SPEAKER - INDEPENDENT COMPOUND VOWELS RECOGNITION USING NEURAL NETWORK BASED ON PHONETIC KNOWLEDGE

Cai Dehe

面向语音学基于音素的利用 NN 识别非特定人复合元音的研究

蔡德和

[提要] 本文主要讨论利用 NN 对非特定人的复合元音的识别，本文所述系统的特点是：NN 的输入层刺激采用语音的功率谱，并在复合元音的识别过程中运用了语音学知识。

本系统对复合元音的识别分两个层次完成的。第一层次是对音素的识别；利用 NN 将自然语音中的各音素映射为与其对应的规一化的共振峰模式，从而解决了音素的特征多变性问题。然后将规一的共振峰模式转换为相应的音素符号。因为在自然语音中复合元音也具有多变性，也就是说组成某个复合元音的音素序列会因人而异，所以第二层次是在音素符号识别的基础之上再对复合元音识别，就要运用语音学知识将第一层次得到的多变的音素符号串转变为相应的复合元音符号。从而解决多变的复合元音的识别问题。

目前大部分 NN 的语音识别系统是以倒谱系数作为 NN 的输入刺激，多以音节或词作为识别单元。但本系统是以语音的功率谱作为 NN 的输入刺激，这样有利于运用语音学知识，因为语音学是建立在对语音功率谱的分析、研究之上的。本系统的识别单元选用的是音素，如果加上辅音和鼻尾音的识别在原则上就能实现全部音节的语音识别。另一方面音节的切分工作可以放在音素识别完成之后，由于已有音素识别的结果，就可以利用语音学知识帮助切分，使切分的正确率提高。还有汉语普通话的音素只有十几个，比音节数(1200 个左右)要少得多，这对缩小 NN 的规模和加速 NN 的收敛都非常有利。

NN 的训练用的是单元音，然后用这单元音训练的网络识别未经训练的复合元音。运用了语音学知识之后，使复合元音的识别率从 54% 提高到 90%。

ABSTRACT

In this paper, a strategy for independent speaker of recognizing the compound vowels of Standard Chinese (SC) using neural network (NN) based on phoneme was primarily discused. The stimuli to NN input layer were the speech power spectra, and the outpt layer was represented by the phonemes on the acoustic vowel chart. The NN trained by monophthongs were used to recognize the compound vowels. Since the phonetic knowledges were added to analyze the result of initial recognition that had not to be trained, the accuracy for recognizing of the compound vowels had greatly imployed from 54% to 90%.

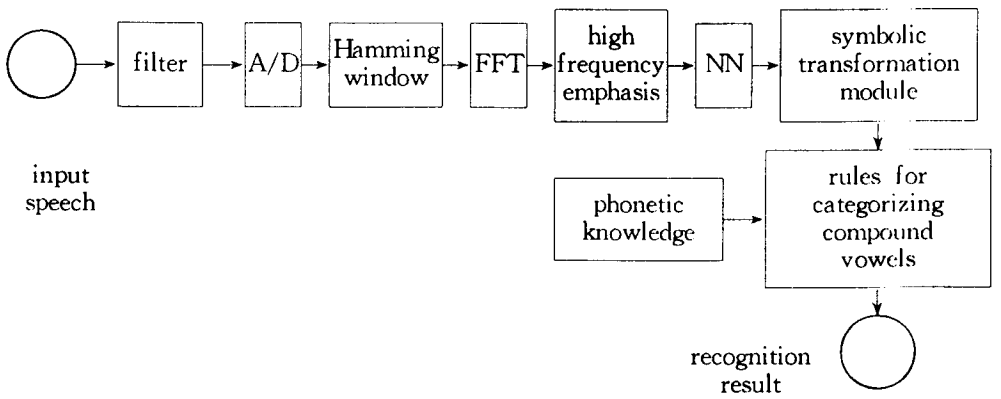
I. INTRODUCTION

In the purpose of resolving the problem of the recognition for the variability of phonemes in SC, the recognizing of the compound vowels was accomplished in two steps. The first step was the recognition of phonemes, for which NN is used to mapping the phonemes of natural speech with their corresponding normalized formant models, by which the problem of the variabilities of phonemic features was solved. Then the normalized fromant models are transformed into corresponding phonemic symbols. The second step was the recognition of compound vowels based on ponetic symbols which are varied very much in natural speech, also their phonemic sequence constructed by compound vowels are also varied according to different speakers. So that the phonetic knowledge had to be added in order to transform the various phonemic symbols obtained from the firat step into the symbols of corresponding compound vowels. Thus the problem of the variability of compound vowels can be solved.

