PAPER *Special Issue on Continuous Speech Recognition and Understanding*

Connectionist Approaches to Large Vocabulary Continuous Speech Recognition

Hidefumi SAWAI†, *Member*, Yasuhiro MINAMI††, *Nonmember*,
Masanori MIYATAKE†††, *Member*, Alex WAIBEL††††, *Nonmember*
and Kiyohiro SHIKANO†††††, *Member*

SUMMARY This paper describes recent progress in a connectionist large-vocabulary continuous speech recognition system integrating speech recognition and language processing. The speech recognition part consists of Large Phonemic Time-Delay Neural Networks (TDNNs) which can automatically spot all 24 Japanese phonemes (i.e., 18 consonants /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /N/, /s/, /sh/ ([ʃ]), /h/, /z/, /ch/ ([tʃ]), /ts/, /r/, /w/, /y/ ([j]) and 5 vowels /a/, /i/, /u/, /e/, /o/ and a double consonant /Q/ or silence) by simply scanning among input speech without any specific segmentation techniques. On the other hand, the language processing part is made up of a predictive LR parser 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. Time alignment between the predicted phonemes and a sequence of the TDNN phoneme outputs is carried out by the DTW matching method. We call this 'hybrid' integrated recognition system the 'TDNN-LR' method. We report that large-vocabulary isolated word and continuous speech recognition using the TDNN-LR method provided excellent speaker-dependent recognition performance, where incremental training using a small number of training tokens is found to be very effective for adaptation of speaking rate. Furthermore, we report some new achievements as extensions of the TDNN-LR method: (1) two proposed NN architectures provide robust phoneme recognition performance on variations of speaking manner, (2) a speaker-adaptation technique can be realized using a NN mapping function between input and standard speakers and (3) new architectures proposed for speaker-independent recognition provide performance that nearly matches speaker-dependent recognition performance.

1. Introduction

In this paper, we describe recent progress in the connectionist large-vocabulary continuous speech recognition system we have developed at ATR. First, we review our research achievements: phoneme recognition and phoneme spotting techniques using Time-Delay Neural Networks (TDNNs). Second, we propose a large-vocabulary and continuous speech recognition method using TDNN phoneme spotting and LR parsing technique. Finally, we propose some new connectionist approaches to *robust* speech recognition, *speaker-independent* phoneme recognition and *speaker-adaptation* techniques as extensions of the current system.

In Sect. 2, we review that a TDNN performs excellent phoneme recognition for a small but difficult task, i.e., bdg-phoneme recognition, and for all consonant task. Scaling up connectionist models to larger connectionist systems is difficult, because large networks require increasing amounts of training time and data, and the complexity of the optimization task quickly reaches computationally unmanageable proportions. We trained several small TDNNs aimed at all phonemic subcategories (nasals, fricatives, vowel, etc.) and then integrated those sub-networks into an all-consonant network[4]-[6].

In Sect. 3, we describe phoneme spotting techniques. Phoneme spotting if reliably achieved, provides a good solution to the spoken word and/or continuous speech recognition problem. Training the Large TDNN is performed based on a back-propagation procedure[7],[8] using shifted training tokens[12] extracted from training-word speech and/or training continuous speech. We then report excellent spotting rates and effectiveness of adaptive training using continuous speech.

In Sect. 4, we propose an integration of speech processing and language processing. The speech recognition part consists of the Large Phonemic TDNN which can automatically spot all 24 Japanese phonemes by simply scanning among an input speech without any specific segmentation techniques. On the other hand, the language processing part is made up of a predictive LR parser[17] 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. Time alignment between the predicted phonemes and a sequence of the TDNN phoneme outputs is carried out by a DTW

Manuscript received January 10, 1991.
Manuscript revised March 25, 1991.

† The author is with ATR Interpreting Telephony Research Laboratories, Kyoto-fu, 619-02 Japan.

†† The author is with the Faculty of Science and Technology, Keio University, Yokohama-shi, 223 Japan.

††† The author is with Information and Communication Systems Research Center, Sanyo Electric Company, Hirakata-shi, 573 Japan.

†††† The author is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213-3890 U.S.A.

††††† The author is with NTT Human Interface Laboratories, Musashino-shi, 180 Japan.

matching method.

In Sect. 5, two kinds of recognition experiments i.e., large-vocabulary isolated word recognition and continuous speech recognition, were performed using the TDNN-LR method. In the large-vocabulary isolated word recognition, 5,240 common Japanese words are used. In the continuous speech recognition, 278 test phrases in the "ATR conference registration task" are recognized.

In Sect. 6, we propose some new connectionist approaches to robust speech recognition, speaker-adaptation and speaker-independent recognition as extensions of the TDNN-LR method: two proposed NN architectures provide robust phoneme recognition performance on variations of speaking manner; a speaker-adaptation technique can be realized using a NN mapping function between input and standard speakers; and new architectures proposed for speaker-independent recognition provide performance that nearly matches speaker-dependent recognition performance.

2. Phoneme Recognition Using TDNN

2.1 TDNN Architecture

For the recognition of phonemes, a three-layer net is constructed. Its overall architecture and a typical set of activities in the units are shown in Fig. 1 based on one of the phonemic subcategory tasks (BDG).

At the lowest level, 16 melscale spectral 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[1] and coefficients adjacent in time collapsed resulting in an overall 10 msec frame rate. The coefficients of an input token (in this case 15 frames of speech centered around the hand labeled vowel onset) were then normalized to lie between −1.0 and +1.0 with the average at 0.0. Figure 1 shows the resulting coefficients for the speech token "BA" as input to the network, where positive values are shown as black squares and negative values as grey squares. The detailed architecture is described in Ref. (1). The TDNN achieved a recognition rate of 98.5% averaged for three male speakers[1].

