SPEAKER RECOGNITION WITH RECURRENT NEURAL NETWORKS

Shahla Parveen[1,2], Abdul Qadeer[2] and PhilGreen[1]
1 Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Sheffield S14DP, UK
2 Department of Applied Physics, University of Karachi, University Road, Karachi-75270, Pakistan
(s.parveen,p.green@dcs.shef.ac.uk)

ABSTRACT

We report on the application of recurrent neural nets in a open-set text-dependent speaker identification task. The motivation for applying recurrent neural nets to this domain is to find out if their ability to take short-term spectral features but yet respond to long-term temporal events is advantageous for speaker identification.

We use a feedforward net architecture adapted from that introduced by Robinson et.al. We introduce a fully-connected hidden layer between the input and state nodes and the output. We show that this hidden layer makes the learning of complex classification tasks more efficient. Training uses back propagation through time. There is one output unit per speaker, with the training targets corresponding to speaker identity.

For 12 speakers (a mixture of male and female) we obtain a true acceptance rate 100% with a false acceptance rate 4%. For 16 speakers these figures are 94% and 7% respectively.

We also investigate the sensitivity of identification accuracy to environmental factors (signal level, change of microphone and band limitation), choice of acoustic vectors (FFT, LPC or Cepstral), distribution of speakers in the training database, inclusion of fundamental frequency. FFT features plus fundamental frequency give the best results.

This performance is shown to compare favorably with studies reported on similar tasks with Hidden Markov Model technique.

1.INTRODUCTION

In automatic *speaker recognition* the problem is to classify speakers based upon their speech acoustics. Speaker recognition research subdivides into *speaker identification* (who is speaking?) and *speaker verification* (is the speaker who s/he claims to be?). In turn, speaker identification may be *closed-set* (the speaker is known to be one of those included in the training database) or *open-set* (the speaker may be outside the training corpus). All these problems have text-dependent, text-independent and text-prompted variants [Furui, 1994], [Furui, 1997].

A number of techniques have been applied to speaker recognition, including hidden Markov models [Siohan, 1998], vector quantisation [Qin et al., 1998], gaussian mixture modelling [Reynolds, 1995], multi-layer perceptrons [Altosaar and Meister, 1995], Radial Basis Functions [Finan et al., 1996] and genetic algorithms [Hannah et al., 1993]. In this paper we investigate an alternative technique: Recurrent Neural Networks. Our task is open-set speaker identification task in text-dependent mode: this is more useful than close-set identification in applications such as criminal investigations [Qin et al., 1998].

The motivation for applying recurrent neural nets to this domain is to take advantage of their ability to process short-term spectral features but yet respond to long-term temporal events. Previous research has confirmed that speaker recognition performance improves as the duration of utterance is increased [He and Liu, 1999]. In addition, it has been shown that in identification problems RNNs may confer a better performance and learn in a shorter time than conventional feedforward networks [Gingras & Bengio, 1998].

We report on a study in which RNNS were trained and tested for some 70 different conditions using a small speaker identification database.

2.RNN ARCHITECTURE

We use a feedforward net architecture adapted from that introduced by Robinson [1994]. In his work, the output layer takes connections both directly from the input layer and from a delayed version of an internal state vector. We introduce a fully-connected hidden layer between the input and states and the output (Fig 1). In section 4.1 below we demonstrate improved performance with this hidden layer.

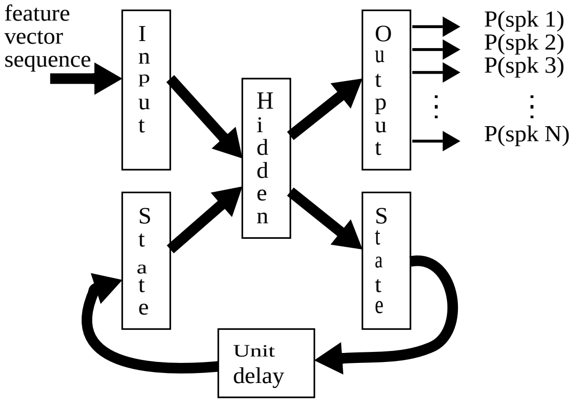

Fig 1. RNN architecture for speaker identification

The input layer has 22 nodes (20 mel scaled fft coefficients, total energy and fundamental frequency). There is one output unit per speaker, with the training targets corresponding to speaker identity. A similar strategy was used by Altosaar and Meister [1995].

In the output layer we use a normalised exponential function to guarantee that the outputs sum to one, whereas input, hidden and state layer unit activation is calculated with a sigmoidal function.Training uses back propagation through time. During forward propagation the average output of all frames for each utterance is calculated and then compared with the utterance target (1 for the correct speaker and 0 for all others). RNN weights are updated if the error expressed by this comparison is greater than a training error threshold, θ . Training continues until the output error for all utterances is less than θ .

