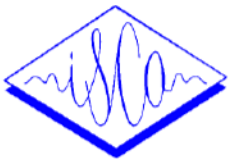

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

Evaluation of Speaker-Independent Phoneme Recognition on TIMIT Database Using TDNNs

Nobuo Hataoka*
Hitachi Dublin Laboratory
Trinity College, Dublin, Ireland

Alexander H. Waibel
School of Computer Science
Carnegie Mellon University, Pittsburgh, USA

Abstract

This paper describes evaluation results and a new structure of Time-Delay Neural Networks (TDNN) for speaker-independent and context-independent phoneme recognition. The proposed new structure is based on the integration of TDNNs which have several TDNNs separated according to the duration of phonemes, so that it deals with phonemes of varying duration more effectively. In the experimental evaluation of the proposed new structure, 16-English vowel recognition was performed using 5268 vowel tokens picked from 480 sentences spoken by 140 speakers (98 males and 42 females) on the TIMIT (TI-MIT) database. A 60.5% recognition rate, which was improved from 56% in the single TDNN structure, and stability improvement of recognition rate showed the effectiveness of the proposed integrated TDNNs.

1. INTRODUCTION

Recently, quite a few efforts have been made to develop speech recognition systems using promising connectionist models (Lippmann *et al.*[1], Waibel *et al.*[2], [3], Leung *et al.*[4], Bourlard *et al.*[5], Franzini *et al.*[6]). This is due to the fact that Neural Networks may have the ability to overcome limitations of conventional techniques in speech recognition. Speech recognition is one of the excellent abilities of human beings. So, new approaches, which are based on human cognitive mechanisms, should be explored to further advance this field. Neural Networks (NNs), whose basic idea is motivated by processing mechanisms of the nervous system, may be a good scheme for pattern recognition, including speech recognition.

However, current structures of NNs must be improved to better cope with the temporal nature of speech. Especially, usual NNs show poor performance in the case of speech features which are quite similar, and where the duration information might be the only cue in distinguishing this speech, such as single vowels and diphthongs. To overcome these problems, variable duration input patterns should be used in order to minimize training and improve generalization in the case of short phonemes (single vowels etc.) and to provide enough input information in the case of long phonemes (diphthongs etc.).

In this paper, firstly we evaluate TDNN ability for speaker-independent and context-independent phoneme recognition from the view points of speech parameters, TDNN input window length, and the number of hidden units. Finally, we propose an integrated TDNNs structure which has several TDNNs separated according to the duration of phonemes. As a result, the proposed structure has the advantage of dealing with varying duration information more effectively. Experimental evaluation of the proposed new structure was performed using 16 English vowels picked from continuously uttered sentences in the TIMIT (Lee *et al.*[7]) database.

2. SPEAKER-INDEPENDENT PHONEME RECOGNITION USING TDNN

2.1 A Brief View of the System

First, sentence length speech, which has been labelled at the phoneme level, is analysed and transferred to speech feature coefficients. We are using an FFT analysis method. (At the parameter evaluation stage, a cepstral analysis method for NNs has been evaluated to compare with an FFT method.) Subsequently, speech intervals, which have vowel parts of a sentence, are picked up using labelling information. In the training mode of NNs, training patterns are used to obtain weighted values of the connections between units in the TDNN. And in the testing mode, other test patterns are used for evaluation of the NNs which have these weighted values. These training and testing modes are carried out by a speaker-independent and context-independent recognition method. This means the patterns, which are used in each mode, are picked from completely different speakers and sentence contents.

The training and testing modes are executed by "DyNet" (Haffner [8]), a software package for the fast training of Neural Nets. The learning algorithm of DyNet is based on the Error Back-Propagation (Backprop, Rumelhart *et al.*[9]), though DyNet is using an optimised search strategy and is controlling the "step size" and the "momentum" of NNs' parameters dynamically. As a result, DyNet can get very fast convergence.

* The author was a visiting researcher at CMU from Central Research Laboratory, Hitachi, Ltd., Japan. This work has been done on a collaborative research project between the Centre for Machine Translation of CMU and Hitachi, Ltd.

2.2. Experimental Conditions

1. TIMIT Database

We use the TIMIT speech database in this research. This is because the TIMIT database has so many and various speakers and sentences that this database is most suitable in evaluating speaker-independent speech recognition performance. Moreover, comparison with other speaker-independent speech recognition systems, which are using the same TIMIT database[1] (e.g. SPHINX system (Lee *et al.*[7]) and NN system (Leung *et al.*[4])), will be possible and effective for the evaluation of our proposed system. We selected a task of 16-English vowel recognition.

