SPEAKER VERIFICATION USING SELF SEGMENTING LINEAR PREDICTORS

E. Ambikairajah*, M. Keane** and G. Tattersall***

* Speech Research Group, Regional Technical College, Athlone, Ireland.
** D.S.P. Research Unit, Dept. of Electronic Engineering, University College, Galway, Ireland.
*** School of Information Systems, University of East Anglia, England.

Abstract

A new speaker verification model is proposed in this paper. The model uses self aligning linear predictors to represent the temporal structures of speech. Conventional Linear Predictive Coefficient (LPC) methods use short frames for analysis, resulting in neighbouring frames having very similar coefficients. This paper proposes a model that uses variable length segments. This considerably reduces the number of coefficients required to represent the true speaker utterance. Furthermore, the fact that these segments self align in the time domain eliminates the need for time warping. The training algorithm is based on a combination of dynamic programming and steepest decent techniques. The self segmenting model was evaluated on a database of 70 utterances, taken over a period of three weeks. Two different criteria were used in the verification decision, one was based on the accumulated prediction residual and the other was based on the optimal segmentation of the utterance. Individually these two tests yielded verification accuracies of 97% and 95%. An accuracy of 100% was achieved when both the accumulated prediction residual and the optimal segmentation were used in the verification decision.

INTRODUCTION

The task of a speaker verification system is to ascertain whether the speaker being tested is who he/she claims to be. Usually a test utterance is spoken by the claimant and the speaker verification system compares the features of this spoken utterance with the previously stored template for the claimed true speaker. Verification models currently in use, such as Multi-layer Perceptrons or cluster analysis, require speech feature vectors as inputs (Iso & Watnabee 1990). For these methods pre-processing is required to generate the speech feature vectors.

Speaker verification using non-linear predictors has also been implemented (Ambikairajah & Kelly 1992). The model uses non-linear neural predictors to predict future speech feature vectors, using two previous feature vectors as inputs. A verification accuracy of 100% was achieved with this model on the same database that is used here.

Linear predictors have previously proven successful in speaker verification where the speech was divided into short analysis frames of equal length. Due to the slow time varying nature of speech, the vocal tract shape is considered to stay constant during the duration of an analysis frame. Hence the speech signal for that frame can be considered a stationary signal and be adequately represented by the Linear Predictive Coefficients (LPC). In many cases the vocal tract stays constant for longer than one analysis frame. Sounds such as "ee", "aah" and "sss" are examples of steady state behaviour over a number of analysis frames (Markel & Gray 1976). These long interval sounds can adequately be represented by a longer frame length. In many cases the vocal tract changes within an analysis frame and the resulting LPC coefficients can not be considered accurate.

The iterative process, described in this paper, trains to a set of LPCs for each speech sub-unit during which the vocal tract stays constant. The associated analysis frame used in conventional LPC analysis is replaced with a speech segment temporally aligned along the speech sub-unit best represented by the LPCs. These segments are of variable length and are not related to the conventional analysis frames. The number of these segments used for a given utterance is much less than the number of frames in conventional analysis. The model produces a set of optimised LPCs by firstly eliminating the repetition present in conventional LPC analysis during long interval sounds, and secondly, aligning along speech sub-units, hence solving the problem of within frame vocal tract changes. This model proposed here uses time domain samples as inputs and requires no pre-processing.

THE SELF SEGMENTING LPC MODEL

The model for speaker verification is composed of a series of M, p^{th} order Linear Predictors and a state transition network, where each state in the network has a linear predictor associated with it (Figure 1). This configuration is similar to a Hidden Markov Model with state skipping not allowed. Transitions from left to right in this network represents transitions through the speech sub-units of the test utterance. A prediction of the test utterance at each sample is generated as a linear combination of p previous samples, where the n^{th} predicted sample is given by

$$\hat{s}_n = \sum_{k=1}^{p} s_{n-k} a_{k,m}$$

where s_{n-k} is the k^{th} previous samples, $a_{k,m}$ is the p^{th} LPC associated with the m^{th} state and p=12.

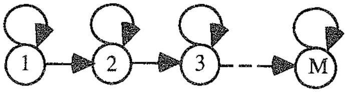

Figure 1: Self Segmenting LPC Model with M states

VERIFICATION ALGORITHM

This section presents the verification algorithm for the self segmenting LPC model. The model for the prediction of the test utterance by an M state self segmenting LPC model requires that the test utterance be divided into M segments such that in state m the model makes a prediction for the m^{th} segment. The optimal segmentation of the utterance is determined by the accumulated prediction residual, D,

$$D = \min \sum_{n=1}^{N} [s_n - \hat{s}_{n,m}]^2$$

where N is the total number of samples in the utterance, s_n is the n^{th} sample and $s_{n,m}$ is the sample predicted by the m^{th} state, where m must be equal to 1 for the start of the utterance and follow the state diagram through the M states. Under these constraints the minimisation can be accomplished using the Dynamic Programming (DP) recursion formula. Figure 2 illustrates a plane visualising the DP computation where the horizontal axis represents time and the vertical axis the model state.

