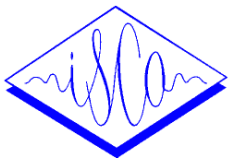

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

# The TDNN-LR Large-Vocabulary and Continuous Speech Recognition System

Hidefumi SAWAI
ATR Interpreting Telephony Research Laboratories
Inui-dani, Sanpei-dani, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan

## Abstract

This paper describes an integration of speech recognition and language processing. The speech recognition part consists of the Large Phonemic Time-Delay Neural Networks (TDNN) which can automatically spot all 24 Japanese phonemes by simply scanning among an input speech. The language processing part is made up of a predictive LR parser which predicts subsequent phonemes based on the currently processed phonemes. We call this 'hybrid' integrated recognition system 'TDNN-LR' method. The TDNN-LR recognition system provides large-vocabulary and continuous speech recognition. Two kinds of recognition experiments i.e., large-vocabulary isolated word recognition and continuous speech recognition were performed using the TDNN-LR method. Speaker-dependent recognition rates of 92.6% for the first choices and 97.6% for the top two choices were obtained for 5,240 Japanese common words, and rates of 65.1% for the first choices and 88.8% within the fifth choices were attained for phrase recognition.

## 1. Introduction

We have demonstrated that a phoneme spotting approach is effective for recognizing continuous speech. A phoneme-based recognition method is more adequate for a large vocabulary system than word-template-based recognition methods. Since Time-Delay Neural Networks (TDNN)[1] have superior phoneme recognition performance and time-shifting invariance, an accurate and efficient speech understanding system could be accomplished by adapting the TDNN spotting method to continuous speech recognition[2][3].

A Large Phonemic TDNN which can simultaneously recognize all 24 Japanese phonemes is constructed in a modular fashion[4][9]. Since the TDNN is trained using all 24 categories of phoneme tokens simultaneously,it is expected that false alarms could be reduced by the effect of lateral inhibition.

The speech recognition part consists of the Large Phonemic TDNN which can automatically spot all 24 Japanese phonemes (i.e., 18 consonants /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /N/, /s/, /sh/, /h/, /z/, /ch/, /ts/, /r/, /w/, /y/ and 5 vowels /a/, /i/, /u/, /e/, /o/ and a double consonant /Q/ or silence) by simply scanning among an input speech without any specific segmentation techniques. The Large Phonemic TDNN architecture is constructed as 4-layered back-propagation type networks in a modular fashion where a group of confusable phonemes are integrated into sub-network, and each sub-network is also integrated into one hidden layer. Training the Large TDNN is performed based on a fast back-propagation procedure[5] using shifted training tokens extracted from training word speech and/or training continuous speech, because the shift-invariance property of the Large TDNN is found to be effective in the region of 20-30 ms through a preliminary experiment.

On the other hand, the language processing part is made up of a predictive LR parser[7] in which the LR parser is guided by the LR parsing table automatically generated from context-free grammar rules, and proceeds left-to-right without backtracking. The predicted LR parser predicts subsequent phonemes based on the currently processed phonemes which are produced from the output units of the Large Phonemic TDNN scanning among an input speech along with it. Time alignment between the predicted phonemes and a sequence of the TDNN phoneme outputs is carried out by a DTW matching method. A duration control technique is applied for the predicted phonemes during the DTW matching to appropriately constrain the alignment.

We call this 'hybrid' integrated recognition system 'TDNN-LR' method. The TDNN-LR recognition system provides vocabulary-independent, large-vocabulary and continuous speech recognition because the Large Phonemic TDNN is trained by phoneme tokens extracted from various contexts of training word and/or continuous speech. Two kinds of recognition experiments i.e., large-vocabulary isolated word recognition and continuous speech recognition were performed using the TDNN-LR method.

