signal to have better quality or to have the appropriate characteristics for the next processing step. Preprocessing covers digital filtering and endpoint detection. Filtering is to filter out any surrounding noise using several algorithms of digital filtering. Endpoint detection is a process of clamping only a desired speech interval. Several endpoint detection algorithms have been proposed for speech processing tasks (e.g. energy-based, zero-crossing, and fundamental frequency). Energy-based endpoint detection probably causes error with too small loudness of speech; protection of this error can be done during a speech recording process.

Feature Extraction
This step is to extract a set of essential characteristics that can identify or represent whole speech signal. DFT, LPC, AMDF coefficient is conducted in our system since it has shown an impressive performance on both state-of-the-art speech recognition and speaker identification tasks.

LPC Function
The LPC (Linear Predictive Coding) calculates a logarithmic power spectrum of the signal. It is used for formant analysis. The waveform is used to estimate the settings of a filter. The filter is designed in a way to block certain frequencies out of white noise. With the correct settings, the result will match the original waveform.

AMDF Function
The AMDF is mainly used to calculate the fundamental frequency of a signal. The starting point of the segment is chosen and compared with the value of the signal within a certain distance and the difference between the absolute values is calculated. This is done for all points and the differences are accumulated. This is repeated for several distances.

V. Neural Networks

Multi-Layered Perceptron (MLP)
Multi-layered networks are capable of performing just about any linear or nonlinear computation, and can approximate any reasonable function arbitrarily well. Such networks overcome the problems associated with the perceptron and linear networks. However, while the network being trained may be theoretically capable of performing correctly, back propagation and its variations may not always find a solution.

There are many types of neural networks for various applications multilayered perceptrons (MLPs) are feed-forward networks and universal approximators. They are the simplest and therefore most commonly used neural network architectures.

In this project, MLPs have been adapted for voice recognition. A general neural structure used in this work is shown in Figure

An MLP consists of three layers:

- an input layer
- an output layer
- an intermediate or hidden layer

Processing elements or neurons in the input layer only act as buffers for distributing the input signal xi to neurons in the hidden layer. Each neuron j in the hidden layer sums up its input signals xi after weighting them with the strengths of the respective connections wji from the input layer and computes its output yj as a function f of the sum, viz.,

$$Y_j=f\,(\Sigma\ w_{ji}\ x,) \qquad (1)$$

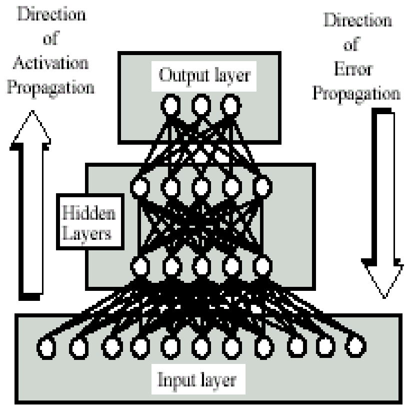

Fig. 2 Multilayer Perception

Training a network consists of adjusting its weights using a training algorithm. The training algorithms adopted in this study optimize the weights by attempting to minimize the sum of squared differences between the desired and actual values of the output neurons, namely:

$$E= \tfrac{1}{2}\ \sum_{f} (Ydj-Yj)^2 \qquad (2)$$

where ydj is the desired value of output neuron j and yj is the actual output of that neuron.

Each weight wji is adjusted by adding an increment .w_{ji} to it. W_{ji} is selected to reduce E as rapidly as possible. The adjustment is carried out over several training iterations until a satisfactorily small value of E is obtained or a given number of epoch is reached. How w_{ji} is computed depends on the training algorithm adopted.

Training process is ended when the maximum number of epochs is reached, the performance has been minimized to the goal, the performance gradient falls below minimum gradient or validation performance has increased more than maximum fail times since the last time it decreased using validation. The learning algorithm used in this work is summarized briefly.

MLP identifiers used in this work are trained with the Levenberg-Marquardt (LM) learning algorithms. The LM is a least-square estimation method based on the maximum neighborhood idea. The LM combines the best features of the Gauss-Newton technique and the steepest-descent method, but avoids many of their limitations. In particular, it generally does not suffer from the problem of slow convergence.

Backpropagation

Training Set
A collection of input-output patterns that are used to train the network.

Testing Set
A collection of input-output patterns that are used to assess network performance.

Learning Rate
A scalar parameter, analogous to step size in numerical integration, is used to set the rate of adjustments.

Network Error
Total-Sum-Squared-Error (TSSE)

$$TSSE = \frac{1}{2} \sum_{patterns} \sum_{outputs} (desired - actual)^2 \qquad (3)$$

Root-Mean-Squared-Error (RMSE)

$$RMSE = \sqrt{\frac{2 * TSSE}{\# \, patterns * \# \, outputs}} \qquad (4)$$

Apply Inputs from a Pattern

- Apply the value of each input parameter to each input node
- Input nodes computer only the identity function

Calculate Outputs for each Neuron based on the Pattern

- The output from neuron j for pattern p is O_{pj} where k ranges over the input indices and W_{jk} is the weight on the connection from input k to neuron j

$$O_{pj}(net_j) = \frac{1}{1 + e^{-\lambda net_j}} \qquad (5)$$

Calculate the Error Signal for each Output Neuron

- The output neuron error signal d_{pj} is given by d_{pj}=(Tpj-pj) Opj (1-Opj)
- T_{pj} is the target value of output neuron j for pattern p O_{pj} is the actual output value of output neuron j for pattern p

$$net_j = bias * W_{bias} + \sum_k O_{pk} W_{jk} \qquad (6)$$

Calculate the Error Signal for Each Hidden Neuron

- The hidden neuron error signal d_{pj} is given by k ranges over the input indices and W_{jk} is the weight on the

$$\delta_{pj} = O_{pj}(1 - O_{pj}) \sum_k \delta_{pk} W_{kj} \qquad (7)$$

connection from input k to neuron j neuron k and Wkj is the weight of the connection from hidden neuron j to the post-synaptic neuron k.