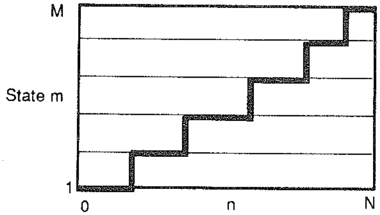

Figure 2 : Prediction Residual Minimisation by DP

The DP recursion formula is given by

$$g(n,m) = d(n,m) + \min\{g(n-1,m), g(n-1,m-1)\}$$

where the local distance measure d (n,m) is defined as

$$d(n,m) = (s_n - \hat{s}_{n,m})^2$$

At the end of recursive application of DP, backtracking of the optimal trajectory provides the utterance segmentation.

The accumulated prediction residual is normalised to give the Mean Squared Prediction Error (MSPE),V, by dividing D, the accumulated prediction residual by the sum of the squared utterance samples over the length of the utterance

$$V = \frac{D}{\sum_{n=1}^{N} s_n^2}$$

The MSPE is used as a test statistic in the speaker verification decision by comparing it with a pre-set threshold .

The segmentation table resulting from the backtracking is considered to reflect the temporal composition of the utterance. It was noted that for true speakers, the utterance is segmented into approximately equal segments, reflecting the fact that the consecutive predictors in the M state verification model accurately predicted the test utterance, as would be expected. However, in the case where the test utterance was that of an impostor the segmentation pattern is considerably different, usually with one segment dominating the utterance (figure 3). A 3-layer Multi-Layer Perceptron (MLP) was trained with the true speaker segmentation pattern and was used as a further test in the speaker verification decision.

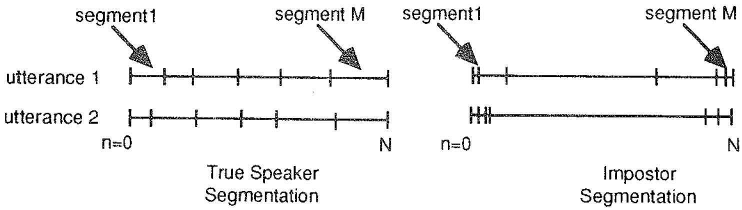

Figure 3 : Optimal segmentation patterns for two true speakers and two impostors utterances

TRAINING ALGORITHM

The training goal is to minimise the total accumulated prediction residual across all the utterances in the training set. This optimisation can be achieved by an iterative procedure combining dynamic programming and LPC coefficient update.

Each of the p LPC coefficients a_k are updated at each sample, n, according to

$$a_k(n+1) = a_k(n) - \mu \nabla e(n)^2$$

where k=1 to p for the p LPCs associated with each segment, μ is the learning rate and

$$\nabla e(n)^2 = -2e(n)s(n-k)$$

Each LPC set is updated only across the segment of the utterance with which it is associated. This reduces the local prediction error for this predictor across this segment and, hence, the global prediction error. For the first iteration of the training process the utterance is divided in to M equal segments and the LPCs are updated using the above equation. The coefficient update is repeated 10 times for each utterance in the training set. This is done because it is necessary to have the LPCs semi-trained before the first segmentation of the utterances using DP, otherwise the segmentation does not reflect the stationary sections of the speech utterance.

Thus, the training algorithm is as follows:

```
1. Initialise all LPC coefficients to zero
   Divide each training utterance into M equal segments
   Semi-train each LPC to it's associated utterance segment

2. Set m=1 {m is the current state in segmentation model}

3. Compute the accumulated prediction residual, D, using DP and determine the optimal
   trajectory using it's backtracking for each utterance in the training set

4. Update m th LPC, using Steepest Decent Algorithm, across associated utterance
   segment from backtracking for every utterance in the training set

5. Increase m by 1

6. Repeat 3-5, while m<M+1

7. Stages 2-6 comprise a single training epoch
   Repeat training epochs until prediction residual stops decreasing
```

At the end of the training process the normalised MSPE, for each of the training utterances is used to set a threshold for that utterance. The optimal segmentation of each of the utterances is stored for use in training a 3 layer MLP to recognise the true speaker segmentation pattern.

SPEAKER DATABASE

For the purpose of developing and testing this model a small database of speakers repeating the same utterance was taken. This consisted of a total of five different speakers all repeating the utterance "Six..three..nine". These utterances were spoken in a laboratory environment with a high quality 16-bit speech acquisition system. Before the sampling process, the speech was pre-emphasised and bandpass filtered from 70 Hz to 3.4 kHz. The true speaker repeated the utterance a total of 50 times over the period of a few weeks and the four impostors each repeated the utterance a total of five times each over the same time period. All of these utterance were spoken in a normal conversational tone which often resulted in co-articulation. The complete utterances were each manually endpointed. For the purpose of these experiments, each of the three words within the utterances were roughly endpointed, to give three separate sets of single word utterances. The model was trained for each word separately.

THE MLP CLASSIFIER

An MLP based classifier is used in one of the experiments to classify a test utterance as a true speaker or an impostor. A 3-layer MLP with 9 hidden nodes in each hidden layer and 1 output was used. The number of inputs to the MLP depends on the experiment being carried out.