## 2. TDNN-LR Recognition System

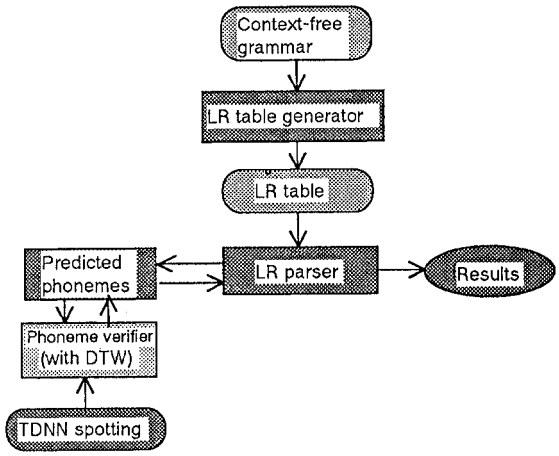


Fig.1. The TDNN-LR System

One of the great advantages in the integrated training is that it makes it possible to spot any phonemes without specific segmentation techniques. Another advantage is that it provides a *vocabulary-independent* recognition method, because the network was trained by phoneme tokens extracted from various contexts of speech, independent from training word contexts. To extend the high performance spotting results to *vocabulary-independent* large vocabulary speech recognition, a *"hybrid"* method combining a predictive LR parser[7] with a DTW alignment technique was proposed. We applied this method to 5,240 common Japanese words and phrases[10] uttered by a male speaker.

## 2.1. The TDNN Architecture for Spotting Phonemes

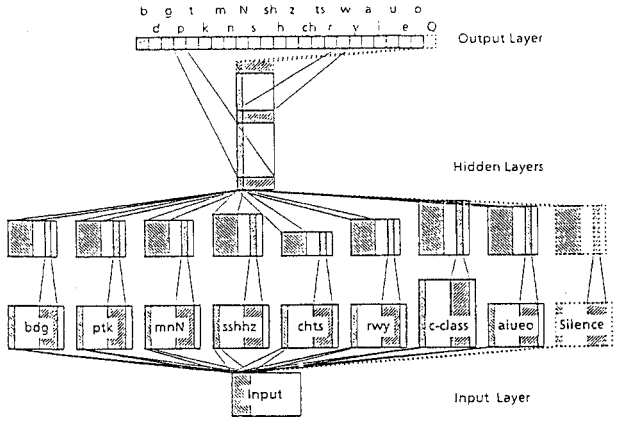


**Fig.2. The Large Phonemic TDNN Architecture**

The Large Phonemic TDNN has 24 output units corresponding to Japanese phonemes, i.e., /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /N/, /s/, /sh/, /h/,/z/, /ch/, /ts/, /r/, /w/, /y/, /a/, /i/, /u/, /e/, /o/, and double consonant (or silence) /Q/, as shown in Fig.2. The input layer consists of 240 units, i.e., 16 melscale filterbank coefficients * 15 frames (10 ms frame rate)[1].

To clarify the lateral inhibition effect, the TDNN is trained using training tokens (up to 200 tokens per category) which are extracted from the center positions of each phoneme in training words uttered by a male speaker. During the training, the fast back-propagation learning method[5] is used. Applying the trained TDNN to 2,620 test words uttered by the speaker, it was found that the TDNN time-shift-invariance is effective at at least 20 - 30 ms intervals from the center position of its input window.

## 2.2. Spotting Phonemes by TDNN

Applying the criterion shown in Table 1 to the results of phoneme spotting, it was demonstrated that 95.8% of the phonemes in the words were correctly spotted. On the other hand, false alarms were still to be found, especially in the vowel parts. It seems that training tokens should be extracted not only from phoneme centers but also from many different parts of the whole training word utterances.

**Table 1. A criterion to get spotting rates**

| Result | Definition |
|---|---|
| Correct | A firing pattern corresponding to the phoneme segment with 30ms wider than the label |
| Deletion error | No firing pattern exists in the above corresponding segment. |
| False alarm | A firing pattern that doesn't correspond to any segments. |

The Large Phonemic TDNN is retrained using tokens newly extracted at each 20 ms interval in 2,620 training word utterances. Each token is made of 15 frame data (150 ms), and overlaps with other tokens. Since it is difficult to categorize phoneme boundary data, those points within 10 ms from hand-labeled phoneme boundaries are omitted if there are other points for the phonemes. Tokens for silence are extracted from the points 20 ms before the beginning or after the end of words. The upper limit of the number of tokens, i.e., 1,000 tokens per category, are used.