2.2 Consonant Recognition by Modular TDNN Design

Our consonant TDNN (shown in Fig. 2) was constructed modularly from networks aimed at the consonant subcategories, i.e., the bdg-, ptk-, mnN-, sshhz-, chts- and the rwy-tasks. Each of these nets had been trained before to discriminate between the consonants within each class. In addition, an interclass discrimination net that distinguishes between the consonant subclasses was trained. This hopefully provides missing feature information for interclass discrimination. Three connections were then established to each of the 18 consonant output categories (/b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /N/, /s/, /sh/ ([ʃ]), /h/, /z/, /ch/ ([tʃ]), /ts/, /r/, /w/ and /y/ ([j])): one to connect an output unit with the appropriate interclass discrimination unit in hidden layer 2, one with the appropriate intraclass discrimination unit from hidden layer 2 of the corresponding subcategory net and one with the always-activated threshold unit (not shown in Fig. 2). The overall network architecture is illustrated in Fig. 2 for the case of an incoming test token (e.g., a /g/). The performance of the network yielded 96.0% correct consonant recognition over the test data[3]-[5]. Furthermore, a fast back-propagation method later developed at ATR made it possible to train the consonant network from random weight values at the same time, and yielded a better recognition rate of 96.7%[8].

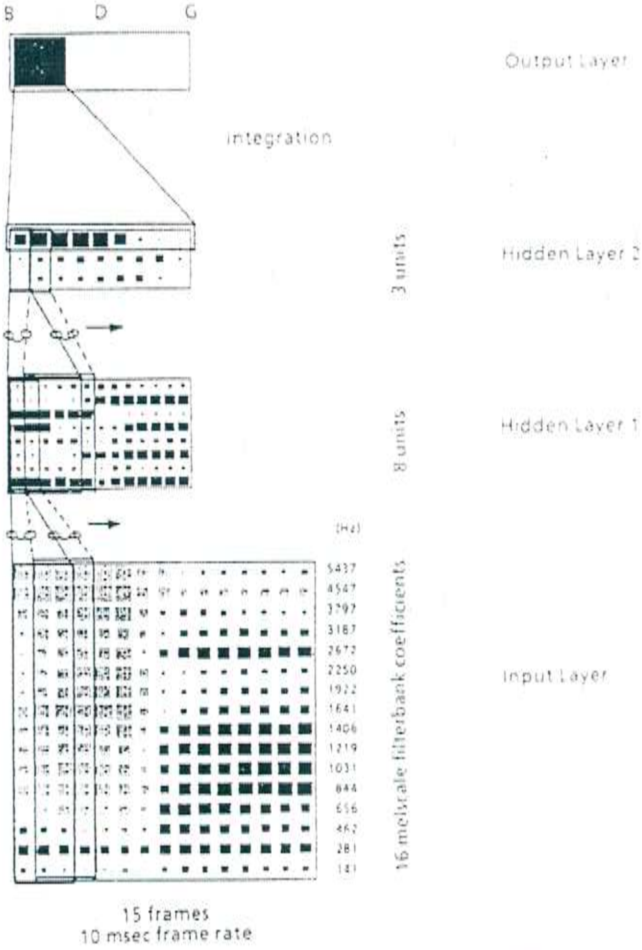

Fig. 1 The TDNN architecture (input: "BA").

