ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

SPONTANEOUS MANDARIN PRODUCTION: RESULTS OF A CORPUS-BASED STUDY

Shu-Chuan Tseng

Institute of Linguistics, Academia Sinica

ABSTRACT

This paper presents empirical results of a corpus-based study attempting to characterize linguistic features of spontaneous Mandarin, which has been difficult to obtain before due to the lack of suitable speech material. Starting from linguistic considerations, these results of word frequency as well as syllable frequency should provide important cues to spontaneous speech production. Frequent words or syllables need special investigations into their phonetic forms in real production. Examinations of syllable structures also show that the distribution of onset consonant, nucleus and coda consonant in syllables which are often used in spontaneous Mandarin is similar across different speakers. And results of a segmental analysis also clearly indicate the likelihood of a segment being produced in spoken Mandarin.

1. INTRODUCTION

Conventionally, linguistic studies mainly rely on field works to document the use of languages often with a research focus on pronunciation, lexicon and sentence grammar. With the database construction methodology developed in corpus linguistics, new approaches to analyzing spoken language have become possible recently. It has an essential influence on spontaneous speech studies, because due to limitations of data size and database management it has been difficult to investigate and model spontaneous speech using the traditional research methods. This paper uses a corpus of spontaneous data to examine Mandarin, which is spoken in Taiwan. What we report in this paper is a new attempt to obtain linguistic characteristic of spontaneous Mandarin. The results are primitive, but with a great potential to be developed into a deep and systematic understanding of spoken language production. Frequency of actively used words and syllables in spoken language provides useful cues to a correct lexical selection, when the available acoustic information is not clear enough to select words in the lexicon. A lexicon for speech recognition systems, probably similar to the mental lexicon of a speaker, needs to store different phonetic forms of words for instance reduction, assimilation and contraction [4]. It is not realistic to consider all phonetic variations of all words listed in a standard dictionary, so frequent words are no doubt the most important ones we need to take into account first. Word and syllable frequency as well as segmental analysis can be of great use and this information can be systematically obtained by using spoken corpora. In addition, for notation used in this paper, lexical tones in Taiwan Mandarin have four marked realizations: (1) high level tone, (2) rising tone, (3) contour tone and (4) falling tone and the unmarked neutral tone (5). Different Chinese dialects have different numbers of lexical tones associated with different melodic values [2]. Throughout this paper, we use Pinyin to transcribe Mandarin words.

2. DATA AND GENERAL STATISTICS

Mandarin Conversational Dialogue Corpus (MCDC) was collected at the Institute of Linguistics, Academia Sinica from 2000 to 2001 [3]. It consists of eight transcribed conversations between strangers. The recorded speech data has a total length of approximately eight hours (the corpus will soon be released for public use). Because no blanks are available in the writing system of Mandarin to separate individual words, we have to segment the transcripts into words first. In order to ensure that the segmentation results are consistent, the automatic word segmentation and tagging system developed by the Chinese Knowledge Information Processing Group at Academia Sinica [1] is adopted to automatically segment word boundaries and syntactically tag the segmented words. General statistics are listed in Table 1.

Table 1: General statistics of MCDC

Speaker	Sex	Age	Syllables	Word types	Word tokens	Syllable/word ratio
S-01	F	29	4,789	921	3,334	1.44
S-02	M	25	9,262	1,445	6,913	1.34
S-03	F	37	8,522	1,140	5,853	1.46
S-04	M	35	6,202	965	4,234	1.46
S-05	F	16	9,273	1,093	6,339	1.46
S-06	F	17	6,659	874	4,497	1.48
S-07	M	40	8,887	1,283	6,946	1.28
S-08	F	46	7,360	1,140	5,497	1.34
S-09	F	30	2,687	572	1,967	1.37
S-10	F	35	13,534	1,577	9,103	1.49
S-11	M	35	7,140	1,104	4,399	1.62

S-12	M	23	6,057	882	3,723	1.63
S-13	M	43	7,847	1,066	5,301	1.48
S-14	F	45	7,808	864	4,859	1.61
S-15	F	37	4,437	858	3,255	1.36
S-16	M	24	6,751	1,150	4,833	1.40
∑			**117,215**		**81,053**	**1.45**