Applying the newly retrained TDNN to the 2,620 test words, it is found that most of phonemes are correctly spotted throughout the each phoneme segment, and that the outputs corresponding to the other categories are well inhibited. Fig.3 shows an example of spotting phonemes in the word /toridasu/.

It is demonstrated that 98.0% of phonemes in the words are correctly spotted as shown in Table 2, and that more than 90% of the false alarms of /r/ and 60% - 80% of the false alarms of /g/ and /k/ are eliminated.

**Table 2. Phoneme spotting results**

| | |
|---|---|
| #Training tokens/cat. | 1,000 |
| #Phonemes | 13,974 |
| %Correct | 98.0% |
| %Deletion error | 2.0% |
| %False alarm rate | 23.2% |

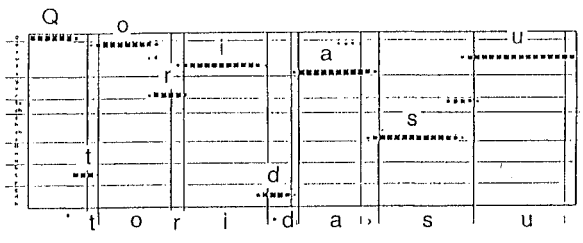


**Fig.3. An example of spotting results: input utterrance is /toridasu/.**

## 2.3. A Predictive LR Parser

LR parsing is well known in the field of program languages, and is applicable to a large class of context-free grammars. Generalized LR parsing[6] is a kind of LR parsing, and has been extended to handle arbitrary context-free grammars. For an ambiguous grammar, the LR parsing table has *multiple entries*. The LR parser is guided by the LR parsing table automatically created from context-free grammar rules, and proceeds left-to-right without backtracking. These parsing algorithms are very efficient for natural language processing.

A predictive LR parsing method predicts the next phonemes in input speech based on the currently processed phonemes. Up until now, an HMM continuous speech recognition system using an predictive LR parsing has been proposed[7]. This technique is also applicable for spotting results from the TDNN and a word or phrase grammar describing a large vocabulary or phrase database[10], respectively.

## 2.4. Integration of the TDNN and the Parser

The basic diagram of the recognition system using the TDNN spotting method and the predictive LR parser (we call this method "TDNN-LR") is shown in Fig.1. An input speech is converted to phoneme sequences of spotting results obtained via the Large Phonemic TDNN as shown in Fig.3. These phoneme sequences are processed by a DTW matching method, and then are evaluated by the predicted LR parser based on the grammar. The grammar rules are described by the context-free grammar as an LR table.

The LR parser predicts subsequent phonemes based on the currently processed phonemes. In the case of predicting plural phonemes, the parser processes the phonemes in parallel.

The predicted phonemes are evaluated by matching with the phoneme spotting results along a DTW alignment. This procedure continues until the input phonemes come to an end. However, because it requires plenty of time to process all predicted phonemes, a beam search taking the top ten candidates for each phoneme was adopted for reducing the computation.

The Likelihood of a similarity between a predicted phoneme and an input phoneme is defined as a logarithm of the activation value in outputs of the TDNN. The length of reference patterns (predicted phoneme patterns) is the average length of training phoneme tokens extracted from the training words of the large vocabulary. The slope constraint in DTW alignment is 1/2 to 2. The matching algorithm is as follows;

j: a predicted phoneme

p(t, j): an output value of a phoneme "j" at a frame "t"

D0(t): a table#0 for saving likelihood

D1(t): a table#1 for saving likelihood

[Initialization]

$Q(0, t) = D0(t)$, $Q(1, t) = D1(t)$, $t = 1,2,...,N$.