3. Phoneme Spotting Using TDNN

A large TDNN architecture for discriminating 24

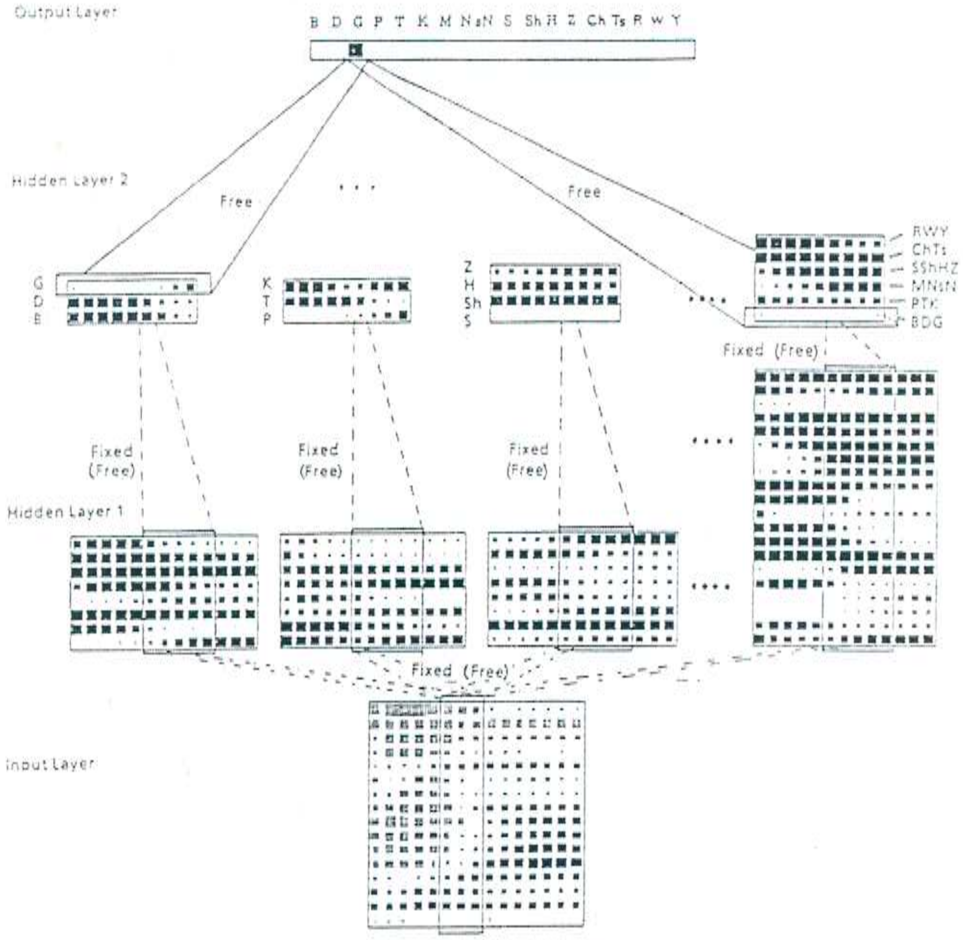

Fig. 2 Modular construction of all consonant network.

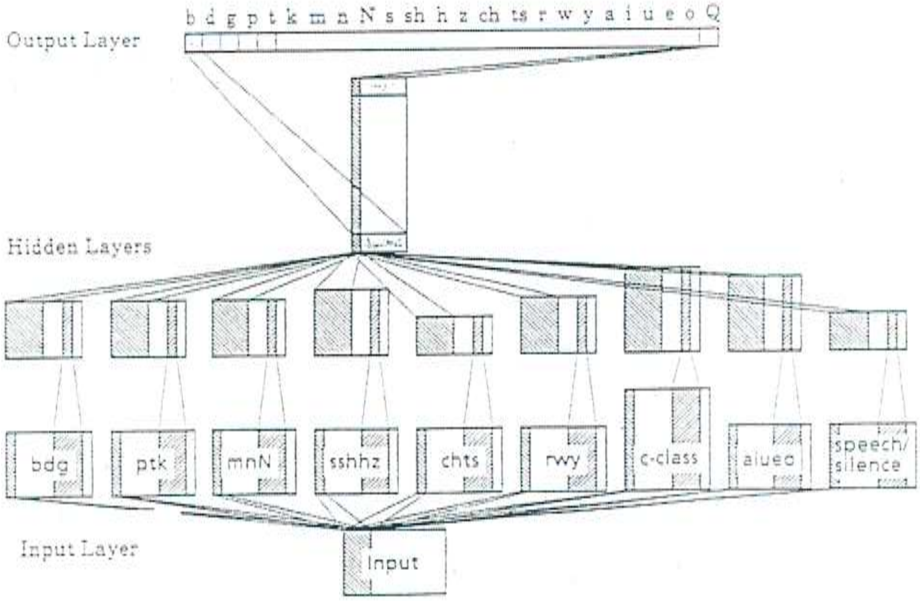

Fig. 3 The large phonemic TDNN architecture.

Japanese phonemes (18 consonants: /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /N/, /s/, /sh/ ([ʃ]), /h/, /z/, /ch/ ([tʃ]), /ts/, /r/, /w/, /y/ ([j]), and 5 vowels /a/, /i/, /u/, /e/, /o/, and silence) was constructed as shown in Fig. 3[3],[6],[11],[12]. This TDNN is modulay constructed by 6 intra-class subnetworks discriminating among "bdg", "ptk", "mnN", "sshhz", "chts" and "rwy", an intra-class subnetwork discriminating among consonant groups, a vowel network, and a silence network discriminating between silence and speech. These subnetworks are integrated into a third hidden layer which has 24 units so that their corresponding output units can be laterally inhibited.

Phoneme tokens for training the TDNN are classified into 24 phoneme categories, based on hand labeling, extracted from even-numberd words of the 5,240 common words uttered by a male speaker. The number of training tokens per phoneme category ranges up to 1,000, randomly selected from the extracted tokens. Tokens are duplicated when the number per category can not reach 1,000. Training the TDNN is performed using a fast back-propagation learning procedure[8].

Phoneme spotting outputs are obtained as recognition results by shifting the input layer among input speech frame by frame. This phoneme spotting method does not require any phoneme segmentation techniques and can get spotting results merely by scanning the network[2],[3],[11],[12].

Table 1 shows spotting results for 2,620 test words when using up to 400 and 1,000 training tokens/category, respectively. It is demonstrated that 98.0% of the phonemes in the test words are correctly spotted for the latter case, yielding a false alarm rate of 23.2%.

Thus, the Large Phonemic TDNN is already trained by as many as 18,864 training tokens extracted from 2,620 training words. For the first experiment, spotting experiments in continuous speech were conducted using the TDNN. The initial correct phoneme spotting rate in 278 Japanese test phrases was 81.2% with a false alarm rate of 47.8%, as shown in Table 2. Because of the different co-articulatory effects of word speech and continuous speech, incremental TDNN training using a small number of tokens extracted from continuous training speech seemed to be needed. The number of tokens for incremental training is only 100/200 tokens per phoneme category (2,011/3,251 tokens are only 11%/17% of the original tokens extracted from the training words). The correct phoneme spotting rate was significantly improved from 81.2% to 89.0%/89.1% after the adaptive incremental training. More importantly, the false alarm rate decreased from 47.8% to 34.8%/25.8%. Figure 4 shows an example of phoneme spotting results in the phrase/touroku-wo/. The lower layer shows an input spectrogram and the upper shows spotting outputs. We can also expect better phrase recognition rates in continuous speech after the incremental training.