During test sessions we calculate the average output probability of each speaker for all frames in each utterance. The highest scoring speaker is accepted provided there is sufficient contrast compared to the next highest:

$$\frac{(P(max) - P(nextmax))}{(P(max) + P(nextmax))} > 0.4$$

3.DATABASE AND EXPERIMENTAL SETUP

The database of 30 speakers (15 males, 15 females) was recorded in a laboratory environment (average signal to noise ratio 22 dB) and was sampled at 11 kHz. Each speaker was asked to utter the same phrase *"Assalam-o-Alaikum"* six times. Three examples of the phrase from each speaker were used for training and the remaining three were used for testing. Impostors for each testing condition were selected from the speakers which were not included in training for that condition.

Signal processing steps for training and recognition were: start and end point detection; channel equalisation; time-domain signal normalisation and scaling. Spectral analysis was performed at intervals of 46 msec frames with a 92 msec window duration (50% overlap). Most experiments used spectral energy in 20 mel-scaled FFT bins. For comparison purposes (section 4.6) we also performed cepstral analysis and linear predictive (LP) analysis. In addition we added total power in the frame and fundamental frequency to the acoustic vector. For F0 estimation we used Hermes Sub-harmonic Spectrum method [Hermes, 1988]. The drawback of this method is that it makes no distinction between voiced and unvoiced frames, for which it tends to return a value outside the normal F0 range. We therefore replaced these values by initialising them to the standard adult geometric mean pitch (168Hz) and updating this estimate by taking a moving average with the F0 estimate in voiced frames. This smoothing improved the convergence rate of the RNNs.

4.RESULTS

In all our experiments we were able to achieve perfect classification on training data. Here, we report results in terms of test set error rates.

4.1 Introduction of a hidden layer

We varied the number of state units and hidden units in the RNN architecture and examined the effect on convergence speed and classification error rate. Typical configurations were (26 state units, 30 hidden units), (26, 42), (26, 50), (30, 42) and (30, 50). We found (26, 42) to perform best for most of the tests.

Table 1 compares RNN performance (as percentages of true and false speaker acceptance) with and without the hidden layer for a 12-speaker open-set identification task. The total number of training and test utterances were 36 and 60 respectively. In Table 1, the first row corresponds to the RNN with 22 input, 26 states 42 hidden and 12 output units. The remaining rows show results for RNNs without a hidden layer at different values of θ and with varying numbers of state units. The RNN with a hidden layer outperforms RNNs without hidden layers.

TABLE 1: Identification results for RNNs with and without hidden layer

RNN Topology	θ	No. of Epochs	True Accept %	False Accept %
22-26-42-12	0.001	261	100	4.0
22-26-12	0.18	121	77.7	25.0
22-26-12	0.20	105	75.0	25.0
22-26-12	0.25	115	77.7	20.8
22-40-12	0.3	62	80.5	45.8

4.2 Varying the number of speakers

Table 2 gives speaker recognition performance for a net with topology 22-26-42-n where n is the number of speakers. The best result with this topology is for n=12. It seems that the optimal number of hidden units is dependent on n - more units are required for more speakers. For 20 speaker we trained the RNN with 42 and 60 hidden units respectively. Increasing the number of hidden units improved true acceptance rate from 81.7% to 88.3%.

TABLE 2: Effect of number of speakers on identification performance. The terms in parentheses indicate the total numbers of test examples

No. of Speakers	θ	No. of Epochs	True Accept%	False Accept%
8	0.001	267	100 (24)	14.3 (21)
12	0.01	178	100 (36)	4.1 (24)
16	0.01	131	93.7 (48)	7.1 (42)
20	0.1	95	81.7 (60)	16.6 (30)

4.3 Distribution of Speakers

The distribution of speakers in the training database has significant effect on recognition performance. A task in which the database consists of evenly distributed male and female speakers is easier to train and gives higher performance than the case where the database consists of speakers of the same sex, as shown in table 3.

TABLE 3: Effect of distribution of speakers in the training database.

Experiment	No. of Epochs	True Accept%	False Accept%
10M	304	90.0	16.6
10F	350	83.3	16.6
5M+5F	261	96.6	20.0

4.4 Frequency Cut Off

With a view to telephone applications, we studied the effect of band limitation on speaker identification performance. In this experimented we trained the system for 12 speakers. We trained RNN on 36 utterances and tested with a further 36 utterances from valid speakers and 24 from imposters. It is clear from table 4 that recognition accuracy decreases with the narrowing of frequency-band, because there is speaker-dependent voice quality information in the higher frequency region [Yoshida et al., 1999], [Hayakawa and Itakura, 1995].

TABLE 4: Effect of band limitation on identification performance.