$Q(1, 1) = p(1, j)$, otherwise 0.

(Iterative formula)

for $t = 1,...,N$ and $i = 1,...,M$.

$$Q(i, t) = \max [\, Q(i-1, t-1) + \log(p(t, j)),$$
$$Q(i-2, t-1) + \log(p(t, j)) + \log(p(t, j)),$$
$$Q(i-1, i-2) + 0.5 * \log(p(t-1, j)) + 0.5 * \log(p(t, j))]$$

$D0(t) = Q(M-1, t)$, $D1(t) = Q(M, t)$.

This procedure is similar to a level-building DTW matching[8] with its endpoints free, which builds up subsequent phonemes.

# 3. Experiments

## 3.1. Large-Vocabulary Speech Recognition

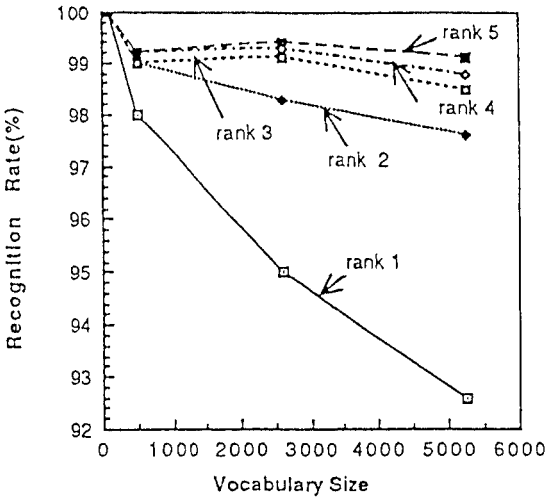


**Fig.4. Results on Large-Vocabulary Recognition**

Training of the Large Phonemic TDNN was performed using training tokens extracted from half of a large vocabulary Japanese database(i.e., 2,620 words).

In recognition experiments of large vocabulary, 5,240 common Japanese words were used. Among those words, another half of the large database which were *not* used for the network training were used as test words. The number of test words was incremented as 100, 500, 2,620 test words. On the other hand, the number of reference words was also incremented as 100, 500, 2,620 and 5,240 words, where in the former three cases, the reference words corresponded to the test words, and in the last case, the 5,240 reference words included the 2,620 test words as a subset. Therefore, note that this experiment is *vocabulary-independent* recognition.

Fig.4 shows the recognition rates of the n-th ($1 \leqq n \leqq 5$) top choices as a function of the vocabulary size of reference words from 100 to 5,240. In the case of the whole 5,240 words, a rate of 92.6% is obtained for the top choices, and rate of 97.6% and 99.1% are obtained for the second and fifth choices, respectively. While the rate of top choices decreases according to the increased vocabulary size, the rate within the top 5 choices is maintained higher than 99.1% for any vocabulary size.

Typical errors occurred in the following cases; (1)insertion errors at the beginning parts of the words such as /t/ or /k/ (ex.:"aisuru" → "taisuru"), (2)confusion between a double consonant and a closure accompanied by stop consonants(ex.: "itai" → "ittai").

### 3.2. Continuous Speech Recognition

Table 3. Phrase recognition results before and after incremental training(%)

| Rank | Before training | After training (100/cat.) | After training (200/cat.) |
|---|---|---|---|
| 1 | 55.0 | 64.4 | 65.1 |
| 2 | 70.1 | 79.5 | 78.4 |
| 3 | 76.6 | 81.7 | 87.1 |
| 4 | 81.3 | 86.0 | 88.1 |
| 5 | 82.7 | 88.8 | 88.8 |

The Large Phonemic TDNN is already trained by as much as 18,864 training tokens extracted from 2,620 training words. For the first experiment, continuous speech recognition experiments were conducted using the TDNN and an LR-parser describing *general* phrase grammar rules (its phoneme-perplexity is 5.9). The initial phrase recognition rate for 278 Japanese test phrases was 55.0% for the top choices and 82.7% for the top 5 choices, respectively. Because of different co-articulatory effects between word speech and continuous speech, incremental training of the TDNN using a small number of training tokens extracted from continuous training speech seemed to be needed.