Table 1 TDNN phoneme spotting results on large-vocabulary.

Phonemes	# of phonemes	400tokens/category			1,000 tokens/category		
		Correct	Deletion	false alarm	Correct	Deletion	false alarm
b	231	228	3	268	225	6	104
d	180	175	5	106	171	9	71
g	265	230	35	198	210	55	57
p	28	25	3	203	26	2	104
t	461	452	9	178	459	2	235
k	1300	1218	82	116	1283	17	245
m	485	482	3	323	479	6	213
n	273	258	15	84	263	10	63
N	488	487	1	161	488	0	163
s	572	570	2	175	572	0	100
sh	387	385	2	52	386	1	81
h	313	312	1	215	310	3	159
z	315	310	5	170	307	8	87
ch	141	140	1	57	141	0	163
ts	220	219	1	205	218	2	235
r	760	709	51	62	730	30	97
w	81	80	1	74	79	2	13
y	573	531	42	124	561	12	171
a	1772	1770	2	108	1771	1	85
i	1333	1282	51	155	1302	31	200
u	1615	1496	119	206	1543	72	200
e	829	822	7	222	827	2	254
o	1352	1337	15	97	1348	4	136
Total	13974	13518 (96.7%)	456 (3.3%)	3559 (25.5%)	13699 (98.0%)	275 (2.0%)	3236 (23.2%)

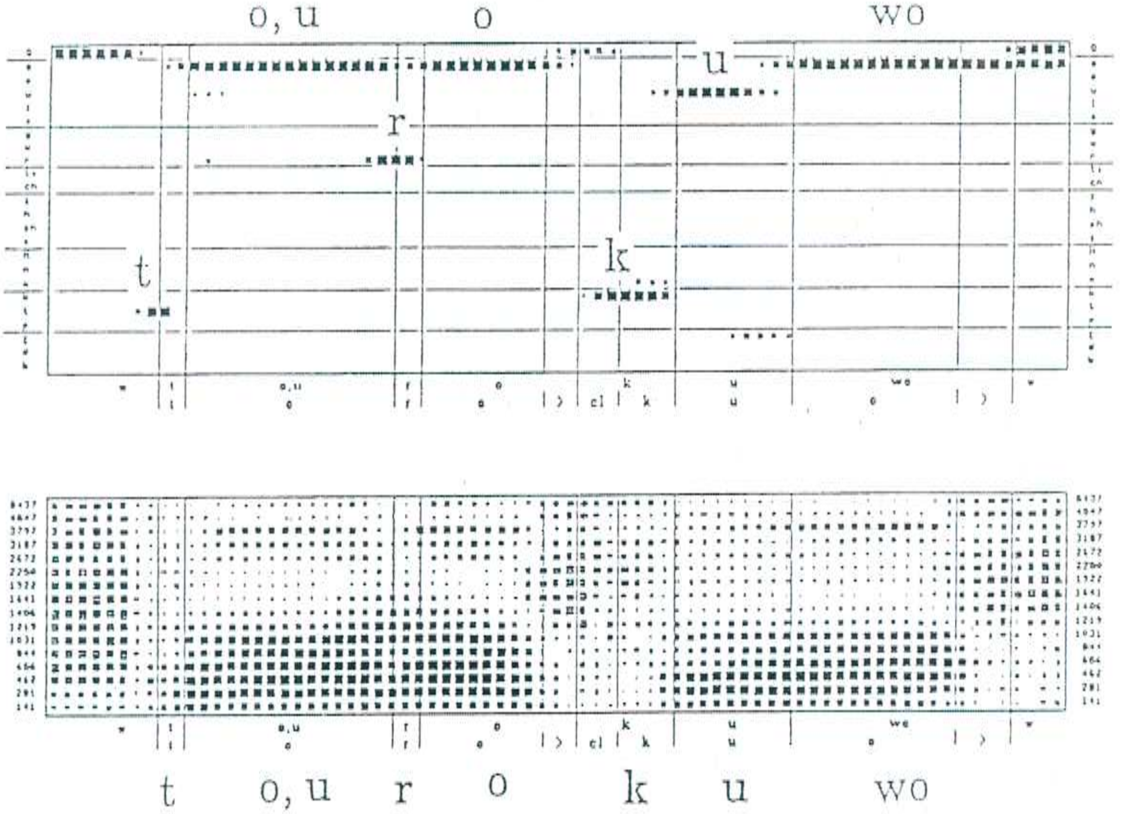

Fig. 4 An example of phoneme spotting results : (phrase name is /touroku-wo/).

4. The TDNN-LR Recognition System

To extend the high performance spotting results to large-vocabulary continuous speech recognition, a "hybrid" method combining a predictive LR parser[17] with a DTW alignment technique was proposed. We applied this method to 5,240 common Japanese words and phrases[19] uttered by the male speaker.

4.1 LR Parsers

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[16] is a kind of LR parsing, and has been extended to handle arbitrary context-free grammars. 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.

An example of sentences in the LR parsing and an LR parsing table are shown in Figs. 5 and 6, respectively[17]. The LR table consists of an ACTION table and a GOTO table. In Fig. 6, lines show grammar symbols, and columns show parser status. The symbols "s" and "r" show "shift" and "reduce" actions, respectively.

```
(1) S    →  NP  V
(2) NP   →  N
(3) NP   →  N   P
(4) N    →  m   a   m   e
(5) N    →  a   r   e
(6) P    →  o
(7) V    →  o   k   u   r   e
(8) V    →  k   u   r   e
```

Fig. 5 An example of context-free grammar.

	a	u	e	o	k	m	r	$	S	N	V	P	NP
0	s2					s3			5	4			1
1				s7	s6						8		
2							s9						
3	s10												
4				s11,r2								12	
5								acc					
6		s13											
7					s14								
8								r1					
9			s15										
10						s16							
11				r6									
12				r3									
13							s17						
14		s18											
15				r5									
16			s19										
17			s20										
18							s21						
19				r4									
20								r8					
21			s22										
22								r7					

Fig. 6 An example of ACTION and GOTO tables.