When applying only the segmentation pattern as input data, there are M entries in the segmentation table and hence M inputs to the MLP are required. When applying the segmentation pattern and the MSPE, M+3 inputs are required, M for the segmentation and 3 for the MSPE. The MSPE is applied to 3 inputs in order to increase its effect on the verification decision. Experimental results showed that repeating this input 3 times had the desired effect on the output of the MLP by accepting true speaker utterances and rejecting impostor utterances.

The MLP is trained using the usual Error Back Propagation algorithm (Rummelhart et al., 1987). Only the utterances in the training set are used in the training phase. Testing involves the application of the test utterance input data, be it segmentation pattern or segmentation pattern plus MSPE to the MLP for verification decision.

EXPERIMENT AND RESULTS

The system which has been described in this paper has been tested with the above database of speakers. Of the 50 true speaker utterances available 40 were used in training and 10 were used as an unseen test set. The impostor set was split into a set of 12 for training and a set of 8 for testing. The self segmenting model was trained with each of the three sets of 40 true speaker utterances separately, according to the training algorithm. The total prediction error for each utterance is shown in figure 4.

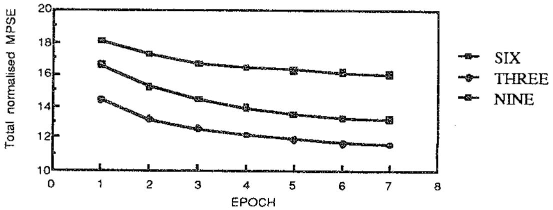

Figure 4: Total MPSE for 40 utterances vs. training iteration

Experiment 1

The MSPE was calculated for both test and training sets of each utterance. The MSPEs for associated utterances were added in each case. For example the MSPEs for the words comprising the first utterance, six1, three1 and nine1 were added to give the total MSPE for utterance 1. Using the total MSPEs for true speakers and impostors in the training set, the optimal threshold for verification was set.

The total MSPEs for utterances in both the training set and the test set are then each compared to this threshold and a verification decision is made. Two types of verification error can occur. Type 1, rejecting a true speaker and Type 2, accepting an impostor. The results for this experiment are summarised in table 1.

	Training set	Test set
True Speakers	0	1
Impostors	0	1

Table 1: Errors for verification using MPSE as test statistic

Experiment 2

The segmentation pattern of each utterance was normalised by dividing each segment length by the utterance length. The normalised utterances in the training set were then used to train a 3-layer MLP with nine nodes in each hidden layer and M input nodes. After the training process the utterances from both training and test sets were applied to the MLP. The results for experiment 2 are summarised in table 2

	Training set	Test set
True Speakers	0	1
Impostors	1	1

Table 2 : Errors for verification using normalised optimal segmentation as input to MLP

Experiment 3

The MSPE and the associated optimal segmentation were used together as inputs to the MLP and the MLP was retrained with the training sets. The number of nodes in the two hidden layers was again nine and the number of input nodes was equal to M+3, where the M normalised segments were applied to M inputs and the MSPE applied to the other 3 inputs. After the training process both training and test sets were applied to the MLP and a verification decision was made for each utterance. The results for this experiment are summarised in table 3.

	Training set	Test set
True Speakers	0	0
Impostors	0	0

Table 3 : Errors for verification using segmentation and MPSE as input to MLP

DISCUSSION AND CONCLUSION

The verification model proposed here has been tested on a small database containing three sets of single word utterances. Of the two test criteria used, the normalised MSPE test results of 97% accuracy were slightly better than the segmentation pattern results of 95%. When both MSPE and segmentation pattern were used as inputs to the MLP a verification accuracy of 100% was achieved.

The results of experiment 1 indicate that the accumulated prediction based measure, the MSPE is not itself enough to make an accurate verification decision. Note however that the verification errors were made only in the previously unseen test set, and that varying the threshold decision level could change the type of errors yielded in this method. For example lowering the threshold would result in less false impostor acceptances but more true speaker rejections. Neither does the criteria used in experiment 2 guarantee an accurate decision, but considering that only temporal alignment information was used in this decision test, the results seem promising. Finally the combination of these two tests, in experiment 3 gives accurate verification on this database in both training and test sets. The test set, which was not used in the training process gives a clear indication of the ability of the system to classify previously unseen utterances. It is hoped to carry out further experiments on this model with a larger database of utterances.

ACKNOWLEDGEMENTS

This research has been funded by EOLAS, the Irish Science and Technology Agency, under the 1990 Applied Research Award.

REFERENCES

Ambikairajah, E. & Kelly, A. (1992) *A speaker verification system based on a neural prediction model,* Proceedings of ISITA '92 Conference, Singapore 16-20 November, 1992.

Iso, K. & Watanabe, T. (1990) *Speaker independent word recognition using a Neural Prediction Model,* CH2847-2/90/0000-0441, IEEE 1990.

Markel, J.D. & Gray , A. H. (1976) *Linear prediction of speech,* (Springer Verlag: Berlin)

Rummelhart, D. et al. (1987) *Learning internal representations by error propagation,* Parallel Distributed Processing, D. Rummelhart and McClelland, J.L. (Eds), MIT Press 1987, 318-362