SPEECH RECOGNITION EXPERIMENTS WITH THE SYLLABLE INVENTORY OF STANDARD CHINESE

Michael Wagner
Department of Computer Science
University College / ADFA
University of New South Wales

ABSTRACT - This papers explores the possibility of using automatic speech recognition as a front-end to a computer for Chinese character processing. A speech recognition experiment has been performed on the complete inventory of second-tone words of Standard Chinese. 2 recordings which were made 48 hours after one another were used as test and reference sets. Distances within word clusters are shown to frequently exceed inter-cluster distances for the inventory of 260 syllables.

INTRODUCTION

There are 2 basic methods of entering Chinese characters into a machine by means of a keyboard: Firstly, a keyboard can be constructed on which each character is represented by a separate key. Such keyboards therefore have several thousand keys in order to process Chinese text. Secondly, an ASCII-type keyboard can be used to encode each input character as a sequence of several keystrokes. Both these basic methods have been and are being used for the computer input of Chinese characters.

Keyboards with a large number of keys not only present the obvious problems of size and weight but also the unsolved problem of arranging the keys in such a way as to enable users to find the required keys without extensive searching of the keyboard. Encoding Chinese characters by means of a ASCII-type keyboard raises a different set of problems which have so far prevented this method from being fully accepted by the Chinese community (Chen & Gong, 1984).

As an alternative method of Chinese character input, it has been suggested that an automatic speech recognition system might be used to recognise the spoken equivalent of each character (Wagner, 1986). This approach would allow the user of the system to speak isolated words into a microphone and select the corresponding character by an interactive process in a word processing environment.

CHINESE LANGUAGE

Each Chinese character corresponds to 1 monosyllabic word of spoken Chinese. There is, however, considerable ambiguity in the mapping between characters and spoken syllables. There are many sets of different characters which are pronounced identically or, in reverse, one spoken syllable

1st tone:	336 words
2nd tone:	265 words
3rd tone:	333 words
4th tone:	357 words
Total:	1291 words

TABLE 1. Chinese vocabulary arranged by tones.

usually represents more than one written character and can represent as many as 20 or 30 different characters. Such ambiguities are usually resolved easily by the native Chinese listener by reference to the context. The resulting inventory of spoken syllables therefore comprises only about 1300 syllables which can be divided into 4 subinventories for the 4 different tones (1) as shown in Table 1.

According to the syllable formation rules governing the Chinese language (Dow, 1972), there are 22 syllable-initial consonants and 36 syllable-final vowels, vowel-nasal compounds and vowel-retroflex compounds not every one of which occurs in every tone. Table 2 shows initials and finals in both IPA and Hanyu Pinyin transcriptions:

GROUP	IPA	Hanyu Pinyin
Zero	-	-
Labials	[b], [ph], [m], [f]	b, p, m, f
Dental Sibilants	[dz], [tsh], [s]	z, c, s
Alveolars	[d], [th], [n], [l]	d, t, n, l
Retroflexes	[dz], [tsh], [s], [z]	zh, ch, sh, r
Prepalatals	[dz], [tch], [c]	j, q, x
Velars	[g], [kh], [c]	g, k, h

TABLE 2a. The 22 syllable-initials in IPA and Hanyu Pinyin transcriptions.

GROUP	IPA	Hanyu Pinyin
Open /a/	[ɑ],[œI],[an],[ɑŋ],[ɑU]	a,ai,an,ang,ao
Open /ə/	[ɣ],[eI],[ən],[əŋ],[oU],[ər]	e,ei,en,eng,ou,er
Spread /a/	[ẓ,ʐ ,i],[Iɑ],[Iæn],[Iɑŋ],[IɑU]	i,ia,ian,iang,iao
Spread /ə/	[IE],[in],[iŋ],[ɪoU]	ie,in,ing,iou
Rounded /a/	[u],[Uɑ],[UæI],[Uan],[Uɑŋ]	u,ua,uai,uan,uang
Rounded /ə/	[Uɣ],[UeI],[Uən],[Uəŋ]	ue,uei,un,ung
Inner-rounded/a/	[y],[Yan]	u, uan
Inner-rounded/ə/	[YE],[yn],[Yəŋ]	ue,un,ung

TABLE 2b. The 36 syllable-final vowel clusters in IPA and Hanyu Pinyin transcriptions.

