First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Speech Recognition Using Sub-Phoneme Recognition Neural Network

Kiyoaki Aikawa † and Alexander H. Waibel ‡

† NTT Human Interface Laboratories
3-9-11 Midoricho, Musashino-shi, Tokyo 180 Japan
‡ Carnegie Mellon University, Pittsburgh, PA 15213 USA

ABSTRACT

This paper proposes a new phoneme-based speech recognition approach using neural networks trained to recognize sub-phonemes. The sub-phoneme is an acoustic unit which is shorter than a phoneme. The sub-phoneme recognition neural networks exhibit a more precise firing pattern and smaller firing gaps around phoneme boundaries than conventional phoneme recognition neural networks. The word or sentence score is given by the normalized highest sum of the output neuron firing score, which is obtained by the Dynamic Time Warping (DTW) algorithm. A Time Delay Neural Network (TDNN) structure is employed for the sub-phoneme recognizer. The proposed method has been evaluated through word recognition using a continuous speech database. The results show that the recognition rate greatly improves when the sub-phoneme is introduced as a recognition unit. The best word recognition rate is obtained when a phoneme period is divided into front and rear sub-phonemes. The recognition rate is further improved by introducing a multiple entry word dictionary.

INTRODUCTION

The phoneme-based recognition approach, which recognizes a word or a sentence as a sequence of phonemes, is especially effective for large vocabulary continuous speech recognition. This paper investigates phoneme-based approach using the TDNN which has recently been reported to be powerful for phoneme recognition [1, 2]. Originally, tied connection was introduced to the neural network to achieve shift-invariant phoneme recognition. The TDNN is applicable to phoneme-based speech recognition because of its firing stability.

In this paper, DTW is used to obtain a word or sentence score. In conventional DTW approaches, the neural network is trained to recognize phonemes, where, the output neural cell should keep firing throughout the phoneme period. A method using a set of shifted phoneme samples for the training has been tried to produce stable firing patterns [3]. However, the feature at the incoming transition of a phoneme and that at the outgoing transition are much different. It is not advantageous to memorize these feature variety into a neural network. To cope with this problem, the neural network would have to be very large to be able to learn all these variations. Moreover, this approach requires a large number of training samples and a large amount of computation time for training.

This paper introduces a sub-phoneme as an acoustic unit which is smaller than a phoneme. The sub-phoneme neural network is responsible only for the recognition of a part of a phoneme. Therefore, it is expected that the firing gap is reduced and that the firing accuracy is improved due to the localization of features to be memorized. This paper applies the sub-phoneme-based approach to word recognition for a continuous speech database.

TDNN

TDNN is a neural network architecture developed to recognize time-shifted phoneme samples [1]. TDNN is characterized by tied connections along the time axis in the network. A tied connection is defined as a set of neural connections which are forced to have the same connection weight. The tied connection also effectively reduces the degree of freedom while keeping a large scale network architecture.

TDNN architecture is robust for the time-shifted patterns, because a feature extracting network is repeated along the time axis. The connection between the output layer and the second hidden layer works like an analog 'OR' gate. Therefore, the output cell can collect the signal indicating the existence of phoneme feature from any position within the input window of the neural network.

In this paper, a neural cell is modeled as a multiple input one output nonlinear function unit. The nonlinear function of the cell is the standard sigmoid defined by

$$f(y_j) = \frac{1}{1 - e^{-y_j}}$$

$$y_j = bias_j + \sum_i w_{ij} x_i,$$

where x_i is the i-th input to the j-th cell. A standard feed forward type TDNN is employed. The back propagation algorithm is available for TDNN training.

The error at the j-th output cell, denoted by E_j, is measured by the square distance between the desired output T_j and the output signal $f(y_j)$ as

$$E_j = (T_j - f(y_j))^2.$$

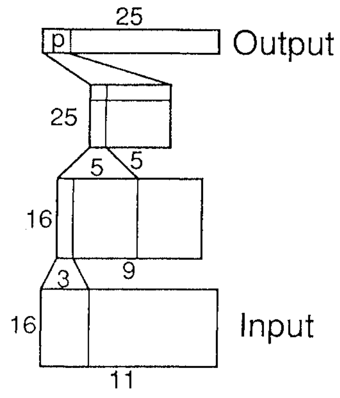

Figure 1: The TDNN for 25 sub-phoneme recognition

The connection weight w_{ij} is modified in proportion to the partial derivative as