Most of the speech recognitions by NN system were using the cepstrum coefficient for the input stimuli now a days, and also syllables or words are used as the recognized units. However, in this paper, speech power spectrograph was used as the input of NN and phonemes as recognizind units among the phonemic recognition steps. If there are consonants and nasals recognitions being added, the recognition of unlimited vocabulary can be realized. Moreover, segmentation of syllables can be done after the phonemic recognition being accompished. Thus, the segmentation process may take advantage of the phonetic knowledge to make the segmentation highly explicit, for the results of phonemic recognition had been obtained already. On the other hand, that around ten monphthongs in SC is far less than the number of syllables which are around 1200. This is an adventage for reducing the structure or quichkening the convergence of NN.

II. SYSTEMATIC STRUCTURE AND ITS PRINCIPLE

The block diagram of this system is illustrated as below:



In this system, sampling frequency was 10 KHz with a frame length of 25ms (256 points). A power spectrum of 128 points was obtained by FFT. The design of high frequency's gain was refered the auditory curve by Fletcher－Monsen. The power spectrum of 128 points was reduced to 64 points and then transfered to 64 nodes of the input layer and 22 hidden nodes, to form 26 output nodes of BP network. The NN output layer was divided into three parts in which each one was corresponding one of the three formants(F1, F2 and F3). In the output layer , the difference of the positions of activated nodes represented the different frequency of the output formants. Diffenrent vowel formants were transformed into normalized models of the three formants by the mapping of NN. This normalized models of the formants were obtained from the statistical data of SC vowels spoken by 15 male speakers.

This BP network was using the above mentioned normalized formant models of single vowels as the target values trained by single vowels/y/, /i/, /e/, /a/, /o/, /u/, /ə/, /ɚ/ spoken by 15 male speakers. Since the phonetic features of the sequencial phonemes are changed continuously, so the transition properties of the adjacent vowels the acoustic vowel chart are changed continuously. However, the output nodes in NN are independent from each other, it's easy to make mistakes in recognition or to reject to be recognized for these phonemes. As a diphthong is concerned, it should be given a continuous change and a directional curve in the acoustic vowel chart. However, it presented the discontinuous effects when we used the above network.

Therefore we adopted the overlapping modes to determine the target values of the output modes, i.e., between the contiguous phonemes there were output nodes of the target phonemes and common nodes of the overlapping ones also. The transition properties of contiguous phonemes were then presented by these common nodes. The recognition of trained NN was 90.3% for the vowels spoken by independent male speakers.

In this system, the NN trained by monophthongs was used to recognixe these compound vowels. In the procedure of recognition, the pre－processing was similar to the training of speech signals. Since the frame length was 25 ms, and was approximated to monophthongs, can be analysed by FFT. Any compound vowel after a NN recognition cauld obtain the data of normalized formants. These data were arranged according to time alignment of 20ms interval. The data of the frames of normalized formants were processed by symbolic transformation module. The principle of minimum distance in three dimension A. V. C. was used in the symbolic module, in which the nomalized data of formants were transformed into vowel's symbols by 1st, 2nd, and 3rd choises, and their distance between the practical phonemes and the displayed symbols in the A. V. C. of three dimension of F1, F2, and F3. The normalized formant data were transformed to symbols through the symbolic transformation module.

Since every frame was corresponding to one symbol while one compound vowel was included around ten frames, it was obtained that some ten vowel symbols were placed accord-

ing to the time sequence. In the symbolic sequence, the same successive symbols were only denoted as one symbol.

Since the signals of compound vowels were varied with time and also with the habit or circumstance of the speaker. The change of compound vowels was more complex than that of single ones.

In the phonetic domain, the complexity of the compound vowels for which the aspects of acoustic manifestation, and the articulatory physiology had been thoroughly studied. Had this knowledge used in the recognition, the accuracy for recognizing of the compound vowel were greatly increased.