All sixteen speakers (nine female and seven male) produced 117,215 syllable tokens in total which correspond to 81,053 segmented words. The average number of syllables per word is 1.45. This implies that modern Taiwan Mandarin is not a monosyllabic language, because a large number of polysyllabic words are actively used in spontaneous conversation. With regard to the vocabulary, a language in principle consists of a core lexicon which may most possibly be highly frequent function words and a peripheral, domain-specific set of words. Our data shows that the word token coverage of 40% is achieved by cumulating occurrences of the first thirty most frequent words. Furthermore, 50%, 60%, 70%, 80% and 90% of the produced word tokens are covered by token occurrences of the first 54, 106, 238, 581 and 1651 most frequent words, respectively. That is, if we can manage to understand the phonetic variations of the first fifty most frequent word occurrences and their acoustic correlates, the automatic speech recognizer built with this understanding will be able to recognize at least the half of the word tokens in conversation.

3. POS AND SYLLABLE FREQUENCY

Table 2 illustrates statistically that the majority of word tokens often used by Mandarin native speakers are mono- and disyllabic words irrespective of content or function words. Monosyllabic words make up 31.4% of the overall word types (averaged for all sixteen speakers), and with respect to word tokens, it is 56.4%. Disyllabic words, in contrast, make up 57.8% of the overall word types, but only 38.9% of the overall word tokens. This means, mono- and disyllabic words form the most substantial part of the core vocabulary used in the overall data. And the difference between mono- and disyllabic words is that high-frequency words tend to be monosyllabic words rather than disyllabic words. But around 60% of distinctive word types are disyllabic. This also provides a clear piece of evidence that modern Mandarin contains a great number of polysyllabic words, especially disyllabic words, and is not a monosyllabic language any more.

Table 2: Mono- and Disyllabic Words in MCDC

Speaker	Sex	Monosyll. Word types (%)	Monosyll. Word tokens (%)	Disyll. Word types (%)	Disyll. word tokens (%)
S-01	F	309(34%)	1,846(55%)	509(55%)	1,321(39%)
S-02	M	416(29%)	3,531(51%)	783(54%)	2,586(37%)
S-03	F	341(30%)	3,398(58%)	659(58%)	2,155(37%)
S-04	M	292(30%)	2,476(59%)	552(57%)	1,523(36%)
S-05	F	413(38%)	3,585(57%)	598(55%)	2,541(40%)
S-06	F	307(35%)	2,450(55%)	510(58%)	1,936(43%)
S-07	M	372(30%)	4,147(60%)	759(59%)	2,434(35%)
S-08	F	348(31%)	3,171(58%)	665(58%)	2,047(37%)
S-09	F	210(37%)	1,152(59%)	316(55%)	728(37%)
S-10	F	461(29%)	5,288(58%)	953(60%)	3,500(39%)
S-11	M	312(28%)	2,323(53%)	665(60%)	1,859(42%)
S-12	M	257(29%)	2,004(54%)	540(61%)	1,528(41%)
S-13	M	308(29%)	3,063(58%)	640(60%)	2,049(39%)
S-14	F	265(31%)	2,757(57%)	495(57%)	1,839(38%)
S-15	F	268(31%)	1,878(58%)	501(58%)	1,199(37%)
S-16	M	372(32%)	2,703(56%)	646(56%)	1,881(40%)
∑		**(31.4%)**	**45,772** (56.4%)	**(57.8%)**	**31,126** (38.9%)

3.1. Word and Syllable Frequency

The first fifty most frequently used words in the MCDC are listed in Table 3 with their percentage; detailed definitions of the part of speech tags cf. [1]. In Table 3, only sixteen frequent words are disyllabic, the others are monosyllabic. This once again proves our hypothesis that frequent words are more likely to be monosyllabic. These can be grammatical particles such as de5 (structure particle), ge5 (classifier), le5 (aspect particle) etc., personal pronouns (wo3, ni3, ta1) and frequent verbs such as shi4 (to be) and you3 (to have), shuo1 (to say). Disyllabic, frequent words are mostly frequent conjunctions such as jiu4shi4 (that is), ran2hou4 (then), suo3yi3 (so), ke3shi4 (but) etc. The first fifty most frequent word types make up 36.43% of all word tokens produced in the corpus. Varieties of these words in different sentential contexts (initial, medial or final) and their full and reduced forms (phonetic) need to be further investigated.

