Similarity Match (SM) Technique for the Development of Client Barcode

Sheikh Hussain Shaikh Salleh

Fakulti Kejuruteraan Elektrik
Universiti Teknologi Malaysia
81310 Skudai, Johor, Malaysia
e-mail: hussain@suria.fke.utm.my

Abstract: A hybrid neural network is proposed for speaker verification (SV). The basic idea in this system is the usage of vector quantization preprocessing as the feature extractor. The experiments were carried out using a neural network model (NNM) with frame labeling performed from a client codebook known as NNM-C. The work also examines how the Neural Network Model with enhance features from the client barcode compares to NNM client codebook with Linear Time Normalization (LTN). Improved performance for NNM (client barcode) with more inputs and proper alignment of the speech signals supports the hypothesis that a more detailed representation of the speech patterns proved helpful for the system. The flexibility of this system allows an equal error rate (EER) of 0.62% (Speaker Specific EER) on a single isolated digit and 1.9% (SI EER) on a sequence of 12 isolated digits.

Keywords
Speaker verification, vector quantization, neural network.

I. INTRODUCTION

It is generally accepted that humans can identify a person from the sound of his voice and yet two different voices sound alike. The variation in voices has made automatic speech recognition difficult while their similarities provide problem for the speaker recognition system [1]. Speaker Verification - the main concern of this paper involved the use of correlation scores as an added information to train the neural network to classify between the client and the impostor. For example, [2]Itoh studied the correlation between speech signal power and pitch frequency for twenty main languages. The short-term correlation is analyzed to study the relationship between the high correlation values and other characteristics of the speech signals. Other related work by Itch includes the study of correlation between speech signal power and pitch frequency for Japanese. These results indicate that there is strong relationship between speech power and pitch frequency and that it can be used in speech processing system for all spoken languages. Zack and Thomas describe a neural network built around a new objective function correlation + scaling error (COSE). The neural network is trained for articulatory speech recognition with its application to vowel identification. there are two processing stage for the system. In the initial stage the recorded speech acoustics are mapped onto the speakers articulatory movement. This mapping is accomplished by Neural Network. In the second stage heuristic classification algorithms are applied to the articulatory data to rectify the gestures associated with particular vowels or consonant. Rumelhart[3] defined the error term based on the difference between the desired output are the target value and the actual output. The actual output is the result of feed forward calculation. The error for a given input pattern is the summation of the difference between the target and the output over all the output nodes for that pattern. This square error obtained is used to propagate backwards for proper adjustment of weights. Zack and Thomas also used an alternative means f measuring neural network error such as the correlation(r) values while training to maximize r produces the proper relative timing and relative magnitude of the articulator movements. Their study involves evaluating a function, which combines the desirable characteristics of square error and r. Comparison with the traditional neural network shows that COSE trains faster and produces an output, which better represents the shape of the articulatory movements and yields higher recognition rates for vowel gestures.

However, one of the simplest approaches to overcome the variability of the word is the linear time normalization (LTN). LTN is made to correspond as closely as possible to a straight line joining the initial and final points. The approach of using LTN is implemented for the verification system [4]. The problem with LTN is that phonetic events (especially short time events) such as the plosive can be discarded during the process or insertion of feature vector may alter its relative duration. This is more important than the steady state information in the vowel. Care must be taken in selecting the proper LTN values, which preserves the nonlinearity of the speech signals. It is of interest to study the effects of different values of LTN (with respect to