The number of training tokens for incremental training is only 100 tokens per phoneme category (2,011 tokens in total are only 11% of the original tokens extracted from the training words). And then we increased the number up to 200 tokens per category (3,251 tokens in total). The phrase recognition rates are shown in Table 3 as compared with the rates before the incremental training. A phrase recognition rate of 65.1 % for the top choices and 88.8% for the top 5 choices were obtained. Therefore, efficiency in the adaptive incremental training using a small number of training tokens extracted from continuous speech was confirmed through this experiment.

## 4. Conclusion

We described an integration of speech recognition and language processing. The speech recognition part consists of the Large Phonemic Time-Delay Neural Networks (TDNN) which can automatically spot all 24 Japanese phonemes with an excellent spotting rate of 98.0% by simply scanning among an input speech along with it. The language processing part is made up of a predictive LR parser which predicts subsequent phonemes based on the currently processed phonemes. The TDNN-LR hybrid recognition system provides large-vocabulary and continuous speech recognition. Two kinds of recognition experiments i.e., large-vocabulary isolated word recognition and continuous speech recognition were performed using the TDNN-LR method. Speaker-dependent recognition rates of 92.6% for the first choices and 97.6% for the top two choices were obtained for 5,240 Japanese common words, and rates of 65.1% for the first choices and 88.8% within the fifth choices were attained for phrase recognition.


## Acknowledgements

The author would like to express his gratitude to Dr. Akira Kurematsu, president of ATR Interpreting Telephony Research Laboratories, for his support, and Dr. Alex Waibel of Carnegie Mellon University for his valuable suggestions, and to Dr. Kiyohiro Shikano, Mr. Masanori Miyatake and Mr. Yasuhiro Minami for their help in integrating the TDNN-LR system, and to Dr. Shigeki Sagayama for his comments on this paper. He is also indebted to the members of the Speech Processing Department at ATR, for their constant help in various stages of this research.


## References


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

[2] H. Sawai, A Waibel, P. Haffner, M. Miyatake and K. Shikano,"Parallelism, Hierarchy, Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes/CV-Syllables," Int. Joint Conf. on Neural Networks, Proceedings of IJCNN-89, vol. II , pp81-88, June 1989.

[3] H. Sawai, A. Waibel, M. Miyatake and K. Shikano,"Spotting Japanese CV-Syllables and Phonemes Using Time-Delay Neural Networks," IEEE, Proceedings of ICASSP-89, S1.7, May 1989.

[4] A. Waibel, H. Sawai and K. Shikano,"Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks," IEEE, Proceedings of ICASSP-89, S3.9, May 1989.

[5] P. Haffner, A. Waibel, H. Sawai and K. Shikano,"Fast Back-Propagation Learning Methods for Large Phonemic Neural Networks," European Conference on Speech Communication and Technology, pp553-556, Paris, Sep. 1989.

[6] M. Tomita,"Efficient Parsing for Natural Language - A Fast Algorithm for Practical Systems," Kluwer Academic Publishers (1986).

[7] K. Kita, T. Kawabata and H. Saito,"HMM Continuous Speech Recognition Using Predictive LR Parsing," IEEE, Proceedings of ICASSP-89, S13.3, May 1989.

[8] C. S. Myers and R. Labiner:"A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," IEEE Trans. on ASSP, vol.29, No.2, pp284-279 1981.

[9] M. Miyatake, H. Sawai, Y. Minami and K Shikano, "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks," IEEE, Proceedings of ICASSP-90, S8.10, Apr. 1990.

[10] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa and T. Watanabe, "Construction of a Large-Scale Japanese Database and Its Management System," IEEE, Proceedings of ICASSP-89, S10b.12, May 1989.