These 16 English vowels are /ae/(bat), /eh/(bet), /ih/(bit), /iy/(beat), /uh/(book), /ah/(butt), /ax/(the), /ix/(roses), /aa/(cot), /ao/(about), /uw/(boot), /aw/(bough), /ay/(bite), /ey/(bait), /ow/(boat), and /oy/(boy).

2. Training and Test Samples

We carried out the experiments according to the following two phases which are separated from the amount of sample size used. These samples were selected at random from the speech data in the TIMIT database.

(1) Preliminary Experiments (Small Samples)

This data was used for the comparison of speech analysis methods (FFT and cepstral analysis) and the length of input data to decide which would be better for the main experiments. The data size was as follows:

50 speakers, 135 sentences
- Training Samples: 1139 vowel patterns from 35 speakers
- Test Samples: 430 vowel patterns from 15 speakers

(2) Main Experiments (Large Samples)

Main experiments are carried out by the following data:

140 speakers, 480 sentences
- Training Samples: 4326 vowels from 100 speakers (69 males, 31 females)
- Test Samples: 942 vowel patterns from 40 speakers (28 males, 12 females)

3. Speech Processing

The speech input, which was sampled at 16 kHz and pre-emphasized with a filter (transfer function $1-0.97z^{-1}$), was hamming windowed and 256-point FFT coefficients were computed every 5 msec. And then, the 16 melscaled coefficients of the power spectrum were obtained by the melscaled transformation from these 256-point FFT coefficients. Finally, 16 coefficients of 10 msec frame rate were obtained by the average of two adjacent coefficients in time. The coefficients of an input token were then normalized to have the values between -1.0 to +1.0 with the average of 0.0.

[1]In the TIMIT database, 630 speakers uttered five 'sx', three 'si', and two 'sa' sentences. The 'sa' is not used in this research because of its fixed context.

3. EXPERIMENTS USING SINGLE TDNN

3.1 TDNN Architecture

The TDNN structure has been created to cope with many problems, which are substantial in the speech recognition field. And the TDNN has been shown to be powerful, especially for Japanese phonemes, such as */b/*, */d/*, */g/* in speaker-dependent speech recognition tasks. The TDNN consists of four layers, including input and output layers.

The connections between each layer used in this research are completely the same as in the previous report [2]. The differences are an addition of a power coefficient to 16 FFT coefficients and 16 outputs in the output layer.

First, we evaluate the performance of speech coefficients (FFT vs. cepstral) and the duration length of input sample (150 msec vs. 200 msec).

3.2 Experimental Results

1. Preliminary Experiments Using Small Samples

(1) Parameters: FFT vs. Cepstral Coefficients[2]

Fig. 1 shows recognition rates for training samples and test samples according to learning times (epochs). The maximum rates for test data are found within 100 epochs and the rates over 100 epochs did not increase. This is because of overlearning and/or generalisation problem. Table 1 shows comparison results (maximum rates).

From these results, we found that FFT coefficients showed a slightly improved performance, especially in view of overlearning and generalisation problems. In this research, we have decided to use the FFT coefficients from this preliminary comparison. However, this comparisonwas done on small samples, so we need further evaluation before reaching a final conclusion.

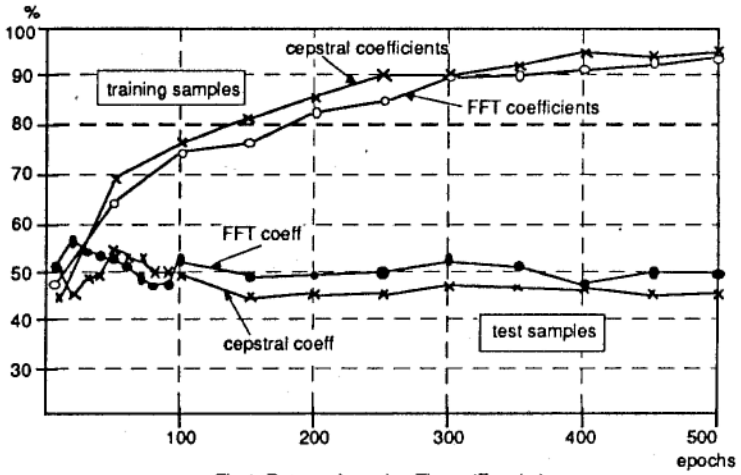

Fig.1 Rate vs. Learning Times (Epochs)

[2]TDNN structures: In the case of FFT, the total number of units is 509 including a bias unit, i.e. 16 input coefficients without power and 16 vertical units in the first hidden layer. In the case of cepstral coefficients, the total number of units is 759 including a bias unit, i.e. 26 input coefficients (including 12 differential cepstral coefficients and one differential power) and 16 vertical units in the first hidden layer.

Table 1 Comparison between FFT vs. Cepstral Coeff.