Table 3: Top 50 POS Words in MCDC

Rank	1	2	3	4	5	6	7	8	9	10
Word POS	de5 (DE)	wo3 (Nh)	shi4 (SHI)	dui4 (P)	ni3 (Nh)	ge5 (Nf)	you3 (V_)	jiu4 (D)	bu4 (D)	ta1 (Nh)
#	3307	3225	2648	1597	1447	1420	1368	1360	1328	1177
%	4.06	3.96	3.25	1.96	1.78	1.74	1.68	1.67	1.63	1.44
∑ %	4.06	8.01	11.26	13.22	15.00	16.74	18.42	20.09	21.72	23.16
Rank	11	12	13	14	15	16	17	18	19	20
Word POS	na4 (Nep)	yi1 (Neu)	jiu4 shi4 (Cbb)	dou1 (D)	hen3 (Dfa)	wo3 men5 (Nh)	zai4 (P)	shuo1 (VE)	ye3 (D)	hui4 (D)
#	1144	1086	1050	957	937	900	882	827	794	783
%	1.40	1.33	1.29	1.17	1.15	1.10	1.08	1.01	0.97	0.96
∑ %	24.56	25.90	27.19	28.36	29.51	24.27	25.35	26.36	27.34	28.30
Rank	21	22	23	24	25	26	27	28	29	30
Word POS	ran2 hou4 (D)	jue2 de2 (VK)	yin1 wei4 (Cbb)	le5 (Di)	zhe4 (Nep)	yao4 (D)	ren2 (Na)	jiang3 (VE)	suo3 yi3 (Cbb)	ke3 shi4 (Cbb)
#	741	679	675	646	626	559	487	482	482	463
%	0.91	0.83	0.83	0.79	0.77	0.69	0.60	0.59	0.59	0.57
∑ %	29.21	30.04	30.87	31.66	32.43	33.11	33.71	34.30	34.89	35.46
Rank	31	32	33	34	35	36	37	38	39	40
Word POS	qi2 shi2	ta1 men5	qu4	she2 me5	hai2	shi2 hou4	xian4 zai4	ta1	mei2 you3	zhe4 yang4

	(D)	(Nh)	(D)	(Nep)	(D)	(Na)	(Nd)	(Nh)	(VJ)	(VH)
#	431	429	425	414	412	410	405	399	346	340
%	0.53	0.53	0.52	0.51	0.51	0.50	0.50	0.49	0.42	0.42
∑ %	30.04	30.56	31.09	31.59	32.10	32.60	33.10	33.59	34.01	34.43
Rank	41	42	43	44	45	46	47	48	49	50
Word POS	zhong3 (Nf)	qu4 (VCL)	Gen1 (P)	zuo4 (VC)	ni3 (Nh)	bi3 jiao4 (Dfa)	hao3 (VH)	dui4 (VH)	na4 bian1 (Ncd)	kan4 (VC)
#	339	322	318	313	308	298	295	291	284	283
%	0.42	0.40	0.39	0.38	0.38	0.37	0.36	0.36	0.35	0.35
∑ %	34.85	35.24	35.63	36.02	36.39	35.01	35.37	35.73	36.08	36.43

