Spotting Japanese CV-Syllables and Phonemes Using Time-Delay Neural Networks

Hidefumi SAWAI, Alex WAIBEL, Masanori MIYATAKE and Kiyohiro SHIKANO

ATR Interpreting Telephony Research Laboratories

Abstract

Syllable or phoneme spotting if reliably achieved, provides a good solution to the spoken word and/or continuous speech recognition problem, . We previously showed that the Time-Delay Neural Network (TDNN) provided excellent recognition performance (98.6%) of the "BDG" consonant task. We would also like to extend the encouraging performance of TDNN to word/continuous speech recognition. In this paper, we show techniques for spotting Japanese CV syllables/phonemes in input speech based on TDNNs. We constructed the TDNN which can discriminate a single CV-syllable or phoneme group. In Japanese, there are only about one hundred syllables, or less than thirty phonemes, which makes it feasible to prepare and train the TDNN to spot all possible syllables or phonemes extracted as training tokens from training words. Syllable and phoneme spotting experiments show excellent results, including a syllable spotting rate of better than 96.7% correct. These spotting techniques are proved to be a good step toward continuous speech recognition.

1. Introduction

Spotting CV-syllables is a good approach to word and continuous speech recognition in Japanese because there are only about one hundred syllables. An earlier work was done with only 7 kinds of English demi-syllables[1]. The architecture of the TDNN[2,3] is extremely suitable for spotting CV-syllables/phonemes because the shift-invariant structure makes it possible to correctly spot them even in the neighboring positions of syllable/phoneme tokens. If we train a neural network which can reliably discriminate one syllable, and also prepare all kinds of neural networks, spotting of any syllables, in principle, can be achieved for any input utterances. However, because there are about one hundred syllables which can be chosen, one significant problem will occur when we choose as training tokens all possible syllables except the specific syllable to be discriminated. As the first step in spotting Japanese CV-syllables, we arbitrarily chose the syllable "BA" as a typical Japanese syllable. As training tokens, the "non-BA" syllables "DA", "GA", "PA", "TA" and "KA" are chosen because they might be confused with "BA". We might be able to automatically discriminate all possible syllables other than the five syllables used as training tokens. If this is true, both training time and the number of tokens will be greatly reduced.

In general, spotting phonemes is also an effective approach to speech recognition in other languages. We study the phoneme spotting approach for comparison with the syllable spotting approach described above. In this case, phoneme group networks which can discriminate one phoneme group (ex.:"BDG") are trained. "BDG", "PTK", "MNsN", "SShHZ", "ChTs", "RWY" and "AIUEO" are chosen as phoneme groups. As well as training phoneme group networks, the corresponding "intra-group" networks are also provided with training tokens of each subcategory (ex.:"B", "D" and "G" for the "BDG" intra-group network). Spotting experiments are performed by determining the phoneme category which is involved in the phoneme group. This "hierarchical" approach to spotting phonemes is critical if an incorrect phoneme group were determined. However, it is advantageous that only 7 phoneme groups and 7 intra-group networks need be prepared rather than preparing about one hundred CV-syllable networks. We will compare the performance of the two approaches in CV-syllable spotting and phoneme spotting.

2. TDNN for Spotting Japanese CV-Syllables

2.1. Architecture

The architecture of a TDNN for spotting a Japanese CV-syllable is shown in Fig.1. The input layer of the network is the same as a "BDG" network for discriminating "B", "D" and "G"[2]. The first hidden layer is composed of 4 units compared to 8 units for the "BDG" network. This is because the number of categories is reduced to two (i.e., one specific syllable and others). The second hidden layer is also composed of two units corresponding to two outputs in the output layer. As in phoneme recognition by the "BDG" network, 16 melscale spectrum coefficients serve as input to the network. Input speech, sampled at 12 kHz, was Hamming-windowed and a 256-point FFT computed every 5 msec. Melscale coefficients were computed from the power spectrum[3] and adjacent coefficients in time collapsed, resulting in an overall frame rate of 10 msec . The coefficients of the input token (in this case 15 frames of speech centered around the hand labeled vowel onset) were then normalized to fall between -1.0 and +1.0, with the average at 0.0. Fig.1 shows the resulting coefficients for the speech token "BA" as input to the network, with positive values shown in black squares and negative values in grey squares.