In this paper, we took the diphthong /ai/ and the triphthong /iou/ for example, in order to show how to use the phonetic knowledge for the increasing of recognizing accuracy.

Firstly, the analysing of /ai/ , which is a diphthong with the vocalic ending "－i" in /ai/ is actually lower and backward in tongue position than its corresponding vowel of monophthong /i/ and approaching to /e/. In SC, there are few distinctions in lexical perception between /ai/, /æ/ and /e/ for they are usually articulated in spoken speech with undershoot more or less. So the contribution of these three "finais" (A Chinese phonological term for part other than consonant in a syllabie) in speech perception might be the same, and can be categorized into one type. After this treatment , the accuracy for recognizing the /ai/ had been increased from 75% to 98%.

Secondly, the analysing of /iou/ , which is a triphthong with ascendig sonorant . The vocolic ending "u " in this compound is usally articulated with less rounded and bacdward than the 2nd vowel or even with the same tongue place as the second vowel /－o－/. Also the tramsition between the vocalic initial /i－/ and the second vowel /－o－/ may be articulatell as an /ə/. So there are no difference in lexical distinction between /iou/, /io/, /iəou/ and /iəo/. Thus these for compounds vowels could be classified into one type.

In this paper, same method was applied to the other compound vowels of SC. The rules of category for each compound vowel were given in the following table.

The ranges and the trajectories of these compound vowels were shown in acoustic vowel chart as the following:

compound vowels	/ai/	/ao/	/ou/	/ia/	/ie/	/ua/	/uo/
allophone	/ae/	/aou/	/əu/	/ea/		/oa/	/uə/
	/e/	/o/					

compound vowels	/iao/	/iou/	/uai/	/uei/
allophone	/iaou/	/iəu/	/uae/	/uəi/
		/io/	/ue/	
		/iəo/		

The locus and the diretions of compound vowels in A.V.C. see Fig.1.

The recognizing results of some speakers (not included the speakers tranined) were given as follows: 54% for the case of that the phonetic knowledge had not been utilized, while 90% for that had been utilized.

In natural speech, the articulation of compound vowels was affected by many features, i.e., the tempo, rythem, coarticulation and habit.

As there is a main requirement that the semantic confuse it should be avoid the in natural seecj, however, the compound vowels were presented many allophones. That's just the question that the accuracy of the recognition for compound vowels had to be increased by using the phonetic knowledge. Of cause there were different rules for vowels categorizing had been used in different dislects and languages.

There was another question: the accuracy of recognition for single vowels was 90.3%, why does that for compound vowels based on phonemes reach as high as high as 90%? It is the reason that the information bearing of compound vowels is greater than that of a phoneme. If this information is fully applied, it is possible that the accuracy of recognilion for the compound vowels might be reached or even surpassed that for the phoneme. There were the same cases for the recognition system based on syllebles , in which if the gramatical ruls would be applid, the accuracy of recognizitic of word and setence should be high than that of syllebl.

III. DISCUSSION:

1. It is feasible that the compound vowels were recognized by NN based on phoneme, in which the speech power spectra were used as its input parameters.

2. The compound vowels were far well recognized when the phonetic knowledge based on phoneme had been added to the NN recognizing system based on phoneme.

3. In the prupose of recognizing the transitions from voiceless consonants or nasal ones to vowels , the discrimination - capacity to phoneme has to be improved in the further works.

REFERENCES

Wu, Zongji, (1992), (edi.), A Fundamental Course of Modern Chinese Phonetics, Foreign Languages Printing House, (in Chinese).