The figure on the right side of "s" is the next status in a "shift" action. The figure on the right side of "r" is the number of grammar rules. The right side shows a GOTO table where the figure indicates the next status value.

A predictive LR parsing method predicts the next phonemes in input speech based on the currently processed phonemes. An HMM continuous speech recognition system using an predictive LR parsing has been evaluated[17]. This technique is also applicable to spotting results from the TDNN and a word or phrase grammar describing a large vocabulary or phrase database[19], respectively. Prediction can be easily realized by referencing an LR table such as Fig. 6. By way of analysis, when the predictive LR parser is in a status, possible phonemes served to this LR parser are the only phonemes described by "shift" and "reduce" on a line of the table. The predictive LR parser regards these phonemes as predicted phonemes.

4.2 Integration of the TDNN and the Parser

The basic structure of the recognition system which utilizes TDNN spotting and predictive LR parsing is shown in Fig. 7 (hereafter: TDNN-LR)[9]–[11],[13]–[15]. First, an input speech is converted to outputs via TDNN phoneme spotting shown in the upper part of Fig. 4. Matching between these outputs and reference words is performed by the predictive LR parser according to a grammar rule. When plural phonemes are predicted, the predictive LR parser analyzes the phonemes in parallel. The predicted phoneme sequences are evaluated by a DP match between predicted phonemes and the TDNN phoneme spotting results. This procedure continues until input phonemes come to an end. However, since it takes considerable time to process all predicted phonemes, a beam search is used to take the first "B" candidates, where "B" is the width of the beam (ex.: $B = 100$).

The likelihood of a similarity between a predicted

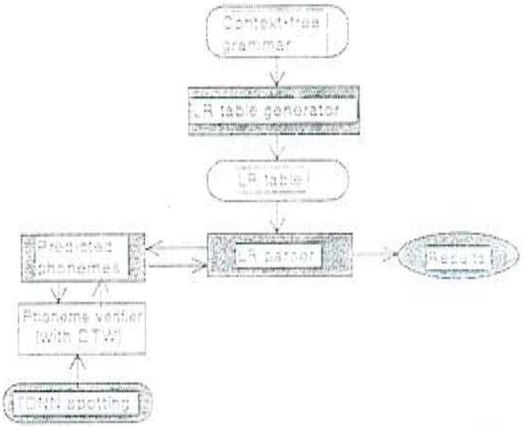

Fig. 7 The TDNN-LR speech recognition system.

Table 2 TDNN phoneme spotting results on test phrases.

Phoneme	# of phonemes	No adaptive training			Adaptive training (200 tokens/cat.)		
		Correct	Deletion	false alarm	Correct	Deletion	false alarm
b	16	14	2	19	10	6	5
d	69	44	25	29	62	7	19
g	34	19	15	36	19	15	17
p	10	8	2	6	10	0	6
t	70	48	22	11	68	2	56
k	210	182	28	94	195	15	7
m	58	36	22	19	18	40	23
n	74	36	38	5	33	41	11
N	34	24	10	25	27	7	19
s	74	67	7	20	72	2	9
sh	53	50	3	15	53	0	17
h	27	14	13	10	24	3	48
z	32	32	0	53	32	0	25
ch	17	10	7	11	16	1	7
ts	24	24	0	79	22	2	13
r	66	53	13	43	61	5	24
w	25	5	20	3	23	2	14
y	96	80	16	30	90	6	30
a	279	238	41	23	272	7	16
i	192	160	32	27	179	13	39
u	97	90	7	287	85	12	50
e	127	107	20	38	121	6	27
o	256	234	22	43	237	19	18
total	1940	1575 (81.2%)	365 (18.8%)	926 (47.8%)	1729 (89.1%)	211 (10.9%)	500 (25.8%)

phoneme and an input phoneme is defined as the logarithm of the activation value of TDNN output, where the likelihood is regarded as a posterior probability for each output. The length of the reference patterns (predicted phoneme patterns) is the average length of the training phoneme tokens extracted from the training words of the large vocabulary. The slope constraint in DTW alignment is 1/2 to 2. The detailed matching algorithm is described in Refs. (10), (14), (18).

5. Recognition Experiments

5.1 Large-Vocabulary Speech Recognition

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.

Figure 8 shows the recognition rates of the n-th $(1 \leq n \leq 5)$ 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(11),(14),(28).

Recognition error in 5,240 common words is classified into the following three cases: (1) insertion of "t" or "k" at the beginning of a word (ex.: "aisuru" → "taisuru"). (2) a short word is misrecognized as a long word (ex.: "aa" → "hanahada"). (3) a double consonant is confused with a silence accompanied by an unvoiced stop (affricate) (ex.: "itai" → "ittai").