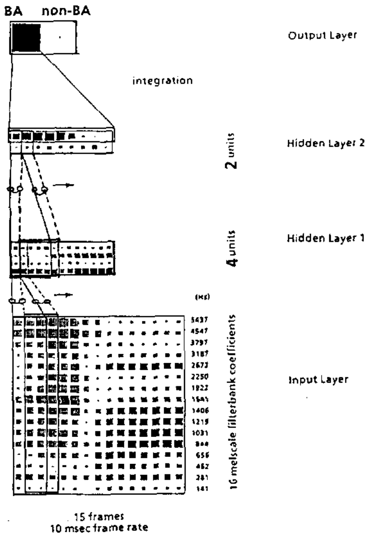

Fig.1. TDNN for spotting a Japanese CV-syllable

CH2673-2/89/0000-0025 $1.00 © 1989 IEEE

2.2. Experiment Conditions and Database

Experiments in spotting a Japanese CV-syllable are carried out by scanning the corresponding TDNN among input word speech uttered by one male speaker. Spotting syllables in words is an important step toward continuous speech recognition. If the spotting experiments are promising, this work even has the potential to spot any syllable in continuous speech.

For the evaluation of spotting Japanese CV-syllables, a large database of 5,240 common Japanese words were uttered in isolation by a native Japanese speaker. All utterances were recorded in a soundproof booth and digitized at a 12 kHz sampling rate. The database was then split into a training set and a testing set of 2,620 utterances each, from which the actual syllabic tokens were extracted. For the CV-syllable network training, tokens from various contexts of syllables containing "BA", "DA", "GA", "PA", "TA" and "KA" were selected. Both training and testing tokens were analyzed and windowed by a 256 point Hamming window, and then converted to feature parameters represented as 16th melscale FFT outputs every 10 msec.

2.3. Training in a TDNN for Spotting CV-Syllables

Several techniques for neural network training exist[4,5]. Training for our TDNN was done by a fast program of the back-propagation procedure[6] which iteratively adjusted all the weights in the network so as to decrease the error between the desired outputs and the network real output values. Although there are about one hundred "non-BA" syllables, it is inefficient to use all of them. Thus, only five "non-BA" syllables such as "DA", "GA", "PA", "TA" and "KA" are used, because these syllables are particularly confusable with "BA". The decision hyperplanes for discriminating between "BA" and "non-BA" may be formed without excessive training tokens for the "non-BA" syllables shown in Fig.2. Actually, as training tokens for "BA" syllables, we used 53 words with one or more "BA" syllables chosen from the even-numbered words of the 5,240 common Japanese word database. Similarly, as training tokens for "non-BA", 752 tokens which contain the syllables "DA", "GA", "PA", "TA" and "KA" were used. The boundaries between the consonants and the following vowels in the training syllables are centered at the input layer in TDNN. Training took only about several minutes using a fast program of the back-propagation procedure[6] on an 8-processor Alliant super-minicomputer.

2.4. Results of Spotting CV-Syllables

For testing CV-syllable spotting, 61 words containing one "BA" syllable each were taken from the testing word set. 138 "non-BA" CV-syllables (not only "DA", "GA", "PA", "TA" and "KA" but also other possible CV-syllables) were included in the 61 words. Given an unknown input CV-syllable, "BA" and "non-BA" were recognized according to the following criteria:

if o("BA") > o("non-BA"),
then the input CV-syllable is "BA",
if o("non-BA") > o("BA"),
then the input CV-syllable is "non-BA".

Where, o("BA") and o("non-BA") are the activation values of two outputs corresponding to "BA" and "non-BA", respectively.