Wu, Zongji & Lin Maocan, (1988), An Outline of Experimental Phonetics, Higher Education Press, Beijing (in Chinese)

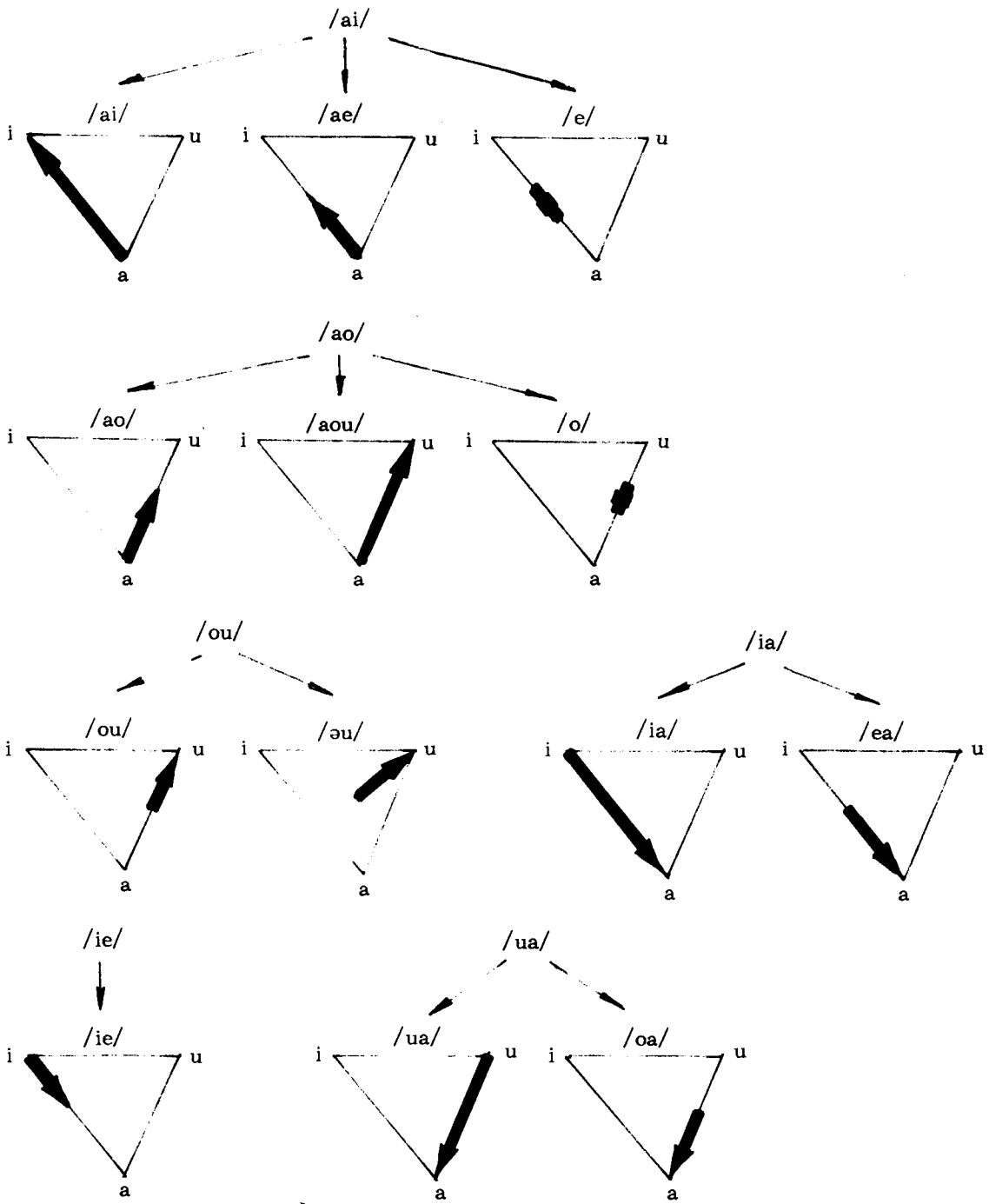
Cao, Jianfen & Yang, Shun'an (1984), "An experimental inversigation: diphthongs in Standard Chinese." Zhongguo Yuwen, 1984, No. 6:426 - 433.

