# Speaker-Independent Phoneme Recognition on TIMIT Database Using Integrated Time-Delay Neural Networks (TDNNs)

Nobuo Hataoka [1] and Alex H. Waibel

Hitachi Dublin Laboratory
Trinity College
Dublin 2, Ireland

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.

**Abstract**

This paper describes a new structure of Neural Networks (NNs) for speaker-independent and context-independent phoneme recognition. This structure is based on the integration of Time-Delay Neural Networks (TDNN) which have several TDNNs separated according to the duration of phonemes. As a result, the proposed structure has the advantage 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. 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 was a 60.5% 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 to use temporal information.

## 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. From this point of view, 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 in the field of speech recognition. 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 in the case of single vowels and diphthongs. To overcome these problems in phoneme recognition, variable duration input patterns should be used in order to minimize training and improve generalization in the case of short phonemes (e.g. single vowels) and to provide enough input information in the case of long phonemes (e.g. diphthongs).

In this paper, we propose a new algorithm for NNs which is quite useful for speaker-independent and context-independent phoneme recognition. This structure is based on the integration of Time-Delay Neural Nets (TDNN, Waibel *et al.*[2], [3]) which have 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(TI-MIT, Lee *et al.*[7]) database. We report here on the details of the algorithm and experimental evaluation results for speaker-independent and context-independent phoneme recognition.

## 2 SPEAKER-INDEPENDENT PHONEME RECOGNITION USING TDNN

### 2.1 A Brief View of the System

First, sentence length speech, which has been labeled at the phoneme level, is analyzed and transferred to speech feature coefficients. We are using an FFT analysis method. (Moreover, a cepstral analysis method for NNs has been

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

evaluated in preliminary experiments to compare with an FFT method.) Subsequently, speech intervals, which have vowel parts of a sentence, are picked up using labeling information. We use only the beginning information of vowels. In other words, some portion of speech from this beginning is being used in the training and testing (recognition) modes. 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 testing 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 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 optimized search atrategy and is controlling the "step size" and the "momentum" of NNs' parameters dynamically. As a result, DyNet can get very fast convergence.

### 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[2] (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 Testing 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.

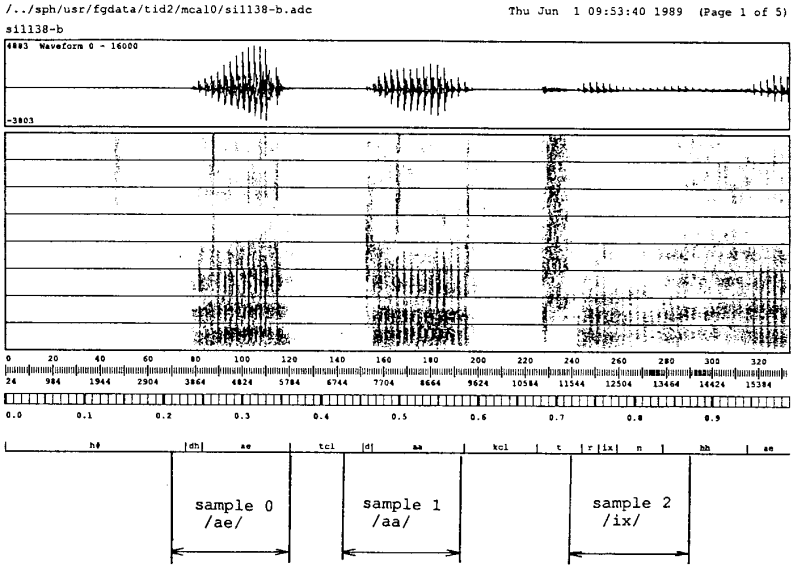


Fig. 1 An Example of Input Speech

[2]The TIMIT database consists of speech data uttered by 630 speakers. Each speaker uttered the following sentences.
- Five "sx" sentences which were read from a list of 450 phonetically balanced sentences selected by MIT,
- Three "si" sentences which were randomly selected by TI,
- Two "sa" sentences which are the same over all speakers.

The "sa" sentences were not used in this research because the evaluation of the context-independent recognition performance using NNs is one of the main purposes of this research. So, the "sa" sentence set is not applicable for this research because of its fixed context.

(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
- Testing 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 vowel patterns from 100 speakers (69 males, 31 females)
- Testing 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. Fig.1 shows an example of speech samples picked from a continuously uttered speech. This speech is the beginning portion of a sentence whose content is "That doctrine has been accepted by many, but has it produced good results?"

## 3. PRELIMINARY 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. As a result, the input layer has 255 input units (15 frames x 17 coefficients). The numbers of the first hidden layer and the second hidden layer are 208 units (13 horizontal and 16 vertical units) and 144 units (9 horizontal and 16 vertical), respectively.

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) Comparison of Parameters: FFT vs. Cepstral Coefficients

Table 1 shows comparison result[3]. From this result, we found that FFT coefficients showed a slightly improved performance, especially in view of overlearning and generalization problems. However, this comparison was done on small samples, so we need further evaluation before reaching a final conclusion. In this research, we have decided to use the FFT coefficients from this preliminary comparison.

Table 1 Comparison between FFT vs. Cepstral Coefficients

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

[3]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.

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

Table 2 shows comparison result[4]. 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 neighbor vowels and consonants, especially in short duration vowels such as /ax/ and /ix/, and as a result, the generalization 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)<br>93.5% (500th) | 59.5% (50th)<br>90.8% (500th) |
| testing Data (430 patterns) | 55.6% (20th)<br>50.0% (500th) | 48.8% (50th)<br>41.5% (500th) |

2. Experiments Using Large Samples

(1) Comparison of 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 case of 24 vertical units showed the best performance. Overlearning and generalization problems might have occurred in the case of 28 vertical units.