0-7803-6355-8/00/ $10.00 © 2000 IEEE

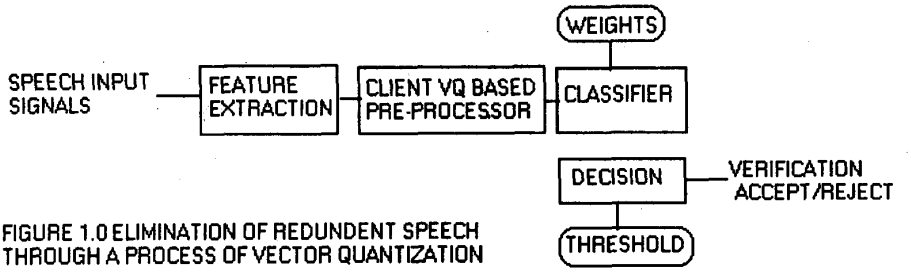

FIGURE 1.0 ELIMINATION OF REDUNDENT SPEECH THROUGH A PROCESS OF VECTOR QUANTIZATION

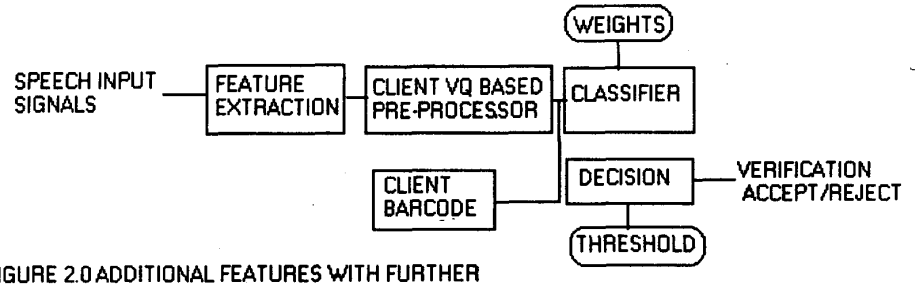

FIGURE 2.0 ADDITIONAL FEATURES WITH FURTHER EMPHASIS ON SIMILARITY/DISSIMILARITY BETWEEN THE CLIENT AND THE REST OF THE POPULATION.

information being discarded or inserted) and compared to the method described in this present paper to the speaker verification performance.

II. SPEAKER VERIFICATION METHOD

Oglesby and Mason [5] proposed a text dependent speaker verification (SV) using one MLP per speaker. These approaches used raw features to feed into the neural network. The larger and more complex the input space the more training samples are needed for training before the network can learn to generalize. There is also the possibility that large number of hidden nodes are required to solve the problem. If this is the case then training may be difficult as not only will the MLP takes a long time to learn but it will also increase the chance to get trapped in a local minimum, which may not yield a good solution to the problem. In the method described in the present paper, vector quantization is used in a preprocessing stage to reduce the number of input features. A self organization network is combined with the LBG technique to design the vector quantizer. Once the codebook is generated, the preprocessing stage uses the vector quantizer to select the index. The indices of the winner nodes are fed to a neural network classifier in which the system can be trained and evaluated. The use of a preprocessing stage allows a smaller network configuration. This can eliminate the difficulties in the training phase and facilitates training on limited data.

SPEAKER VERIFICATION (PRE-PROCESSOR)

The verification system is shown in Figures 1.0 and 2.0. The initial stage used the commonly used feature set, cepstral coefficients for the speaker modeling stage as the speech signals. For each frame of the input speech, the output of the preprocessor would contain the index j of the codevector with the minimal distortion and the corresponding distortion value d. The input pattern is linear time normalized (LTN) either by linear compression or expansion so that the total number of frames becomes a constant regardless of the word duration. Through the preprocessing stage, the highly redundant speech data are reduced so that only the useful information regarding codevector and the distance measure is retained in the feature vector to feed the MLP. For example, if the number of frames after LTN=40, two coefficients per frame will fit the 80 input units. By using the client codebook each hidden unit is fed with 80 input units resulting in architecture of 80-N-1 where N is the number of hidden units. The classifier system is based on a three layer perceptron trained using the back-propagation algorithm. The training scheme used a separate net for each digit for each speaker. Separate nets were trained for each of the 12 digits for each of the 11 speakers.

SPEAKER VERIFICATION with CLIENT BARCODE

A speaker recognition system using word recognition was developed based on the similarity match(SM). During the

recognition process, the SM values are time continuously computed for word recognition through pattern matching between an input vector and reference pattern vectors. An end point candidate t_j is assumed for each analysis frame. Using the maximum and minimum duration ($d_{\{max\}}$, $d_{\{min\}}$), a series of start point candidates (t_i, t_{i+1} ..... t_n) are determined corresponding to the end point t_j for reference word *l*. The SM is applied to obtain the maximum similarity value for that word *l* at t_j. The process is repeated for all the vocabulary words and if the maximum similarity value is found then this value is used to train the neural network for classification. It was found that the SM technique used is a powerful tool to aid the learning process of neural network. A similarity technique is developed in this paper for speaker verification systems. The focus of the work is to determining intra-speaker variation based on this similarity match. Experimental results and system implementation are given to show the effectiveness of the proposed method.

description of developing the client barcode as well as a new approach to SV system design is given in Figure 3.0. The approach has been tested with the same database used in the previous experiments and has been shown to produce improvements in the overall performance.