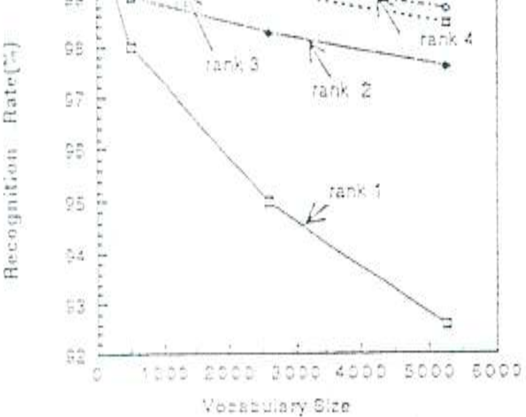

Fig. 8 Results on large-vocabulary recognition.

5.2 Continuous Speech Recognition

The Large Phonemic TDNN is already trained by as many as 18,864 training tokens extracted from 2,620 training words. As an first attempt, continuous speech recognition experiments were conducted using the trained TDNN and an LR-parser describing *general* phrase grammar rules (its phoneme-perplexity is 5.9). Table 3 shows the features of the ATR "Conference Registration" task we used. 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). We then increased the number up to 200 tokens per category (3,251 tokens in total). The phrase recognition rates are shown in Table 4 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

Table 3 Features of the task.

Number of words	1,035
Number of rules	1,656
Number of states in LR	5,015
Phoneme perplexity	5.9
Entropy/phoneme	2.6 bit
Average number of phonemes/phrase	7.32

Table 4 Phrase recognition rates (%).

Rank	Before adaptive training (without duration control)	Before adaptive training (with duration control)	After adaptive training (100 cat)	After adaptive training (200 cat)
1	52.9	55.0	64.4	65.1
2	70.1	70.1	79.5	78.4
3	77.7	76.6	81.7	87.1
4	81.7	81.3	86.0	88.1
5	82.4	82.7	88.8	88.8
6~10	86.3	87.1	93.2	91.0
11~15	87.4	87.4	93.5	92.4
16~	12.6	12.6	6.5	7.6

choices were obtained. Therefore, the efficiency of adaptive incremental training using a small number of training tokens extracted from continuous speech was confirmed through this experiment.

Typical errors are as follows:

(1) Substitution errors between /n/ and /m/.
ex.: /saNka-no/→/saNka-mo/,
/syotei-no/→/syotei-mo/.

These errors occurred due to the fact that the number of /m/ and /n/ phoneme tokens for incremental training was too small (17 tokens for /m/ and 13 tokens for /n/) compared with the original training tokens extracted from training words (1,000 for /m/ and 460 for /n/).

(2) Phoneme insertion errors.
ex.: /zyuusho/→/zyuusho-o/,
/happyou/→/happyou-o/.

These errors occurred due to difficulty of precise duration control at the end of the utterances.

6. Extensions

In this section, we describe extensions in the TDNN-LR speech recognition system: *robustness for variations of speaking manner, speaker-adaptation and speaker-independent phoneme recognition.*

6.1 Neural Network Architectures for Robust Speech Recognition

Until now, Time-Delay Neural Networks (TDNN) architecture has been applied to several speaker-dependent recognition stages, such as phoneme recognition (described in Sect. 2), Japanese phoneme spotting (in Sect. 3), and the TDNN-LR large-vocabulary continuous speech recognition system with integrated training for spotting Japanese phonemes (in Sect. 4). If we extend these recognition methods based on TDNN to a continuous, speaker-independent speech recognition system, a novel robust recognition strategy should be developed. This section introduces several novel TDNN architectures for robust speaker-independent, continuous speech recognition[20],[21].

One novel architecture for a Frequency-shift-invariant TDNN (FTDNN) is based on the frequency-time-shift-invariance as well as the time-shift-invariance by constructing the same weighting values between the input layer and the hidden layers of the TDNN. Speech features from the input layer of the FTDNN are individually extracted along the time-axis and the mel-scaled frequency-axis by each corresponding first hidden layer. The extracted features are then integrated into a single second hidden layer. The final decision is made based on the activation patterns whose property is invariant from both the time- and frequency-shift of input phoneme tokens.

Another novel architecture is a Block-Windowed NN (BWNN), based on windowing each layer of the NN with local time-frequency windows. This architecture makes it possible for the NN to capture global features from the upper layers as well as precise local features from the lower layers, because the local windows in the upper layers can integrate more global features than those in the lower layers. A five layered BWNN is constructed for a phoneme recognition experiment. These architectures yielded significantly better recognition performance than the original TDNN[20],[21].

6.2 Speaker-Adaptation Using Neural Networks

Speaker-adaptation is one good approach to a *speaker-independent* recognition problem. It is neccessary to use a small amount of training data uttered by an input speaker to adapt a speech recognition system. A speaker-adaptation technique using neural networks have been proposed[22]. It is also possible to use segmental speech for speaker-adaptation by building a mapping function from an input speaker to a standard speaker. We proposed a segmental approach using neural network identity mapping as a supervised learning method[23]. In this approach, segmental speech including a phoneme or syllable can be mapped between two speakers through a neural network and DTW matching method[22],[23]. This mapping network can be used as a front end of the TDNN-LR speech recognition system[29]. This technique is being applied to other phoneme categories including all consonants and phonemes. Also, an unsupervised speaker-adaptation technique using neural networks is being investigated[24].

6.3 Speaker-independent Recognition

In this section, we compare several TDNN architectures applied to speaker-dependent and multi-speaker's phoneme recognition with respect to their capabilities in a *speaker-independent* recognition problem.