Different from word frequency, Table 4 lists the first fifty most frequently used syllables in the MCDC irrespective of the associated lexical tones. These make up 62.4% of the overall syllable tokens. From the results summarized in Table 3 and 4, our observation is that frequent syllables are in principle frequent words. "Shi" for instance is the most frequently produced syllable in the whole corpus, with 6305 tokens, and is at the same time the third most frequently produced word shi4 (to be), with 2648 tokens. That is, in addition to shi4 (to be), 3657 word tokens have or contain the syllable "shi", for instance ke3shi4 (but, 463 occurrences) and qi2shi2 (in fact, 431 occurrence). Thus, it is important to model this specific syllable in different word contexts to be able to detect with which frequent word it is associated for a speech recognition system. Differently, "wo" is the second most frequent syllable with 4148 syllable tokens, but wo3 (I) has already 3225 word tokens and wo3men5 900 occurrences. This means that merely twelve "wo" syllable tokens are other words or part of other words. When a speech recognition system detects a syllable "wo", it is most likely to be the word wo3 or its plural form wo3men5. A interim conclusion is, word and syllable frequency needs to be considered in parallel, i.e. their relationship, to achieve an optimal modelling strategy for a speech recognition system, for instance for "shi" and "wo", different models should be constructed.

Table 4: Top 50 Syllables in MCDC

Rank	1	2	3	4	5	6	7	8	9	10
Syllable	shi	wo	yi	de	dui	you	jiu	bu	ta	ni
#	6305	4137	3889	3681	2864	2857	2742	2438	2040	1884
%	5.4	3.5	3.3	3.1	2.4	2.4	2.3	2.1	1.7	1.6
∑ %	5.4	8.9	12.2	15.3	17.8	20.2	22.5	24.6	26.4	29.0
Rank	11	12	13	14	15	16	17	18	19	20
Syllable	men	hou	zai	ge	zhe	hui	ke	xiang	hen	mei
#	1618	1598	1463	1459	1376	1341	1256	1200	1185	1174
%	1.4	1.4	1.2	1.2	1.2	1.1	1.1	1.0	1.0	1.0
∑ %	29.4	30.7	32.0	33.2	34.4	35.5	36.6	37.6	38.6	39.6
Rank	21	22	23	24	25	26	27	28	29	30
Syllable	wei	dao	Qu	zhi	hao	ye	yao	shuo	qi	zi
#	1103	1079	1051	1046	1028	998	995	961	951	942
%	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.8
∑ %	40.6	41.5	42.4	43.3	44.2	45.0	45.9	46.7	47.5	48.3
Rank	31	32	33	34	35	36	37	38	39	40
Syllable	hai	ji	guo	yang	le	ren	na	ran	dou	lai
#	939	930	928	922	921	921	902	899	891	859
%	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.7
∑ %	49.1	49.9	50.7	51.5	52.3	53.0	53.8	54.6	55.3	56.1

Rank	41	42	43	44	45	46	47	48	49	50
Syllable	jue	me	zhong	da	yin	xiao	jiao	xian	jia	li
#	817	815	789	788	759	728	725	689	684	669
%	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.6
∑ %	56.8	57.5	58.1	58.8	59.5	60.1	60.7	61.3	61.9	62.4

4. SPEECH SOUNDS OF MANDARIN

Given a language, different linguists may set up different systems of speech sounds by applying criteria of minimizing phonemic contrast. Our principle is to use the smallest number of symbols in our system. Thus, we have twenty two consonants and twenty five distinctive vowels and vowel combinations (including diphthongs). The alveolo-palatal group [ɕ], [tɕ] and [tɕʰ] can only be followed by [i] or [y] including diphthongs starting with [i] or [y]. The alveolar group [s], [ts] and [tsʰ], the retroflex group [ʂ], [tʂ], [tʂʰ] and [ʐ] and the velar ones [x], [k] and [kʰ] can be followed by any vowels in Mandarin except [i] and [y]. The coda position can only be occupied by [n] or [ŋ], where [ŋ] never occurs in the onset position. All Mandarin consonants are listed as follows.

bilabial		b:[p]	p:[pʰ]	m:[m]	
labiodental	f:[f]				
alveo-dental		d:[t]	t:[tʰ]	n:[n]	l:[l]
alveolar	s:[s]	z:[ts]	c:[tsʰ]		
retroflex	sh:[ʂ]	zh:[tʂ]	ch:[tʂʰ]		r:[ʐ]
alveolo-palatal	x:[ɕ]	j:[tɕ]	q:[tɕʰ]		
velar	h:[x]	g:[k]	k:[kʰ]	ng:[ŋ]	