It has been shown elsewhere, e.g. by Chen & Pao (1985), that the tone of a syllable can be determined reliably. Therefore, the word recognition task for the complete vocabulary of Chinese reduces to subtasks with reference sets of 336, 265, 333 and 357 words respectively.

A closer inspection of Table 2 shows that there are many pairs of syllables which can be expected to be acoustically and parametrically very close. Since it is well known that even small vocabularies of 10 to 50 words can produce low recognition rates if they contain word pairs that are very similar, for example the English digits and letters of the alphabet, the question arises whether standard isolated-word recognition techniques can be successfully employed to perform speech recognition on the given syllable inventory. The following experiment is designed to shed some light on the parametric distances between syllables in one of the tone inventories.

SPEECH DATA RECORDING AND PROCESSING

The entire vocabulary of spoken Chinese has been recorded by 2 male and 1 female speakers of the Beijing dialect of Standard Chinese. Each speaker recorded the vocabulary of 1291 words 5 times in sessions spaced between 2 days and 1 month. Speakers read the words from reading cards arranged in either random order or Hanyu Pinyin alphabetical order with each word read in all its possible tones sequentially. Recordings were made in a reasonably quiet room with a good-quality cassette recorder.

Recordings were sampled at 16 kHz after low-pass filtering at 7.6 kHz. Samples were 12-bit quantised using a Data Translation analog-to-digital conversion board with an IBM PC-AT microcomputer. Word boundaries in the resulting data files were determined automatically (Rabiner & Sambur, 1982).

For this experiment, 2 randomly recorded sessions by a male speaker spaced 2 days apart were selected. The speech was processed using a Hamming window of length 32ms with an overlap of 16ms thereby producing a frame rate of 62.5 frames per second. Linear prediction analysis of order p=18 was performed on the data (Wagner & Fulcher, 1986).

NUO2	LUO2	0.000	3	1.05	DI2	NI2	0.043	4	2.06
ZHUO2	RUO2	0.000	1	0.52	DAO2	AO2	0.044	2	0.52
NG2	M2	0.007	2	0.69	DUO2	GUO2	0.045	3	1.89
DI2	BI2	0.008	3	0.95	BU2	GU2	0.046	4	1.72
ZHU2	RU2	0.012	2	0.76	MU2	NU2	0.046	2	0.68
BO2	GUO2	0.015	3	1.33	M2	NG2	0.047	3	1.60
BA2	DA2	0.018	2	1.14	NG2	LENG2	0.047	4	1.55
DU2	BU2	0.018	2	0.53	DIE2	BIE2	0.047	2	1.00
BO2	O2	0.019	3	1.06	GEN2	REN2	0.048	6	4.08
O2	HUO2	0.022	3	1.06	O2	MOU2	0.049	5	2.57
BU2	GUO2	0.029	5	1.53	NI2	LI2	0.049	4	2.33
GU2	MU2	0.029	5	2.59	NI2	MI2	0.050	5	2.15
DAO2	MOU2	0.030	4	1.86	DAO2	MAO2	0.050	3	1.25
O2	BO2	0.033	2	0.71	DUO2	RUO2	0.051	3	1.86
BO2	HUO2	0.033	2	0.89	PANG2	HANG2	0.052	5	2.22
AI2	MAI2	0.036	3	1.91	ZHAI2	LAI2	0.052	2	0.96
GU2	WU2	0.036	5	2.82	BIE2	LIE2	0.056	4	2.10
HUN2	WEN2	0.038	1	0.38	DU2	NUO2	0.057	4	1.58
NANG2	LANG2	0.038	4	1.18	GU2	BU2	0.057	3	1.00
LU2	NU2	0.039	3	0.86	QI2	PI2	0.057	6	3.38
NU2	MU2	0.039	2	0.61	DAO2	LAO2	0.058	4	2.18
LUO2	NUO2	0.039	3	1.33	WEN2	HUN2	0.058	1	0.32
FENG2	NENG2	0.040	5	2.20	FENG2	LENG2	0.058	5	1.95
O2	LUO2	0.041	3	1.42	FENG2	NG2	0.059	5	2.20
BO2	MO2	0.042	3	1.24	PO2	TUO2	0.061	3	1.33

Table 3. Average distance, maximum deviation from diagonal warping function in frames and mean deviation from diagonal warping function in frames for the 50 closest word pairs.