III. SPEECH DATA

The database consists of the isolated digits from a large number of speakers. Twelve isolated digits (digits 'one' to 'nine' plus 'zero','nought' and 'oh') were used in the experiments. A group of 11 speakers are modeled by the system and an independent set of 83 impostor speakers is used for testing. The data are all end-point detected to remove excess silence and minimize storage requirements. The framesize was 20ms with 15ms overlap. The training templates consisted of 5 tokens from the client speaker and

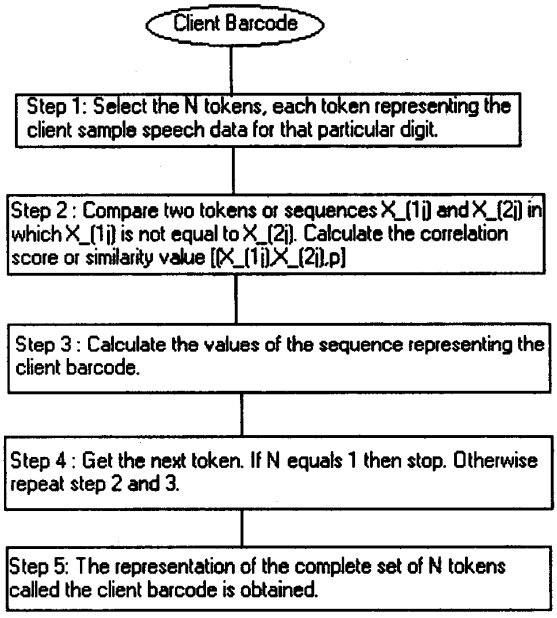

FIGURE 3.0 DEVELOPMENT OF CLIENT BARCODE

Thus the initial stage for the new approach of the SV system starts through a process of generating a barcode base on the technique mentioned above. The barcode is obtained from client training tokens. The training or the test utterances of the preprocessed speech signals are matched with the client

barcode resulting in a correlation score. This can be easily implemented. The new score acts as added information to the existing data index (j) and the minimum distortion value (d) used as an input to train the neural network. A detailed

19 from the impostors (different from the impostors used in testing). The templates from the target group and the impostor group were alternated in the training set. The implemented verification system used another set of data (not used during training) for further evaluation of its performance. It was tested on 20 true speaker tokens and 83 impostor tokens for each digit for each speaker. In the evaluation of the verification system the use of equal error rate (EER) thresholds means that all thresholds are determined a posteriori. This approach sets the proportion of

false acceptance equal to the proportion of false rejection resulting in the said EER.

IV. RESULTS and DISCUSSION

RESULTS

Digit Sequence Performance by Individual Clients

This section looks in more detail at the errors produced by different values of LTN on the 12 digit sequence using speaker specific Equal Error Rate (EER). The performance of the NNM-C model can then be assessed and the best model for each of the client speakers can be determined. The breakdown of errors by client for the 12 digit sequence is shown in Table 1.0. The table shows the EERs for the different LTN values used by client speakers. The best EER over all values of LTN NNM-C is entered in the column BLTN while the worst result is shown in the column WL'IN. The MEAN column has the average EER over the four LTN NNM-C for each of the clients. The best average EER is obtained with LTN60. NNM-C trained with this architecture is error free for 6/11 (54%) of the clients. From the BLTN column, if proper selection of LTN values is made for each of the client speaker then only 4(36%) of the clients have errors. LTN30 has the worst average EER. NNM-C trained with this architecture has 8 clients with errors. Table 1.0 also shows independence between the different values of LTN used and close examination of the table reveals more examples of such independence. Client 2, 9 and 10 had no errors with different LTN values used. Client 5 had errors with LTN30 and LTN40 and no errors with LTN50 and LTN60. Client 6 had no errors with only LTN40. It can be seen that the different values of LTN used in the system provide significant improvement for the different clients. Proper selection of the L'IN values proved an advantage to the technique, which compensate the time scale variation of the words that differ among the client speakers. In order to take advantage of this fact, perform a closed test on the training data using each of the LTN values. The LTN value that is most useful to the client can be selected and used in the NNM-C for that client. The difficulty of determeaning the LTN values is the limited amount of data available. Testing on the training data will not give an accurate estimate of the likely LTN. Extra data could be obtained during enrolment that could be set aside for proper selection of LTN. If the selection of the L'IN values could be done without too much increase of the enrolment data, it can be beneficial to the design of the SV system. In order to improve the initial model trained on a small amount of training data, model adaptation over time will be an essential part of the SY system. The extra data obtained during this enrolment process could also be used for proper selection of LTN.