We verified performance of several architectures: (1) single TDNN, (2) SID (Stimulus Identification) network, (3) Meta-Pi network, (4) Modular TDNN and (5) Modular Speaker ID network, where the single TDNN is an original architecture, the SID network is constructed by both each speaker's module and a speaker ID module which selects outputs in each speaker's module, the Meta-Pi network is reported as the network most suitable for *multi-speaker* phoneme recognition[25]. However, it has not been demonstrated how the Meta-Pi network is effective for a *speaker-independent* phoneme recognition problem. Furthermore, two novel modular TDNN architectures ((4) & (5)) are proposed to improve the performance. The modular TDNN is a network which is constructed by integrating each speaker's module (i.e., a single TDNN) trained on the first stage, and retrained on the second stage to recognize each phoneme, regardless of training speakers. The Modular Speaker ID network comprises of a speaker ID module in addition to the Modular TDNN, thus explicitly classifying each speaker ID as in the Meta-Pi network.

Speaker-independent phoneme experiments for recognizing voiced stops /b, d, g/ using six and twelve training speakers showed high recognition rates of 92.1% for the modular TDNN and 95.6%, respectively for the Modular Speaker ID network. These results are significantly better than the rates of 82.0% and 85.9%, respectively for the Meta-Pi network. As a result, it is found that the Meta-Pi architecture suitable for *multi-speaker* recognition is not necessarily robust for a *speaker-independent* recognition task. The recognition rate for the Modular Speaker ID network nearly matches the *speaker-dependent* recognition rate of 98.0% for the single TDNN[26],[27].

7. Conclusion

We described an integration of speech recognition and language processing. The speech recognition part consists of the Large Phonemic Time-Delay Neural Networks (TDNNs) 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. In the case of continuous speech recognition, adaptive incremental TDNN training using a small number of continuous training tokens is found to be very effective for adaptation of speaking manners.

We also proposed several new connectionist approaches as extensions of the TDNN-LR speech recognition system: (1) two proposed NN architectures (FTDNN and BWNN) provided robust phoneme recognition performance on variations of speaking manner, (2) a speaker-adaptation technique can be realized using a NN mapping function between input and standard speakers and (3) the Modular Speaker ID architecture provided high phoneme recognition performance that nearly matches speaker-dependent recognition performance. These techniques should be implemented in the TDNN-LR method in the future.

Acknowledgement

The authors would like to express their gratitude to Dr. Akira Kurematsu, president of ATR Interpreting Telephony Research Laboratories, for his support, and to Mr. Shigeki Sagayama, Head of Speech Processing Department for his valuable discussions on this research. They are also indebted to the members of the Speech Processing Department at ATR, for their constant help in various stages of this research.

References

(1) Waibel A., Hanazawa T., Hinton G., Shikano K. and Lang K.: "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. on ASSP, **37**, 3, pp. 328-339 (March 1989).

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

(3) Sawai H., Waibel A., Haffner P., Miyatake M. and Shikano K.: "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, pp. 81-88 (June 1989).

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

(5) Waibel A., Sawai H. and Shikano K.: "Modularity and Scaling in Large Phonemic Neural Networks", IEEE Trans. on ASSP, **37**, 12, pp. 1888-1898 (Dec. 1989).

(6) Sawai H., Waibel A., Miyatake M. and Shikano K.: "Phoneme Recognition by Scaling Up Modular Time-Delay Neural Networks", IEICE Technical Report, **SP88**-105 (1988).

(7) Rumelhart D. E., McClelland J. E. and the PDP Research Group: "Parallel Distributed Processing", MIT Press (1986).

(8) Haffner P., Waibel A., Sawai H. and Shikano K.: "Fast Back-Propagation Learning Methods for Large Phonemic Neural Networks", European Conference on Speech Communication and Technology, pp. 553-556, Paris (Sept. 1989).

(9) Minami Y., Miyatake M., Sawai H. and Shikano K.: "Continuous Speech Recognition Using TDNN Phoneme Spotting and Generalized LR Parser", Proceedings of ASJ Fall Meeting, 3-1-11 (1989).

(10) Minami Y., Sawai H. and Miyatake M.: "Large Vocabulary Spoken Word Recognition Using Time-Delay Neural Network Phoneme Spotting and Predictive LR Parsing", Trans. IEICE, **J73-D-II**, 6, pp. 788-795 (June 1990).

(11) Miyatake M., Sawai H., Minami Y. and Shikano K.: "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks", IEEE Proc. ICASSP-90, S8. 10 (April 1990).

(12) Miyatake M., Sawai H. and Shikano K.: "Training Methods and Their Effects for Spotting Japanese Phonemes Using Time-Delay Neural Networks", Trans. IEICE, **J73-D-II**, 5, pp. 699-706 (May 1990).

(13) Sawai H.: "Effect of Incremental Training in the TDNN-LR Phrase Speech Recognition System", Proceedings of ASJ Fall Meeting, 2-P-11 (Sept. 1990).

(14) Sawai H.: "TDNN-LR Large-Vocabulary and Continuous Speech Recognition System", Proc. ICSLP-90, **2**, pp. 1349-1352 (Nov. 1990).

(15) Sawai H.: "TDNN-LR Continuous Speech Recognition System Using Adaptive Incremental TDNN Training", Proc. ICASSP-91 (May 1991), to be presented. S2.4, pp. 53-56 (May 1991).

(16) Tomita M.: "Efficient Parsing for Natural Language—A Fast Algorithm for Practical Systems", Kluwer Academic Publishers (1986).

(17) Kita K., Kawabata T. and Saito H.: "HMM Continuous Speech Recognition Using Predictive LR Parsing", IEEE, Proc. ICASSP-89, S13. 3 (May 1989).

(18) Myers C. S. and Labiner R.: "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", IEEE Trans. ASSP, **29**, 2, pp. 284-279 (1981).