Table 1. Results of Spotting CV-syllables (ex.: "BA")

Syllable	"BA"	"non-BA"	Total
#of syllables	61	138	199
#correct/total	59/61	137/138	196/199
Recognition rate(%)	96.7	99.3	98.5

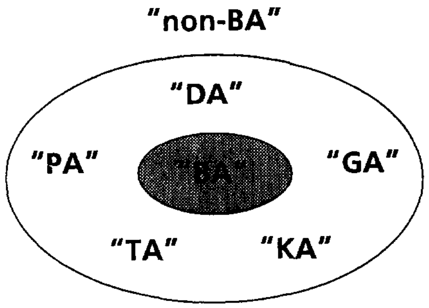

Fig.2. Training syllable tokens confusable with "BA" in Japanese

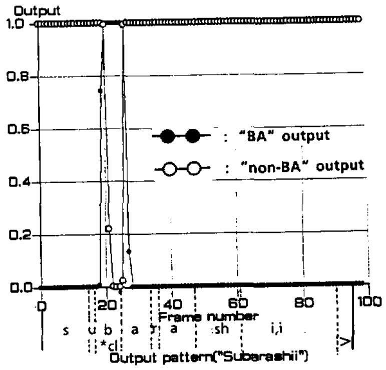

Fig.3. An example of CV-syllable spotting results: input utterance is "SUBARASHII".

In Table 1, spotting performance is shown. The syllable spotting rate is defined as the rate of correct output decisions of the network. The network achieved a "BA" spotting rate of 96.7% and performed well in inhibiting all possible CV-syllables (not only "DA", "GA", "PA", "TA" and "KA" but also all other syllables) belonging to the "non-BA" category at a rate of 99.3%.

Fig.3 shows an example of spotting results, where the input utterance is the Japanese word "SUBARASHII". The output of "BA" is robustly fired at the position of "BA", and all other syllables are well inhibited as shown in the "non-BA" output values.

Spotting errors occured at the beginning of the word "ASHIBA" where "A" misfired as "BA", and in two intermediate parts of words such as "KEIBATSU" and "SHIBARAKU" where the "BA" outputs were not robustly fired.

3. TDNN for Spotting Phonemes

3.1. Hierarchical Architecture

For a good comparison with the CV-syllable spotting approach described in Section 2, we examined a phoneme

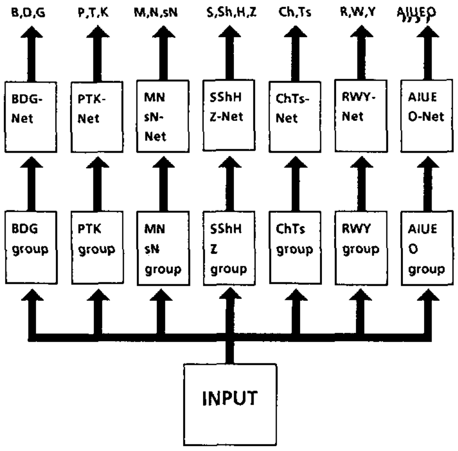

Fig.4. A Hierarchecal structure of TDNNs for spotting phonemes

spotting approach which has the hierarchical decision criterion shown in Fig.4. This criterion is based on first spotting phoneme groups, and then spotting phonemes determined by the results of the phoneme group spotting.

For the phoneme group TDNNs, networks similar to the one shown in Fig.1 were constructed with a slight modification, using 8 units in the first hidden layer instead of the 4 units shown in Fig.1, because phoneme group TDNNs need a greater capacity in the hidden units than the syllabic TDNN shown in Fig.1. All phonemic categories are divided into 7 phoneme groups, i.e., "BDG", "PTK", "MNsN", "SShHZ", "ChTs", "RWY" and "AIUEO". These 7 phoneme group TDNNs were trained by the back-propagation procedure. The 7 intra-group TDNNs[7] were also trained to discriminate phonemes, for example, between "B", "D", and "G" corresponding to the "BDG" phoneme group TDNN.

For testing phoneme spotting, a "hierarchical" decision was made by first determining one or more phoneme groups and after that, by discriminating phonemes in the phoneme group determined.

3.2. Experiment Conditions and Database

As with the experiment conditions of the CV-syllable spotting described in Section 2.2, the even-numbered Japanese common words were used as training tokens. These tokens were analyzed with the same experiment conditions, and then all phonemic tokens were extracted from a variety of training word phoneme contexts by inspection.