DATA	FFT Coeff.	Cepstral Coeff.
training data (1139 patterns)	63.8% (50th epoch) 93.5% (500th)	69.1% (50th) 95.1% (500th)
testing Data (430 patterns)	55.6% (20th) 50.0% (500th)	55.4% (50th) 46.6% (500th)

Table 3 Comparison of the Number of Units in the 1st Hidden Layer

DATA	the number of units in the 1st hidden layer			
	16	20	24	28
training data (4326 patterns)	59.8% (150th epoch)	60.6% (150th epoch)	63.9% (150th epoch)	65.0% (150th epoch)
testing data (942 patterns)	54.1% (70th)	55.5% (30th)	57.3% (30th)	54.9% (30th)

(2) Input Window Length: 150msec vs. 200msec

Table 2 shows comparison result. The input sample of 150 msec has produced better results than that of 200 msec. We can imagine that the 200 msec data is including a lot of unnecessary neighbour vowels and consonants, especially in short duration vowels such as /ax/ and /ix/. As a result, the generalisation for these short vowels is so poor that the decreased performance of these short vowels is affecting the total performance.

Table 2 Comparison of Input Window Length (150msec vs. 200msec)

DATA	150 msec (15 frames)	200msec (20 frames)
training data (1139 patterns)	63.8% (50th epoch) 93.5% (500th)	59.5% (50th) 90.8% (500th)
testing Data (430 patterns)	55.6% (20th) 50.0% (500th)	48.8% (50th) 41.5% (500th)

2. Experiments Using Large Samples

(1) the number of units in the first hidden layer

The number of the vertical units were evaluated using large samples. Table 3 shows recognition results. The recognition rates are maximum ones within 150 epochs. The case of 24 vertical units showed the best performance. Generalisation problem might have occurred in the case of 28 vertical units.

3.3 Consideration on Single TDNN

The experimental results of the single TDNN show the following problems:

(1) errors between single vowels and diphthongs (e.g. /ax/ and /ay/, /ih/ and /ey/ etc.)
(2) necessary to use more input information for diphthongs
(3) generalisation problems, especially for short duration vowels

4. NEW STRUCTURE OF INTEGRATED TDNNs

4.1 Integrated TDNNs

Fig. 2 shows the proposed structure based on the integration of TDNNs. The various intervals of speech are put into each TDNN's input layer in the first NNs. The outputs of first NNs are put into the second NNs' input layer. Each TDNN has an output for the counter category and the training procedure of these NNs is carried out separately. These Integrated TDNNs can manage the duration difference between each vowel, especially between single vowels and diphthongs, because the input da' can be separated by the duration difference, by putting the data into the different TDNN-n in a training mode. As a result, each TDNN-n can share recognition abilities for specified phonemes. Each TDNN can be designed to cover at least sum of duration average and standard deviation of assigned vowels.

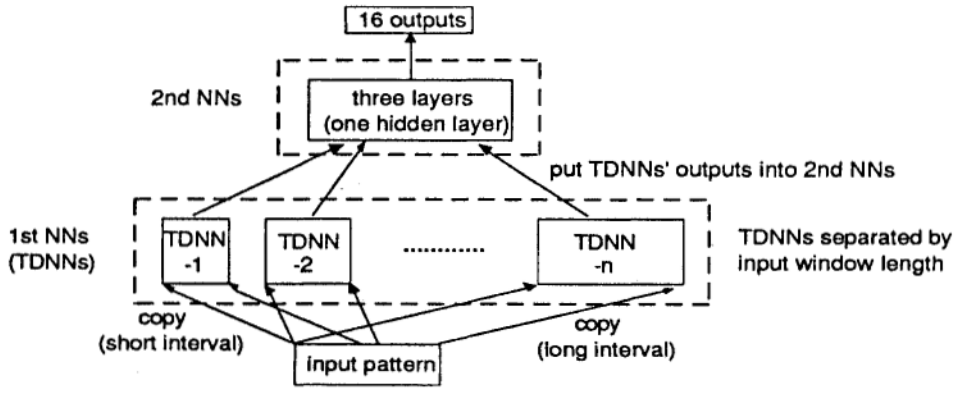

Fig.2 Structure of Integrated TDNNs

4.2 Evaluation Results

Currently, two TDNNs and three TDNNs are being used which are distinguished by the duration difference between each vowel, especially between single vowels and diphthongs. In two TDNNs[3], the TDNNs for single vowels and diphthongs have 150 msec and 200 msec input intervals, respectively. In three TDNNs[4], three TDNNs have 100 msec, 150msec, and 200 msec, respectively.