In the following section, the potential of NNM-CM for ASV is explored directly using experimental procedures mentioned earlier. The normal practice reported in the literature for measurement of system performance is the EER. The speaker acceptance or rejection decision in this section is carried out by comparing the results of the new approach to the previous work of the NNM SV systems. Both speaker specific (SS) and speaker independent (SI) thresholds were used to evaluate the SV system. These thresholds are determined by the EER criterion. The use of EER in both experiments provides a standard set of measurements that detail the performance of SV systems. In this experiment the normalization procedure made use of LTN60 for the evaluation of the NNM-CM SV system.

DISCUSSIONS

The approach of having a fixed input to the NNM-C is one of the simplest methods of time aligning in a linear fashion of the speech signals. One advantage of this approach is that it does give proper alignment of the beginning and the end of the patterns. Improved performance for NNM-C with more inputs and proper alignment of the speech signals supports the hypothesis that a more detailed representation of the speech patterns proved helpful for the system. This paper established the relative performance of the different LTN values used in the experiments. It also suggests the possibility of selecting the best LTN values in order to improve the robustness of the NNM-C model.

The main concern of this paper is the use of correlation scores as added information to train the neural network to classify between the client and the impostor. Speech patterns from the same client speaker should be similar when matched although speech signals of each utterance will not be exactly the same. In the work carried out here the same practical strategy of cross match was employed to evaluate local similarities within an analyzing frames of the speech signals. The aim is to compare the client speech patterns in order to group them into corresponding samples. Having found a sequence representing the group it seems sensible to attempt to form a 'barcode' representing the client.

The performance results of NNM SV systems are likely to be affected by a number of factors. Firstly, the use of multiple information sources obtained from different types of codebooks. Secondly, the size and network complexity is under the constraint of limited training data. Thirdly, the capability of the preprocessor in handling the temporal structure of the speech signals. Finally, the input representation to the neural network varies and this is dependent on the type of preprocessor used for the SV system. These preprocessors contain vital information of the speaker

V. REFERENCES

[1] Luck, J.E. Automatic speaker verification using cepstral measurement. Pages 1026-1032 of: Journal of the acoustic of America, vol. 46.

[2] Hussain, S. An evaluation of preprocessors for neural network speaker verification. Ph.D. thesis, University of Edinburgh, 1997.

[3] Rumelhart, D.E., & McClelland, J.L. Parallel distributed Processing, Exploration in the Microstructure of Cognition, vol. 1. Foundation MIT press.

[4] Hussain, S. Comparison of neural network techniques for speaker verification. In Proceedings of the Sixth Australian International Conference on Speech Science and Technology. Adelaide, December 1996.

[5] Oglesby & Mason., Optimization of neural network for speaker identification. Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. Pages 261-264 of Proceedings of the IEEE International conference on Acoustic Speech and Signal Processing, vol. 1.

	Speaker Specific equal error rate (%) L T N						
CLIENT	30	40	50	60	WLTN	BLTN	MEAN
1	0.6	0	0	0	0.6	0	0.15
2	0	0	0	0	0	0	0
3	0.6	0.6	1.2	0.6	1.2	0.6	0.75
4	4.2	1.8	5.0	1.2	5.0	1.2	.05
5	0.6	0.6	0	0	0.6	0	0.3
6	0.6	0	1.8	2.5	2.5	0	1.22
7	0.6	0	0.6	0	0.6	0	0.3
8	8.5	5.5	6.1	1.8	8.5	1.8	5.47
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	3.6	3.0	3.6	1.2	3.6	1.2	2.85
Mean	1.75	1.04	1.70	0.7	2.05	0.43	1.27

Table 1.0: Comparison of Error Rates on 12 Digit Sequences between Eleven Client Speakers with a range of LTN NNM-C