Testing on spotting phonemes was performed for 50 words of the phoneme balanced words selected from the Japanese database and additional words, so that all phonemes would appear equally often.

3.3. Training in the TDNN for Spotting Phonemes

As with the CV-syllable spotting networks, the phoneme group networks were trained by the back-propagation procedure using both the corresponding training tokens in a phoneme group (ex.:"BDG") and the residual phoneme group tokens (ex.: tokens other than "BDG"). The residual training tokens were selected by easily confused phonemic group tokens based on the phoneme group recognition confusion matrix. This confusion matrix is shown in Table 2. This selection of residual training tokens is necessary to reduce the number of tokens and time for phoneme group network training.

Table 2. Confusion matrix in phoneme group recognition (%)

OUT IN	BDG	PTK	MNsN	SShHZ	ChTS	RWY	AIUEO
BDG	97.7	0.3	1.0			0.7	0.3
PTK		94.0	0.5	0.5	0.5		4.7
MNsN	1.3		97.3				1.3
SShHZ			1.7	93.3	2.3	0.3	2.3
ChTs				5.1	94.9		
RWY	0.7	1.1		1.1		88.5	8.6
AIUEO	0.4	0.5	0.4	0.1	0.1	0.6	97.9

For intra-group network training (ex.:"BDG" net), up to 600 training tokens per phoneme category were used[3]. These intra-group networks consist of "BDG", "PTK", "MNsN", "SShHZ", "ChTs", "RWY" and "AIUEO". The 7 kinds of phoneme group networks and intra-group networks trained are constructed in a hierarchical fashion as shown in Fig.4.

3.4. Results of Spotting Phonemes

Typical spotting results are shown in Fig.5, where the Japanese word "TORIATSUKAU" was spotted. In Fig.5, the input layer is shown at the bottom, the output activation pattern of the phoneme group networks is shown in the middle, and the final output activation pattern of the intra-group networks is shown at the top. The word utterance is manually labeled by inspection to check coincidences between labels and spotting results.

In the Fig.5 phoneme group spotting results, although some misfired activation patterns such as "BDG", "SShHZ", "RWY" are observed, correct spotting results were obtained for all input phonemes except a devocalized "U" after "TS". The spotting rate of correct phoneme groups is better than 96%, averaged over the 50 testing words.

In the final phoneme spotting results (i.e.,intra-group spotting), activation outputs belonging to correct phoneme categories are also observed in the phoneme group spotting results. The spotting rate of correct phonemes is 92%, averaged over the same 50 testing words. Some misfired activations in wrong categories still occurred because the networks were trained only from training tokens extracted from phoneme center positions. Nonetheless, the results of the first phoneme spotting trial are excellent.

4. Conclusion

We have performed two kinds of spotting experiments: CV-syllable spotting and phoneme spotting, and showed that high performance Japanese CV-syllable and phoneme spotting are indeed possible using the Time-Delay Neural Networks (TDNNs). The syllable spotting approach uses a simple TDNN architecture and is easily extended to other CV-syllable networks only by making training tokens of a specific syllable and its easily confused syllables.

In the network for spotting CV-syllables, "BA" syllables and only "DA", "GA", "PA", "TA" and "KA" as "non-BA" syllables are trained because they are easily confused with "BA". The spoting experiments on "BA" syllables showed that a rate of 96.7% was achieved, and other possible syllables except "BA" (not only "DA", "GA", "PA", "TA" and