$$\Delta w_{ij} \propto -\frac{\partial E_j}{\partial w_{ij}}.$$

Figure 1 shows the network architecture for the sub-phoneme recognizer. The input layer forms a matrix of 16 feature parameters by 11 frames. The specific input cell, which supplies a DC *bias* to every neural cell, is implicitly included in the input layer. There are thus 177 input cells in the input layer. The lower hidden layer and upper hidden layer have 16 x 9 cells and 25 x 5 cells, respectively. The full-connection unit between the 16 x 3 input cells and 16 x 1 lower hidden cells is repeated along the time axis as the tied connection. The connections between the hidden layers and between the upper hidden layer and the output layer are tied independently. The connections from the DC bias cell are also tied along the time axis.

SUB-PHONEME

A phoneme is modeled as a sequence of sub-phoneme units. A sub-phoneme could be a very small unit, such as 1/4 or less of a phoneme period. However, equal division into either two or three sub-phonemes is considered to be sufficient to represent acoustical events in a phoneme according to the three state HMM approach [4]. If a phoneme period is divided into either two or three sub-phonemes, at least the incoming transition and outgoing transition are separated into front and rear sub-phonemes, respectively.

Sub-phoneme training samples are collected according to the phoneme labels. A sub-phoneme sample is a sequence of 11 frames centered at the position,

$$center = r \;\; start + (1 - r) \;\; end,$$

where *start* is the beginning of the phoneme and *end* is the end of the phoneme as obtained from the phoneme label, and r denotes the interpolation ratio shown in Table 1. In the case of NSS=2, two samples centered at the different position in a sub-phoneme period are used for training. The NSP is equal to the number of sub-phoneme recognition neural networks used for word or sentence recognition.

Table 1: The interpolation ratio to obtain a sub-phoneme training sample

NSP	NSS	position	r
1	2	-	0.67 0.33
2	1	front	0.67
		rear	0.33
	2	front	0.8 0.6
		rear	0.4 0.2
3	1	front	0.833
		middle	0.5
		rear	0.167

NSP : Number of sub-phonemes per phoneme
NSS : Number of samples per sub-phoneme

DTW

A word or sentence score is given by the normalized highest summation of the firing score of output neurons along the input speech to reference sub-phoneme sequence matching path. The matching is performed by frame synchronous DTW. In frame synchronous DTW, firing score is summed once a frame. The word or sentence score, D, is given by

$$D = \frac{1}{M}\max_{j(i)}\{\sum_{i=1}^{M} d_{i\;j(i)}\}$$

$$1 = j(1) \leq \cdots \leq j(i-1) \leq j(i) \leq \cdots \leq j(M) = N,$$

where $j(i)$ is a function indicating a matching path and d_{ij} is the firing score of the j-th sub-phoneme output neuron at the i-th frame. M is the number of input speech frames and N is the number of sub-phonemes in a reference speech. Figure 2 illustrates the sub-phoneme based DTW.

Four types of local matching path constraints are examined. With local path constraints, the accumulated firing score of the j-th sub-phoneme at the i-th frame, denoted by g_{ij}, is given by,

Type 1-1

$$g_{i\;j} = d_{i\;j} + \begin{cases} \max\left\{\begin{array}{l} g_{i-1\;j} \\ g_{i-1\;j-1} \end{array}\right\} & j \geq 2 \\ g_{i-1\;j} & j = 1 \end{cases}$$

Type 2-1

$$g_{i\;j} = d_{i\;j} + \begin{cases} \max\left\{\begin{array}{l} g_{i-1\;j} \\ g_{i-2\;j-1} + d_{i-1\;j} \\ g_{i-2\;j-1} + d_{i-1\;j-1} \end{array}\right\} & j \geq 2 \\ g_{i-1\;j} & j = 1 \end{cases}$$