Frequency range Hz	No. of Epochs	True Accept%	False Accept%
60-5512	261	100	4.1
60-4000	458	83.3	4.1
60-3000	320	63.6	12.5

4.5 Signal Distortions

Speaker identification performance strongly depends on the quality of the input signal. Distortions which are caused by not using a fixed microphone, changing the microphone in training and recognition and amplitude clipping (recording gain too high) degrade speaker identification performance as shown in table 5. We performed this experiment on a database of 16 speakers with 48 training utterance, some of them distorted. The recognition system was tested on 48 valid utterances and 42 from impostors. For the second line in table 5 the distorted material was replaced by undistorted examples.

TABLE 5: Effect of signal level on identification performance.

Experiment	No. of Epochs	True Accept%	False Accept%
distorted utterances	261	79.1	7.1
undistorted utterances	354	87.5	9.5

4.6 Changing the Acoustic Vector

Inclusion of fundamental frequency in the acoustic vector improves identification performance as shown in table 6. The choice of spectral analysis technique is also important in speaker identification. Our experiments show that cepstal analysis gives improved performance compared to spectral or LPC analysis. Cepstral coefficients with F0 do not give improved performance over FFT coefficients with pitch as given in table 7.

TABLE 6: Effect of inclusion of fundamental frequency on identification performance.

Experiment	No. of Epochs	True Accept%	False Accept%
FFT spect+F0	261	100	4.1
FFT spect	354	86.1	4.1

TABLE 7: Effect of choice of acoustic vectors on identification performance.

Experiment	No. of Epoch	True Accept%	False Accept%
FFT spect	354	86.1	4.1
FFT cepst	249	88.9	8.33
LPC20	366	72.2	12.0
cepst+F0	254	91.7	8.33

5.CONCLUSION

In this paper we have discussed the implementation of RNNs for a small speaker identification task and reported the effect of various variations on identification performance. An RNN with (26, 42) topology performs well on a small (8-16 speakers) open-set speaker identification task. There were no misclassification errors (assignment of identitity of one registered speaker to another) which compares favourably with a similar small closed-set identification task with HMM [Siohan et al.,1998]. In their work average misclassification error for 70 different groups of 10 speakers is reported to be 3.17%.

REFERENCES

[1] He, J. and Liu, L. (1999). Speaker Verification performance and the length of test sentence. Proceedings ICASSP 1999 vol. 1, pp. 305-308.

[2] Gingras, F. & Bengio, Y. (1998). Handling Asynchronous or Missing Data with Recurrent Networks. International Journal of Computational Intelligence and Organizations. vol. 1, no. 3, pp. pp. 154-163, 1998.

[3] Robinson, A. J. (1994). The application of recurrent nets to phone probability estimation. IEEE Transactions on Neural Networks, vol.5 no.2, March 1994.

[4] Siohan, O., Rosenberg, A. and Parthasarathy, S. (1998) Speaker Identification Using Minimum Identification Error Training. Proceedings ICASSP 1998, vol. 1, pp. 109-112.

[5] Altosaar, T. and Meister, E. (1995) Speaker Recognition in Estonian Using Multi-Layer Feed-Forward Neural Nets. Proceedings EUROSPEECH 1995, vol. 1, pp. 333-336.

[6] Qin, J., Luo, S., and Qixiu Hu. (1998) A high performance text-independent speaker identification system based on BCDM. Proceedings ICSLP, vol. 2, pp. 133-136.

[7] Hermes, D.J. (1988). Measurement of pitch by subharmonic summation, Journal of the Acoustical Society of America, vol. 84, pp. 257-264.

[8] Furui, S., (1994). An overview of Speaker Recognition Technology. In Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pages 1-9, 1994.

[9] Furui, S. (1997). Recent advances in speaker recognition. Proc. First International Conference on Audio- and Video-based Biometric Person Authentication, Crans-Montana, Switzerland, March, 1997.

[9] Finan, R.A., Sapeluk, A.T. and Damper, R.I. (1996). Comparison of multilayer and radial basis function neural networks for text-dependent speaker recognition. In Proc. Int. Conf. Neural Networks (ICNN'96), volume 4,pages 1992--1997, Washington, DC. IEEE.

[10] Hannah, M.I., Sapeluk, A.T., Damper, R.I. and Roger, I.M. (1993). Using genetic algorithms to improve speaker-verifier performance. In Proc. Joint Workshop on Natural Algorithms in Signal Processing, volume 2, pages 24/1--24/9, Chelmsford, UK. IEE/IEEE.

[11] Yoshida, K., Takagi, K. and Ozeki, K. (1999). Speaker Identification using Subband HMMs. Proceedings EURSPEECH'99, vol. 2, pp. 1019-1022.

[12] Hayakawa, S. and Itakura, F. (1995). Speaker recognition using speaker individual information in the higher frequency band. The Journal of the Acoustical Society of Japan, vol. 51, pp. 861-868.

[13] Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, vol. 17, pp. 91-108, March 1995.