2—P—12

Phoneme Recognition by Modular Construction of Time-Delay Neural Networks*

Alex Waibel, Hidefumi Sawai and Kiyohiro Shikano
(ATR Interpreting Telephony Research Labs.)

1. INTRODUCTION

A number of studies have recently demonstrated [1-2] that connectionist architectures capable of capturing some critical aspects of the dynamic nature of speech, can achieve superior recognition performance for small but difficult phoneme discrimination tasks. A problem that emerges, however, as we attempt to apply neural network models to the full speech recognition problem is the problem of scaling. In this paper we demonstrate based on a set of experiments aimed at phoneme recognition that is indeed possible to construct large neural networks by exploiting the hidden structure of smaller trained subcomponent networks. A set of successful techniques is developed that bring the design of practical large scale connectionist recognition systems within the reach of today's technology.

2. SMALL PHONEMIC CLASSES by TIME-DELAY NEURAL NETWORKS

For the recognition of phonemes, a four 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). For detailed explanations, see ref.[1].The network is trained using the Back-Propagation Learning Procedure[3]. To evaluate our TDNNs on all phoneme classes, recognition experiments have been carried out for seven phonemic subclasses found in the Japanese database. For each of these classes, TDNNs with an architecture similar to the one shown in Fig.1 were trained. A total of seven nets aimed at the major coarse phonetic classes in Japanese were trained, including voiced stops B,D,G, voiceless stops P,T,K, the nasals M,N and syllabic nasals, fricatives S,SH,H and Z, affricates CH,TS, liquids and glides R,W,Y and finally the set of vowels A,I,U,E and O. Note, that each net was trained only within each respective coarse class and has no notion of phonemes from other classes yet. Table 1 shows the recognition results for each of these major coarse classes.

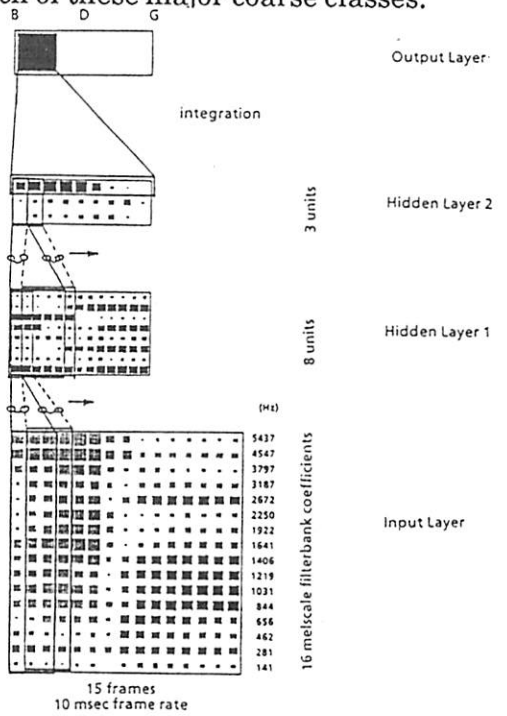

Fig. 1. The TDNN architecture (input : "BA")

3. SCALING TDNNs to LARGER PHONEMIC CLASSES

To shed light on the question of scaling, we consider the problems of extending our networks from the tasks of voiced stop consonant recognition (hence the BDG task) to the task of distinguishing among all stop consonants (the BDGPTK-task). Four experiments were performed for resolving that problem as follows:

(0) First attempt

Maximum activation performed only 60.5% correct.

*時間遅れ神経回路網のモジュール構成による音韻認識
アレックス ワイベル、 沢 井 秀 文、 鹿 野 清 宏
* (ATR Interpreting Telephony Research Laboratories)

Table 1. Recognition Results for 7 phoneme classes

phoneme	TDNN		
	#errors/ #tokens	%correct	total %
b	5/227	97.8	98.6
d	2/179	98.9	
g	2/252	99.2	
p	6/15	60.0	98.7
t	6/440	98.6	
k	0/500	100.0	
m	14/481	97.1	96.6
n	16/265	94.0	
N	12/488	97.5	
s	6/538	98.9	99.3
sh	0/316	100.0	
h	1/207	99.5	
z	1/115	99.1	
ch	0/123	100.0	100
ts	0/177	100.0	
r	0/722	100.0	99.9
w	0/78	100.0	
y	1/174	99.4	
a	0/600	100.0	98.6
i	1/600	99.8	
u	25/600	95.8	
e	8/600	98.7	
o	7/600	98.8	

(1) Exploiting the Hidden Structure of Subcomponent Nets resulted in 98.1% correct recognition.
(2) Class Distinctive Features(in Fig.2).
Adding the voiced/unvoiced(V/UV) distinction layers performed 98.4% correct.
(3) Incremental Learning by Way of "Connectionist Glue" performed 98.4% correct.
(4) All-Net Tuning achieved 98.6% correct.

Table 2 summarized the major results from our experiments. As a strategy for the efficient construction of larger networks we have found that the following concepts to be extremely effective: *modular, incremental learning, class distinctive learning, connectionist glue, partial and selective learning* and *all-net fine tuning*.

4. CONSONANT RECOGNITION by MODULAR TDNN DESIGN

The technique described in the previous section were applied to the task of recognizing *all consonants* (B,D,G,P,T,K,M,N,sN,S,SH,H,Z,Ch,Ts, R,W and Y). After completion of the learning run the entire net achieved a 95.0% recognition accuracy. All net fine tuning yielded 95.9% correct consonant recognition over testing data.

Table2.From BDG to BDGPTK: Modular Scaling Methods.

Method	bdg	ptk	bdgptk
Individual TDNNs	98.3 %	98.7 %	
TDNN:Max. Activation			60.5 %
Retrain BDGPTK			98.3 %
Retrain Combined Higher Layers			98.1 %
Retrain with V/UV-units			98.4 %
Retrain with Glue			98.4 %
All-Net Fine Tuning			98.6 %

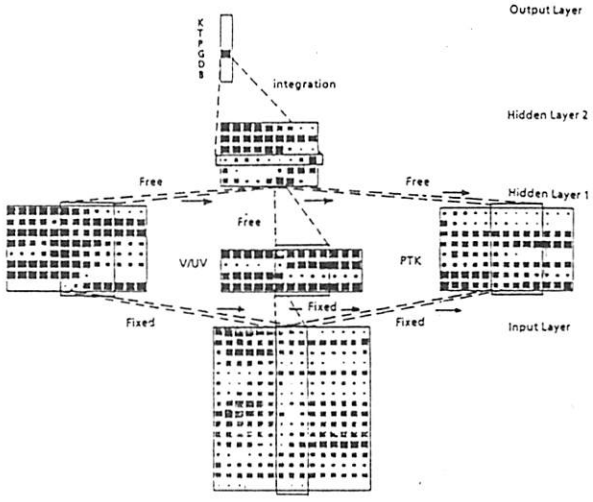

FIg. 2. Combination of a BDG-net, a PTK-net and a class distinctive Voiced/Unvoiced-net.

5. CONCLUSION

We have reported further experimental results from the use of Time-Delay Neural Networks (TDNNs) for recognition in all major phonemic categories in a large vocabulary speech database and have measured excellent recognition performance. Modular design is achieved by several important strategies. This technique was applied to the construction of a large TDNN aimed at the recognition of all consonants which performed 95.9% correct .

REFERENCES

[1]A.Waibel et al. Technical Report TR1-0006, ATR Interpreting Telephony Labs.Oct. 1987.
[2]R. Watrous, PhD thesis, Sep. 1988.
[3]D.E.Rumelhart et al. PDP, MIT Press(1986).