Table 4 shows evaluation results. Rate A is by small samples (50 speakers, 135 sentences) and rate B by large samples (140 speakers, 480 sentences). These results indicate the performance increase according to the increase of the number of TDNNs. Especially, stability of recognition rate within 50 epochs was improved by the integrated TDNNs.

Table 4 Evaluation Results of Integrated TDNNs

structure	rate A (small samples)	rate B (large samples)
Single TDNN	56.1%	57.3% (53.7%)
Integrated TDNNs (two TDNNs)	60.5%	57.8% (57.4%)
Integrated TDNNs (three TDNNs)	-	59.3% (58.7%)

maximum rates within 150 epochs
Rates in the ( ) are average from 10 data within 50 epochs.

5. DISCUSSIONS AND FUTURE WORKS

The evaluation of the Integrated TDNNs shows the performance increase by separated TDNNs. The reasons why the performance has been increased are that the generalisation might become better for short duration vowels, and that sufficient information can be supplied for long duration vowels such as diphthongs.

We obtained around 70% recognition rate (69.1% for small samples) for a collapsed 13-vowel set using the integrated TDNNs trained context independently. Lee and Hon reported context-independent recognition rate of 53.68% and context-dependent of 65.71% for all sonorants which include the collapsed 13-vowel set [7]. Leung and Zue used artificial NNs for the same 16-vowel task, and reported 54% for context-independent recognition and 67% for context-dependent [4].

The future work will be (1) Increase the number of TDNNs, (2) Use of context information, (3) Models for sequential processing, and (4) Hierarchical and feedback type NNs using semantic and syntactic information.

[3]Two TDNNs: separated by the group of single vowels and diphthongs, TDNN-a is for the single vowel group (10 categories; /ae/, /eh/, /ih/, /iy/, /uh/, /ah/, /ax/, /ix/, /aa/, and a counter group) and TDNN-b is for the diphthong group (8 categories; /ao/, /uw/, /aw/, /ay/, /ey/, /ow/, /oy/, and a counter group).

[4]Three TDNNs: separated by duration information, TDNN-x is for the group of 4 categories (/ax/, /ix/, and two counter categories). TDNN-a is for the group of 8 categories (/eh/, /ih/, /iy/, /uh/, /ah/, /uw/, and two counter categories). TDNN-b is for the group of 10 categories (/ae/, /aa/, /ao/, /aw/, /ay/, /ey/, /ow/, /oy/, and two counter categories).

6. CONCLUSION

In this paper, we evaluated the ability of Neural Networks in speaker-independent and context-independent speech recognition on an English database (TIMIT database). And we proposed a new NNs structure (Integrated TDNNs) which can cope with the duration difference problem among vowels and can use the duration information effectively.

In the experimental evaluation of the proposed structure, 16-English vowel recognition was performed using 5268 vowel tokens picked from 480 sentences spoken by 140 speakers (98 males and 42 females) on the TIMIT database. The number of training tokens and testing tokens was 4326 from 100 speakers (69 males and 31 females) and 942 from 40 speakers (29 males and 11 females), respectively. The result on testing data was around 60% recognition rate (around 70% for a collapsed 13-vowel case), which was improved from 56% in the single TDNN structure, showing the effectiveness of the proposed new structure in using temporal information.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Kai-Fu Lee of Computer Science Department at CMU for fruitful discussions which helped us perform this research effectively.

REFERENCES

[1] Lippmann, R.P. *et al.*, *Neural Net Classifier Useful for Speech Recognition*, IEEE ICNN-87, June 1987

[2] Waibel, A. *et al.*, *Phoneme Recognition Using Time-Delay Neural Networks*, IEEE Trans. on ASSP, Vol.37, No.3, March 1989

[3] Waibel, A. *et al.*, *Modularity and Scaling in Large Phonemic Neural Networks*, IEEE Trans. on ASSP, Vol.37, No.12, December 1989

[4] Leung, H. *et al.*, *Some Phonetic Recognition Experiments Using Artificial Neural Nets*, IEEE ICASSP-88, April 1988

[5] Bourlard, H. *et al.*, *Speech Dynamics and Recurrent Neural Networks*, IEEE ICASSP-89, May 1989

[6] Franzini, M. *et al.*, *A Connectionist Approach to Continuous Speech Recognition*, IEEE ICASSP-89, May 1989

[7] Lee, K. *et al.*, *Speaker-Independent Phone Recognition Using Hidden Markov Models*, CMU-CS-88-121,Mar 1988

[8] Haffner, P., *DyNet, a Fast Program for Learning in Neural Networks*, ATR Report TR-I-0059, Nov. 1988

[9] Rumelhart, D.E. *et al.*, *Parallel Distributed Processing; Explorations in the Microstructure of Cognition*, Volume I, MIT Press, Cambridge, MA, 1986