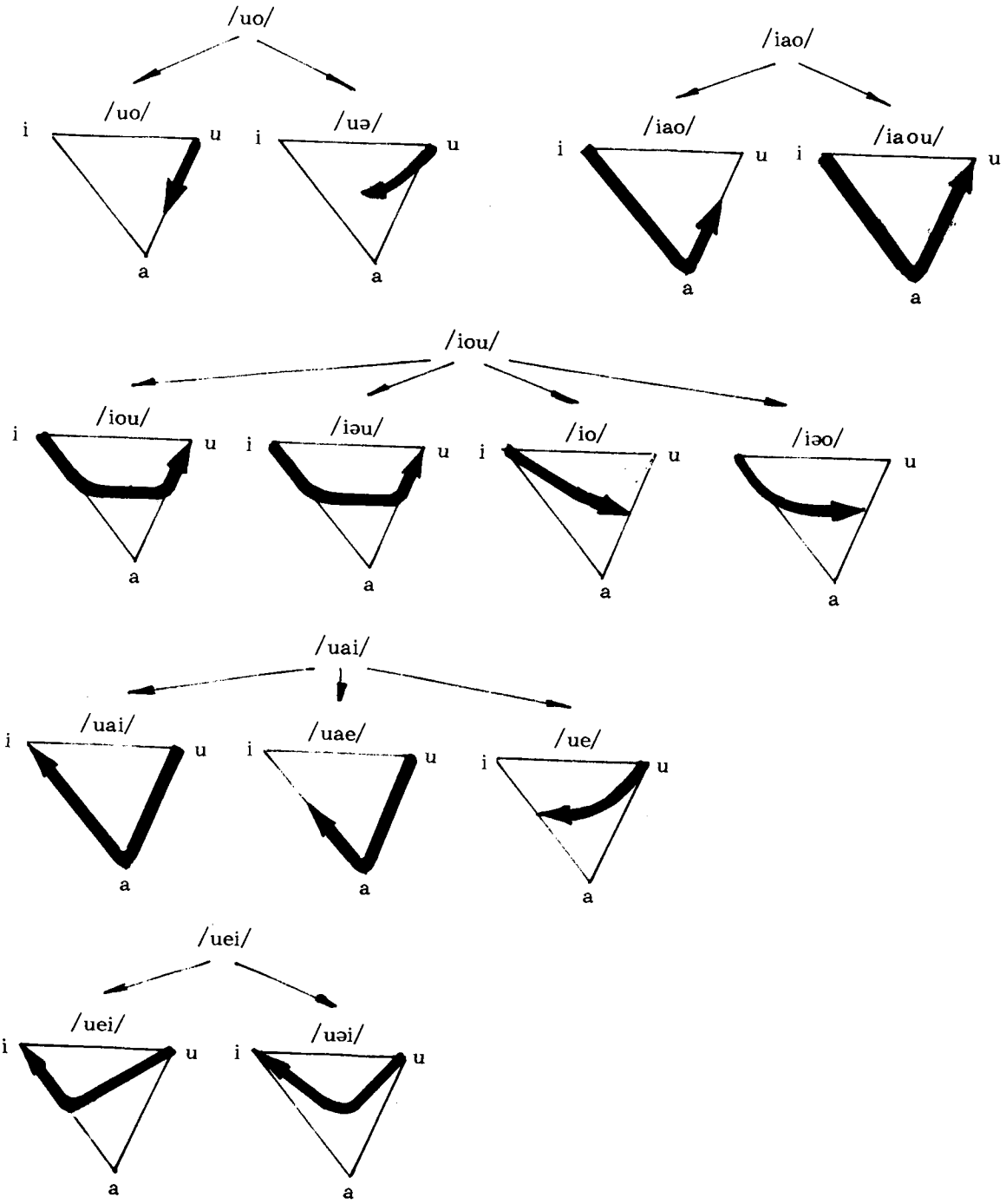
Liu, Ruiting & Dai, Ruwei(1991), Artifitiol neural network system desing. A Collaction of the Articles on Artifitial Neural Network and Its Applications. Institute of Automation Chinese Academy of Sciences.

Jiao Licheng (1990), Systemic Theory of NN, Press of Xi - an Electrinic Technology University, Xi - an, 1990. (In Chinese)

Jeffre :. Elman. Davod Zosr (1988), "Linearing the hidden structure of speech", Tje journal of the acoustical Society Society of America, 1988, Vol. 1. 83 No. 4

FIG.1 The acoustic manifestation of the compound vowels and their "allophone" :




/uo/
/uo/
/uə/
/iao/
/iao/
/iaou/
/iou/
/iou/
/iəu/
/io/
/iəo/
/uai/
/uai/
/uae/
/ue/
/uei/
/uei/
/uəi/
i
u
a