(19) Kuwabara H., Takeda K., Sagisaka Y., Katagiri S., Morikawa S. and Watanabe T.: "Construction of a Large-Scale Japanese Database and Its Management System", IEEE Proc. ICASSP-89, S10b. 12 (May 1989).

(20) Sawai H.: "Frequency-Time-Shift-Invariant Time-Delay Neural Networks for Robust Continuous Speech Recognition", Proc. ICASSP-91 S2.1, pp. 45-48 (May 1991).

(21) Sawai H.: "Time-Frequency-Shift-Tolerant Time-Delay Neural Networks", Proc. ASJ Fall Meeting, 2-P-2 (Sept. 1990).

(22) Iso K., Asogawa M., Yoshida K. and Watanabe T.: "Speaker Adaptation Using Neural Network", Proc. of ASJ Spring Meeting, 1-6-16 (March 1989).

(23) Fukuzawa K., Sawai H. and Sugiyama M.: "Speaker Adaptation Using Identity Mapping by Neural Networks", Proc. ASJ Fall Meeting, 1-8-16 (Sept. 1990).

(24) Sugiyama M., Fukuzawa K., Sawai H. and Sagayama S.: "Unsupervized Training Methods for Set Mappings Using Neural Networks", Proc. ASJ Fall Meeting, 2-P-10 (Sept. 1990).

(25) Hampshire J. and Waibel A.: "The Meta-Pi Network: Connectionist Rapid Adaptation for High-Performance Multi-Speaker Phoneme Recognition", Proc. 1990 IEEE International Conference on Acoustics, Speech and Signal Processing, S3. 9, pp. 164-168 (1990).

(26) Nakamura S. and Sawai H.: "Speaker-Independent Phoneme Recognition Using Time-Delay Neural Networks", ATR Technical Report, TR-I-0178 (Sept. 1990).

(27) Nakamura S. and Sawai H.: "A Preliminary Study on Neural Network Architectures for Speaker-Independent Phoneme Recognition", IEICE Technical Report, **SP90**-61 (Dec. 1990).

(28) Waibel A.: "Connectionist Large Vocabulary Speech Recognition", Trans. IEICE, **J73-D-II**, 8, pp. 1122-1131 (Aug. 1990).

Hidefumi Sawai was born in Kobe on Aug. 23, 1954. He graduated from Dep. of Electrical Engineering in Faculty of Engineering of Keio University in 1977, and also graduated from Doctor course of Keio University in 1982. He joined Ricoh Company in 1983 and has been engaged in research and development of speech recognition, pattern processing in the Research & Development Center. From 1988, he has been working for ATR Interpreting Telephony Research Laboratories where he is studying speech recognition using neural networks. He is currently a senior researcher and Dr. of Engineering. Dr. Sawai was a visitting researcher for Carnegie Mellon University, P. A., in the United States in 1989 and 1990. He is a member of the Editorial Commitee of the Information Processing Society of Japan, currently a member of Acoustical Society of Japan, IEEE, Acoustics, Speech and Signal Processing Society, Neural Network Society and Japan Neural Network Society.

Yasuhiro Minami graduated from Dep. of Electrical Engineering in Faculty of Science and Technology of Keio University in 1986. Currently he is in a doctorial course of the same university and at the same time, he was an intern student of ATR Interpreting Telephony Research Laboratories. His interest mainly includes speech recognition. He is a member of Acoustical Society of Japan.

Masanori Miyatake graduated from Dep. of Electrical Engineering in Faculty of Engineering of Kobe University in 1980. He joined Sanyo Electoric Company in the same year. From 1986 to 1989 he worked for ATR Interpreting Telephony Research Laboratories. Currently he is working for Information and Communication Systems Research Laboratories of Sanyo Electric Company. Since joining Sanyo Company, he has been engaged in speech recognition and synthesis. He is currently a senior researcher and a member of Acoustical Society of Japan.

Alex Waibel received the B. S., M. S. and Ph. D. degrees in Electrical Engineering and in Computer Science in 1979, 1980 and 1986, from MIT, and from Carnegie Mellon University, respectively. In 1986, he joined the faculty of the Computer Science Department as a Research Associate and is now a Research Computer Scientist. From May, 1987 to July, 1988, he has worked as Invited Research Scientist as the ATR Interpreting Telephony Research Laboratories in Japan. His current research interests include Speech Recognition and Synthesis, Neurocomputing, Machine Learning and Machine Translation. He is currently a member of the Technical Commitee of the IEEE ASSP Society, a member of the IEEE, ASSP Society, Computer Society, the Acoustical Society of America, the Association for Computational Linguistics and the International Neural Network Society.

Kiyohiro Shikano received the B. S., M. S. and Ph. D. degrees in electrical engineering from Nagoya University in 1970, 1972 and 1980, respectively. From 1972, he has been working at NTT Laboratories, where he has been engaged in speech recognition research. He is currently the Head of Speech Information Processing Department at NTT Human Interface Laboratories, where he is managing the research of speech recognition and speech coding. During 1984-1986, he was a visiting scientist in Carnegie Mellon University, where he was working on distance measures, speaker adaptation by codebook mapping, and statistical language modeling. During 1986-1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech recognition and speech synthesis research for interpreting telephone systems. He received the Yonezawa Prize from IEICE in 1975. He is a coauthor of "Speech, Auditory and Neural Network Models" (Series of Neuro Science & Technology, Ohm Press). He is a member of the Institute of Electrical and Electronics Engineers, Information Processing Society of Japan, and the Acoustical Society of Japan.