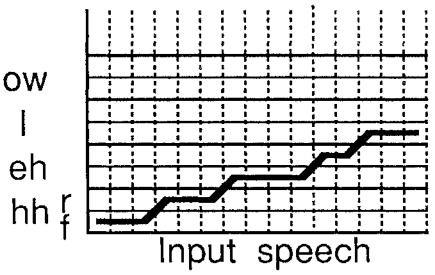

Figure 2: Sub-phoneme based DTW

Type 3-1

$$g_{i\ j} = d_{i\ j} + \begin{cases} \max\left\{ \begin{array}{l} g_{i-1\ j} \\ g_{i-3\ j-1} + d_{i-2\ j} + d_{i-1\ j} \\ g_{i-3\ j-1} + d_{i-2\ j-1} + d_{i-1\ j} \\ g_{i-3\ j-1} + d_{i-2\ j-1} + d_{i-1\ j-1} \end{array} \right\} & j \geq 2 \\ g_{i-1\ j} & j = 1 \end{cases}$$

and
Type 4-1

$$g_{i\ j} = d_{i\ j} + \begin{cases} \max\left\{ \begin{array}{l} g_{i-1\ j} \\ g_{i-4\ j-1} + g_{i-3\ j} + d_{i-2\ j} + d_{i-1\ j} \\ g_{i-4\ j-1} + g_{i-3\ j-1} + d_{i-2\ j} + d_{i-1\ j} \\ g_{i-4\ j-1} + g_{i-3\ j-1} + d_{i-2\ j-1} + d_{i-1\ j} \\ g_{i-4\ j-1} + g_{i-3\ j-1} + d_{i-2\ j-1} + d_{i-1\ j-1} \end{array} \right\} & j \geq 2 \\ g_{i-1\ j} & j = 1. \end{cases}$$

Type m-n indicates that the gradient of DTW local path on the plane spanned by input speech and sub-phoneme sequence is at most n/m. The optimal type depends on the number of sub-phonemes per phoneme. These local path constraints prevent the skipping over a reference sub-phoneme within a short time.

An adjustment window is also used to prevent excessively warped matching paths. The adjustment window at the i-th frame is given by

$$jmin \leq j \leq jmax$$

$$jmax = \frac{N}{M}i + W$$

$$jmin = \frac{N}{M}i - W$$

$$W = \begin{cases} 2 & ;\ \text{number of sub-phonemes} \leq 2 \\ 4 & ;\ \text{otherwise.} \end{cases}$$

This means that the $jmin$-th to $jmax$-th sub-phonemes are allowed to match with the i-th input frame. The local path must run within the adjustment window.

The sentence score can be obtained using the cumulative score, denoted by $g^k_{i\ j}$, based on Ney's method [5].

Type 1-1

$$g^k_{i\ j} = d^k_{i\ j} + \begin{cases} \max\left\{ \begin{array}{l} g^k_{i-1\ j} \\ g^k_{i-1\ j-1} \end{array} \right\} & j \geq 2 \\ \max\left\{ \begin{array}{l} g^k_{i-1\ j} \\ \max_{\forall l} \left\{ g^l_{i-1\ N_l} \right\} \end{array} \right\} & j = 1. \end{cases}$$

$$l = 1, 2, \cdots, N_l$$

Here, $d^k_{i\ j}$ is the firing score of the j-th sub-phoneme at the i-th input frame of the k-th reference word. N_l is the number of sub-phonemes in the l-th reference word.

EXPERIMENTS

The Conference Registration speech database including 204 sentences is used for the evaluation of the proposed method. Each sentence is spoken at natural speed by a male speaker. Neural network training samples are collected according to only the phoneme label, without taking account of the phoneme environment. Therefore, phonemes in various contexts contribute to the training. The Conference Registration database is grouped into 12 sets, C1 to C12. C1-C7 are used for training and C9 for testing. The word samples for testing are collected according to the word labels.

The sampling frequency is 16 kHz. Autocorrelation coefficients are calculated from the speech wave multiplied by the 30-ms wide Hamming window every 10 ms. Cepstral coefficients are derived from autocorrelation coefficients through Linear Predictive Coding (LPC) analysis of order 14. The LPC cepstra for the neural network input are truncated at 15. The other parameter input parameter is the delta power defined by

$$\Delta p_i = \frac{1}{2} \sum_{k=-2}^{2} \log(p_{i+k}) h_k$$

$$h_k = \begin{cases} 1 & k > 0 \\ 0 & k = 0 \\ -1 & k < 0, \end{cases}$$

where p_i is the power at the i-th frame.

The Conference Registration database is labeled with 42 phonemes and a silence symbol. These phonemes are classified into 25 for training stability when using a small database. In fact, some phonemes are rare in the speech data base. There is no linguistic confusion on test words because of this categorization. The silence label /sil/ is not used in the word recognition experiment. Table 2 shows the phoneme groups. The number of output cells in the sub-phoneme recognition neural network is thus 25.