Table 3 Comparison of the Number of Units in the First Hidden Layer

| DATA | the number of units in 1st hidden layer | | | |
|---|---|---|---|---|
| | 16 | 20 | 24 | 28 |
| training data (4326 patterns) | 59.80% (150th epoch) | 60.61% (150th epoch) | 63.92% (150th epoch) | 64.98% (150th epoch) |
| testing Data (942 patterns) | 54.14% (70th) | 55.52% (30th) | 57.32% (30th) | 54.88% (30th) |

### 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 /ai/, /ix/ and /ai/ etc.)
---> how to use duration information explicitly
---> overlapped-category problem (/ax/</ai/, /ix/</ai/)

(2) necessary to use more input information for diphthongs
Quite a few diphthongs have duration length over 200msec.

(3) generalization 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 data 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.

---

[4]TDNN structures: Both cases have same 16 input units without power,and the same 16 vertical units in the first hidden layer.

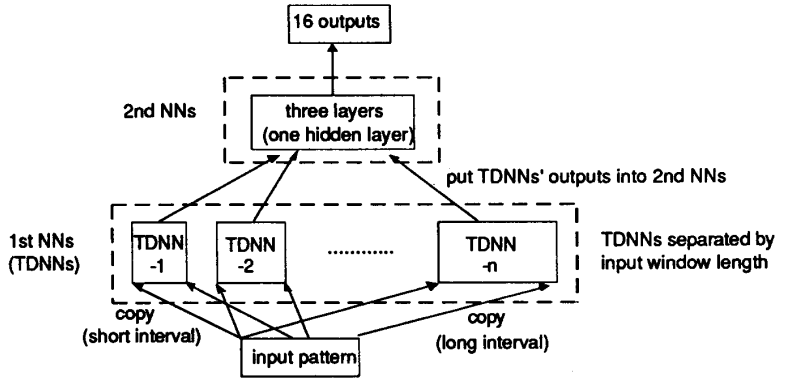


Fig.2 Structure of Integrated TDNNs

### 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[5], the TDNNs for single vowels and diphthongs have 150 msec and 200 msec input intervals, respectively. In three TDNNs[6], three TDNNs have 100 msec, 150msec, and 200 msec, respectively.

Table 4 shows results of preliminary experiments using small samples (50 speakers, 135 sentences) and Table 5 shows recognition results using large samples (140 speakers, 480 sentences). These results indicate the performance increase according to the increase of the number of TDNNs.

(1) Preliminary Experiments Using Small Samples

Table 4 Comparison between Single TDNN and Integrated TDNNs

| structure | rate | comments |
|---|---|---|
| Single TDNN | 56.05% | |
| Integrated TDNNs (two TDNNs) | 60.47% | TDNN-a 62.56% (10 categories)<br>TDNN-b 83.49% ( 8 categories) |

(2)Experiments Using Large Samples

Table 5 Comparison between Single TDNN and Integrated TDNNs

| structure | rate | comments |
|---|---|---|
| Single TDNN | 57.32% | |
| Integrated TDNNs (two TDNNs) | 57.75% | TDNN-a 62.00% (10 categories)<br>TDNN-b 82.38% ( 8 categories) |
| Integrated TDNNs (three TDNNs) | 59.34% | TDNN-x 71.23% ( 4 categories)<br>TDNN-a 71.87% ( 8 categories)<br>TDNN-b 66.14% (10 categories) |

[5]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).

[6]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).

## 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 generalization 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 as follows:

(1) Increase the number of TDNNs: In this report, we are using only two and three TDNNs. The extension to highly separated TDNNs can be possible to obtain better recogniton results.

(2) Use of context information:

(3) Models for sequential processing: This is the most important future work. NN classifiers may not be powerful enough to deal with pattern sequences, and sequential constraints or links with other techniques, such as HMMs or DTW, must be considered in order to advance in the speech recognition field.

(4) Hierarchical and feedback type NNs using semantic and syntactic information: How to use higher level information such as semantics and syntax on the NNs is one of our major tasks.

## 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, and thank Professors Jaime Carbonell and Masaru Tomita, and Business Manager Radha Rao of the Center for Machine Translation at CMU for setting up the research environments. Finally, I greatly appreciate Ms. Marilyn Jones-Bernick's kindness in proofreading this paper.

## REFERENCES

[1] Lippmann, R.P. and Gold, B (1987) *Neural Net Classifier Useful for Speech Recognition*, in IEEE ICNN-87, June 1987

[2] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989) *Phoneme Recognition Using Time-Delay Neural Networks*, in IEEE Trans. on ASSP, Vol.37, No.3, March 1989

[3] Waibel, A., Sawai, H., and Shikano, K. (1988) *Modularity and Scaling in Large Phonemic Neural Networks*, in IEEE Trans. on ASSP, Vol.37, No.12, December 1989

[4] Leung, H. and Zue, V. (1988) *Some Phonetic Recognition Experiments Using Artificial Neural Nets*, in Proceedings of the IEEE ICASSP-88, April 1988

[5] Bourlard, H. and Wellekens, C. (1989) *Speech Dynamics and Recurrent Neural Networks*,in Proceedings of the IEEE ICASSP-89, May 1989

[6] Franzini, M., Witbrock, M., and Lee, K. (1989) *A Connectionist Approach to Continuous Speech Recognition*, in Proceedings of the IEEE ICASSP-89, May 1989

[7] Lee, K. and Hon, H.W. (1988) *Speaker-Independent Phone Recognition Using Hidden Markov Models*, CMU-CS-88-121, March 1988

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

[9] Rumelhart, D.E. and McClelland, J.L. (1986) *Parallel Distributed Processing; Explorations in the Microstructure of Cognition*, Volume I, MIT Press, Cambridge, MA, 1986