The reference patterns consisted of the modified linear prediction coefficients for each frame of each reference syllable. Distances between test and reference frames were computed according to Itakura (1975). Time warping between test and reference patterns (Sakoe & Chiba, 1978) was restricted to a maximum time difference of 6 frames (=96ms).

WORD DISTANCES

The first recognition run used all 260 second-tone words of session 1 as the reference set. The same set of words was then used as a test set and the average interword distance was measured for all 260*(260-1) = 67,340 pairs of different words.

Out of the total of 67,340 word pairs, table 3 shows the 50 word pairs with the smallest interword distances. It can be seen that for most close word pairs the final vowel or vowel-nasal cluster is identical while the initial cluster differs by 1 feature, e.g. frication-plosion, labial-alveolar etc.

In contrast, table 4 shows the 50 word pairs which are separated the most clearly. Most of these pairs feature the contrast between /i/ and /a/ vowels which is clearly recognised by the distance measure used in the experiment.

PA2	QI2	1.947	6	3.12
A2	TI2	1.953	4	1.58
PA2	XI2	1.958	6	3.22
MANG2	JI2	1.958	4	1.37
LA2	XI2	1.961	6	2.52
GA2	JI2	1.965	5	2.28
A2	XI2	1.967	6	2.73
RAO2	JI2	1.968	5	2.11
A2	QI2	1.969	6	2.77
LANG2	QI2	1.971	6	2.62
MA2	XI2	1.974	4	1.74
WANG2	QI2	1.982	6	2.04
DA2	XI2	1.989	6	3.76
WA2	YI2	2.000	5	1.75
MA2	YI2	2.004	4	2.00
RAO2	XI2	2.008	4	1.70
BAO2	JI2	2.010	4	2.11
MA2	QI2	2.013	4	2.12
A2	YI2	2.018	6	3.43
NA2	PI2	2.019	6	2.57
DA2	YI2	2.022	6	3.33
RUA2	JI2	2.026	5	2.53
WA2	PI2	2.037	5	2.68
HUA2	TI2	2.046	6	2.90
DA2	TI2	2.058	5	2.32
A2	JI2	2.061	6	2.89
BA2	TI2	2.067	3	1.40
AO2	JI2	2.069	3	1.53
BA2	PI2	2.077	6	2.36
WA2	XI2	2.083	6	3.18
WA2	TI2	2.084	4	1.50
MA2	JI2	2.088	6	3.44
NAN2	XI2	2.093	5	1.85
LA2	QI2	2.095	6	2.19
ER2	JI2	2.098	5	2.42
ER2	DI2	2.100	4	1.94
DA2	QI2	2.102	6	4.04
BA2	XI2	2.103	6	2.56
WA2	JI2	2.109	6	2.21
HUA2	QI2	2.114	6	3.77
NAN2	QI2	2.117	6	1.96
NA2	XI2	2.119	6	2.58
BA2	YI2	2.138	5	2.50
HUA2	JI2	2.143	4	1.44
RAO2	QI2	2.145	6	2.85
WA2	QI2	2.167	6	2.93
BA2	JI2	2.169	6	2.28
BA2	QI2	2.198	6	3.22
DA2	JI2	2.251	4	1.42
NA2	QI2	2.261	6	2.04

Table 4. Average distance, maximum deviation from diagonal warping function in frames and mean deviation from diagonal warping function in frames for the 50 most distant word pairs.

Word pairs with similar initial consonants and word pairs with similar nasal auslauts were then analysed further. Table 5 compiles the results of that analysis. All word pairs which are identical but for the initial consonant which was "c-" for the test word and "s-" for the reference word were compared and the average distance over this group of word pairs was determined. In this example, there was only one "c-" vs "s-" word pair with a distance of 0.132. Similarly, the "m-" vs "n-" group (10 word pairs), the "q-" vs "x-" group (12 word pairs) and the "p-" vs "f-" group (8 word pairs) show very small average interword distances.