Words which consist of more than one phoneme are chosen for the recognition experiment. The transcriptions of a reference word to phoneme sequences are written in the word dictionary. The single entry word dictionary, which has a unique transcription from a word to a phoneme symbol sequence, is used in this experiment. Table 3 shows the 50 word recognition result. "+" means that the training is continued using the latter training set after the training with the former set. Table 3 shows that the best word

Table 2: Phoneme groups

Group	Phoneme	Group	Phoneme
p	p	t	t
k	k	ch	ch
s	s sh th	f	f hh
m	m	n	n ng
b	b	d	d dx
g	g	z	z dh jh
r	r l	v	v
y	y	w	w
aa	aa ao	ax	ax ah ae er en
ih	ih iy ix	uw	uw uh
eh	eh	ay	ay
ey	ey	ow	ow oy
aw	aw		

Table 3: 50 word recognition results (test set)

Matching Path Type			1-1	2-1	3-1	4-1
NSP	Training Set	NSS				
1	C1-4	2	-	56%	62	62
	C1-4 + C5-7	2	-	72	72	74
2	C1-7	1	86	88	-	-
	C1-4	2	80	82	-	-
	C1-4 + C5-7	2	84	90	78	-
	C1-7	2	84	86	-	-
3	C1-7	1	88	70	-	-

NSP : Number of sub-phonemes per phoneme
NSS : Number of samples per sub-phoneme
C : Training set
+ : Replacement of training set

recognition rate, 90%, is obtained when a phoneme is modeled as the sequence of two sub-phonemes, with two neural networks per phoneme, while the rate is 74% in the case of one neural network per phoneme. Table 4 shows the change of the recognition rate by the number of epochs. The epoch is defined as one cycle of training for all sub-phoneme samples. Table 5 shows the error analysis for the best result.

The pronunciation of the word "the" depends on the initial phoneme of the succeeding word.

"the" = /dh ax/ for consonants
= /dh iy/ for vowels

The word recognition rate is improved to as high as 94% by introducing a multiple entry word dictionary, which allows more than one transcription from word to phoneme sequence.

CONCLUSIONS

This paper has proposed a new neural network based speech recognition approach which uses sub-phonemes as the acoustical unit. The word recognition experiments show that the recognition rate is improved by introducing sub-phoneme recognition neural networks compared with the case of using the phoneme as the acoustical unit. The optimal size of the sub-phoneme was the half of a phoneme period. Almost the same recognition accuracy was obtained using the the size of the 1/3 of a phoneme period. Further improvement was obtained by using a multiple entry word dictionary.

Table 4: The change of recognition rate by the number of epochs (training set : C5-7, after the training with C1-4)

Number of Epochs	0	50	100	150	200
Recognition Rate	80%	84	86	90	86

Table 5: Error analysis for the best word recognition result

True		Misrecognition	
Word	Score	Word	Score
CONFERENCE (k aa n f r ax n s)	0.451	OFFICE (ao f ax s)	0.455
HAVE (hh ae v)	0.529	DRIVE (d r ay v)	0.584
CALLED (k ao l d)	0.292	ALL (ao l)	0.453
LOOK (l uh k)	0.230	WHAT (w ah t)	0.406
SOON (s uw n)	0.494	SURE (sh er)	0.696

ACKNOWLEDGMENT

The authors would like to thank Sadaoki Furui and Kiyohiro Shikano for providing the opportunity to do this research at Carnegie Mellon University.

References

[1] A. H. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, Phoneme Recognition Using Time-Delay Neural Networks, IEEE Trans., ASSP Vol. 37 No. 3, pp328-339, (Mar,1989).

[2] J. B. Hampshire and A. H. Waibel, A Novel Objective Function for Improved Phoneme Recognition Using Time Delay Neural Networks, Proceedings of the 1989 International Joint Conference on Neural Networks, Vol. 1, PP.235-241, (Jun,1989).

[3] M. Miyatake et al., Integrated Training for spotting Japanese Phonemes Using Large Phonetic Time-Delay Neural Networks, ICASSP90, S8.10, pp.449-452, (Apr,1990).

[4] Kai-Fu Lee, The SPHINX speech recognition system, Ph. D. thesis (Apr,1988).

[5] Hermann Ney, The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition, IEEE Trans., ASSP Vol. 32, No.2, pp. 263-271, (Apr,1984).