Compute Weight Adjustments
DW_{ji} at time t by

DW_{ji} (t) = ç d_{pj} O_{pi} (8)

Apply Weight Adjustments According

W_{ji}(t+1) = W_{ji}(t) + DW_{ji}(t)

Some Add A Momentum Term
A*Dw_{ji}(T-1)

VI. Test Results

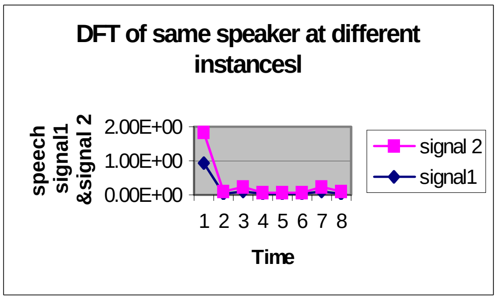

Fig. 3 DFT of same speaker at different instances

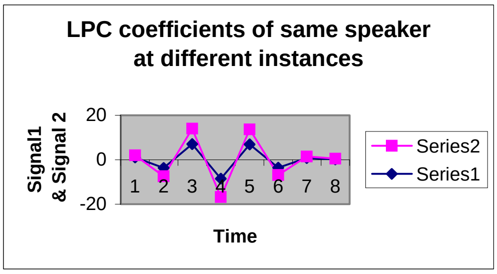

Fig. 4 LPC coefficients of same speaker at different instances

VI. Conclusion

The results obtained show that the deviation in the results for the same speaker speaking at different instances is negligible. The software works fine for identifying speaker from number of different speakers.

- The vocabulary is limited
- Number of users are limited
- The software works only for '.wav 'files

Artificial Neural Network should be very well trained.

VII. Future Scope

A range of future improvements is possible:

- Speech independent speaker identification
- No of user scan be increased
- Identification of a male female child and adult

References

[1] S. kasuriya1, V. Achariyakulporn, C. Wutiwiwatchai, C. Tanprasert, Text-dependent speaker identification via telephone based on dtw and mlp. 22nd floor, Gypsum-Metropolitan Building, Sri-Ayudhaya Rd.,Rachathewi, Bangkok 10400, Thailand

[2] Monte, J. Hernando,X.Miró,A. Adolf Dpt.TSC.Universitat Politécnica de Catalunya Barcelona.Spain ,Text independent speaker identification on noisy environments by means of self organizing maps Dpt. TSC. Universitat Politécnica de Catalunya, Barcelona, Spain.

[3] A.N. Iyer, B. Y. Smolenski, R. E. Yantorno J. Cupples, S. Wenndt, Speaker identification improvement using the usable speech concept. Speech Processing Lab, Temple University 12th & Norris Streets, Philadelphia, PA 19122,Air Force Research Laboratory/IFEC, 32 Brooks Rd. Rome NY 13441-4514.

[4] Lawrence Rabiner- "Fundamentals of Speech Recognition" Pearson Education Speech Processing Series. Pearson Education Publication.

[5] Lawrence Rabiner- "Digital Processing of Speech Signals" Pearson Education Speech Processing Series. Pearson Education Publication.

[6] Picton P.H.- "Introduction to Neural Networks", Mc Graw Hill Publication.

[7] Ben Gold & Nelson Morgan –"Speech and Audio Signal Processing." John Wiley and Sons.

[8] Duane Hanselman & Bruce Littlefield-"mastering Matlab-A comprehensive Tutorial & reference" Prentice Hall International Editions

[9] Proakis Manolakis "Digital Signal Processing Principles, Algorithms and applications " Prentice –Hall India

[10] John G Proakis & Vinay K Ingle-"Digital Signal Processing Using malab "Thomson Brooks/cole

[11] Jacek .M.Zurada "Introduction to artificial Neural Systems"

[12] http/www.electronicsletters.com

[13] http/www.DspGuru.com

[14] http/www.mathworks.com

[15] http/www.bores.com

[16] www.ieee.org/discover

R.V Pawar Working presently with Vishwakarma institute of technology has an experience of working as I/C Head Of Electronics Department at AISSM Women's College Of Engineering.,PuneShe has also worked for Indpro Electronics Pvt. LTD ,Pune and has handled projects for IFFCO, Usha Ispat,Reliance,and various state Electric city Boards.

P.P.Kajave: Presently working with Mahindra British Telecom, as a Training Manager has previously worked for Govt. College Of Engineering Pune. Has a vast knowledge in the field of compiler construction. He is an author of various books in the same field.

S.N. Mali: Presently working as Head Of Computer Department Vishwakarma Institute of Technology has a vast experience in the field of DSP, Microprocessors and Embedded system. He has published paper at various conferences in the same field at national level. He has done his Masters (ME Electronics- Computer) from Walchand college of Engineering, Sangli and is presently pursuing his Ph.D. from Pune University