Similarly, a group of word pairs which differed in final "-n" vs "-ng" was similarly analysed and the 29 word pairs of this group yielded an average interword distance of 0.315.

c- ... s-	n=1	ave=0.132	std=0.000
m- ... n-	n=10	ave=0.158	std=0.074
q- ... x-	n=12	ave=0.201	std=0.076
p- ... f-	n=8	ave=0.215	std=0.075
j- ... q-	n=5	ave=0.274	std=0.085
z- ... c-	n=4	ave=0.303	std=0.056
ch- ... sh-	n=7	ave=0.311	std=0.116
k- ... h-	n=4	ave=0.316	std=0.092
d- ... t-	n=4	ave=0.336	std=0.066
zh- ... ch-	n=6	ave=0.353	std=0.135
b- ... p-	n=7	ave=0.392	std=0.064
g- ... k-	n=1	ave=0.415	std=0.000
-n ... -ng	n=29	ave=0.315	std=0.095

Table 5. Interword distance distributions for groups of word pairs which differ only in their initial consonant or in their final nasal.

A second recognition run used the same reference set with a test set of second-order words which was recorded 2 days later. This run compiled the distances between the 260 identical word pairs. Intraword distances for this run varied from 0.123 (YA2) to 0.698 (GU02).

Figure 1 shows the distribution of interword distances as determined by Run1 in comparison with the distribution of intraword distances as determined by Run 2.

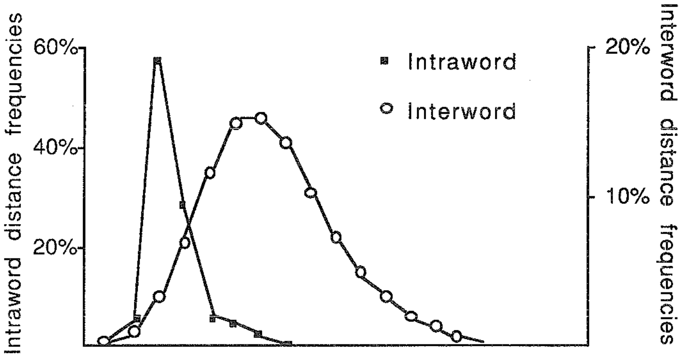

Figure 1. Frequency of interword distances (circles) and intraword distances (squares) between second-tone words.

The overlap between the 2 distributions shows clearly that the 260 word clusters are not well separated and that the recognition error rate for the entire second-tone syllable inventory will be considerable (Wagner et al., 1986).

CONCLUSION

It has been shown that the entire inventory of 260 Chinese second-tone syllables populates an 19-dimensional LPC parameter space very densely and that a pattern-matching approach using an LPC distance measure and dynamic time warping is unable to resolve all 260 word cluster sufficiently to yield user-acceptable recognition rates. Further experiments are currently underway to enhance cluster separation by including additional parameters into the feature space.

ACKNOWLEDGMENT

This project is supported by the Australian Research Grants Scheme (ARGS).

NOTE

(1) The neutral tone would be processed separately by the recognition system.

REFERENCES

CHEN, C.W. & GONG, R.W. (1984) "Evaluation of Chinese input methods", Comp.Proc.of Chinese & Oriental Lang., vol.1, no.4.

CHEN, C.H. & PAO, Y.H. (1985) "Computer recognition of Chinese language (Mandarin) homonyms", Comp.Proc.of Chinese & Oriental Lang., vol.2, no.1.

DOW, F.D.M. (1972) "An outline of Mandarin phonetics", Australian National University Press, Canberra.

ITAKURA, F. (1975) "Minimum prediction residual principle applied to speech recognition", IEEE Trans. Acoust. Speech and Signal Proc., vol. 23, no.1.

RABINER, L.R. & SAMBUR, M.R. (1975) "An Algorithm for Determining the End-points of Isolated Utterances", Bell Syst. Tech. J., 54, 297-315.

SAKOE, H. & CHIBA, S. (1978) "Dynamic programming algorithm optimization for spoken word recognition", IEEE Trans. Acoust. Speech and Signal Proc., vol. 26, no.1.

WAGNER, M., WANG, W., HO, H. & O'KANE, M. (1986) "Isolated Word Recognition of the Complete Vocabulary of Spoken Chinese", Proc. Int. Conf. on Acoust. Speech Signal Processing, Tokyo, 701-704.

WAGNER, M. & FULCHER, J. (1986) "An IBM-PC Based Speech Research Workstation", Proc. 1st Aust. Conf. on Speech Science and Technology, Canberra.