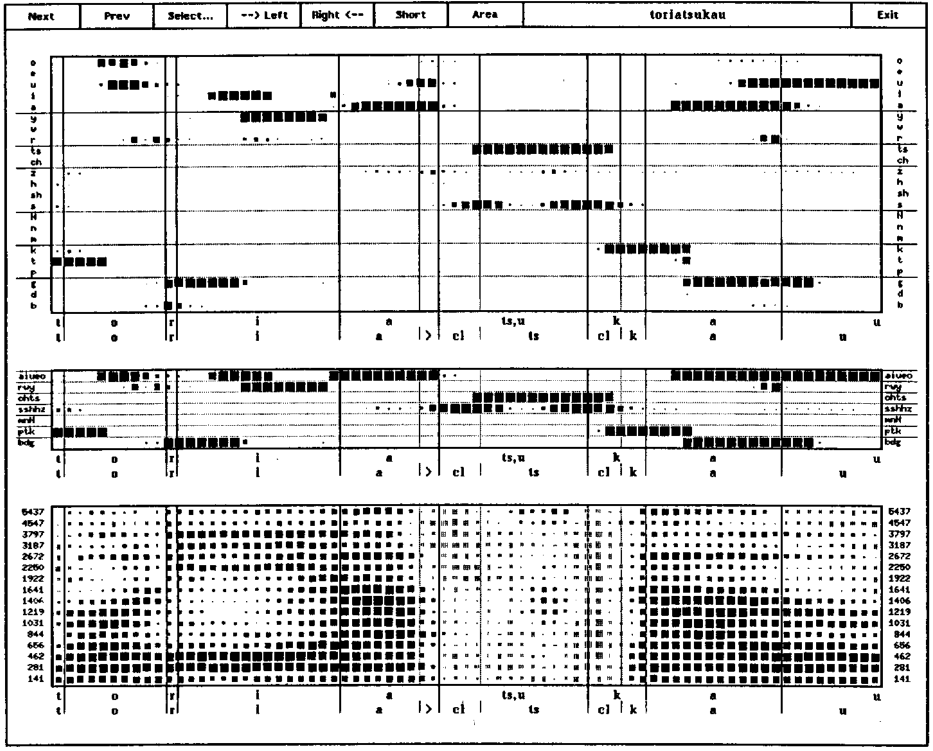

Fig.5. An example of phoneme spotting by a hierarchical TDNN: input utterance is "TORIATSUKAU".

"KA" but also all other syllables) are well inhibited at a rate of 99.3%.

On the other hand, in the phoneme spotting, a hierarchical TDNN architecture was applied to input word utterances where phoneme group spotting was then performed as a first step, and phoneme discrimination was performed within the determined phoneme group. The advantage of this phoneme spotting approach is that the network needed is smaller than the CV-syllable network, and that spotting any phoneme is possible for other languages as well as Japanese.

Even though some misfired patterns appeared in phoneme spotting because the network were trained only with easily confused training tokens extracted from the center positions of phonemic tokens, excellent spotting results were obtained.

These results in spotting Japanese-CV syllables and phonemes in words strongly suggested that these spotting techniques can be applied to recognizing not only spoken words but also continuous speech.

Acknowledgement

The authors would like to express their gratitude to Dr.Akira Kurematsu, president of ATR Interpreting Telephony Research Laboratories, for his encouragement and support, and to Mr.Patrick Haffner who developed the fast back-propagation program, and to Ms. Kitaoka who helped prepare the graphic tools for displaying the results of the spotting experiments. We are also indebted to the members of the Speech Processing Department at ATR, for their constant help in the various stages of this research.

References

[1] C.Kamm, T.Landauer and S. Singhal, "Using an Adaptive Network to Recognize Demisyllables in Continuous Speech," J.Acoust.Soc.Amer., vol.83,suppl.1.X7,May 1988.

[2] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang, "Phoneme Recognition: Neural Networks vs. Hidden Markov Models," International Conference on Acoustics, Speech, and Signal Processing, Apr. 1988.

[3] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang, "Phoneme Recognition Using Time-Delay Neural Networks," Technical Report TR-I-0006, ATR Interpreting Telephony Research Laboratories, Oct. 1987.

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

[5] R.P.Lipmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine, 4-22, Apr. 1987.

[6] P.Haffner, A.Waibel, H.Sawai and K.Shikano, "Fast Back-Propagation Learning Methods for Neural Networks in Speech," Technical Report TR-I-0058, ATR Interpreting Telephony Research Laboratories, Nov. 1988.

[7] A. Waibel, H. Sawai, K. Shikano, "Modularity and Scaling in Large Phonemic Neural Networks," Technical Report TR-I-0034, ATR Interpreting Telephony Research Laboratories, Aug. 1988.