Front vowels are [i], [y], [e], [a], back vowels [u] and [o] and mid vowels [ɨ] (after [s ts tsʰ]), [ɤ]/[ɭ] (after [ʂ tʂ tʂʰ ʐ]), [ə] (after [t tʰ n l x k kʰ] and in [mən]) and the retroflex [ɚ] in the diminutive suffixation. Diphthongs are [aɪ], [ei], [ao] and [ou]. Rising diphthongs such as [ia], [ie] and [uo] are considered combinations of a glide and a nucleus. Triphthongs are regarded as combinations of a prenuclear glide and a diphthong. Mandarin has two glides, the labiovelar [w] and the palatal [j]. They are listed below. [ɤ] (after [ʂ tʂ tʂʰ ʐ]), [ə] (after [t tʰ n l x k kʰ] and in [mən]) and [ɚ] are represented by "e", [ɨ] (after [s ts tsʰ]) and [ɭ] (after [ʂ tʂ tʂʰ ʐ]) by "I".

open	a:[a] ai:[aɪ] ao:[ao] ei:[ei] ou:[ou] o:[o] e:[ɤ], [ə] and [ɚ] I:[ɨ] and [ɭ]
close	i:[i] ie:[je] io:[jo] iou:[jou] ia:[ja] iao:[jao] u:[u] ua:[wa] ai: [waɪ] ue: [wə] uei:[wei] uo: [wo]
rounded	y:[y] ye:[ye]

5. SEGMENTS PRODUNCED IN MCDC

In terms of the percentage of each segment over all segments produced by the same speaker, we calculated the percentage distribution for all sixteen speakers. The

correlation between them in terms of onset, nucleus, coda and tone is very high. There are only three coda consonants and five lexical tones, so the distribution of coda consonant and lexical tones is as expected highly correlated across all sixteen speakers (r ranges from 0.95 to 1). Nevertheless, the correlation matrices for onset consonants and nuclei are also high, even though the variations are much more complicated than tones and coda consonants. For demonstration, Table 5 shows the correlation matrix for onset consonants.

Table 5: Correlation Matrix for Onset Consonants

	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	s-11	s-12	s-13	s-14	s-15	s-16
s-1		0.98	0.97	0.96	0.88	0.92	0.95	0.96	0.95	0.93	0.94	0.93	0.96	0.96	0.95	0.96
s-2			0.96	0.97	0.90	0.91	0.95	0.95	0.95	0.94	0.95	0.93	0.93	0.94	0.95	0.97
s-3				0.96	0.89	0.94	0.98	0.98	0.96	0.92	0.96	0.96	0.95	0.98	0.94	0.96
s-4					0.91	0.92	0.94	0.96	0.97	0.98	0.96	0.90	0.93	0.96	0.97	0.94
s-5						0.97	0.88	0.91	0.91	0.90	0.90	0.86	0.85	0.87	0.91	0.90
s-6							0.91	0.93	0.94	0.89	0.92	0.91	0.87	0.91	0.92	0.90
s-7								0.98	0.95	0.91	0.97	0.96	0.96	0.95	0.96	0.96
s-8									0.97	0.93	0.97	0.95	0.96	0.96	0.96	0.96
s-9										0.94	0.96	0.92	0.91	0.95	0.94	0.94
s-10											0.95	0.87	0.91	0.93	0.95	0.90
s-11												0.95	0.93	0.96	0.95	0.94
s-12													0.93	0.92	0.89	0.94
s-13														0.95	0.95	0.96
s-14															0.95	0.92
s-15																0.94
s-16																

Irrespective of individual speakers and conversation topics, the distribution of the produced segments and tones is greatly similar across speakers. So using the symbols in Section 4, Table 6 lists the percentage of each segment. We can see that syllables produced by Mandarin speakers have mostly an initial onset consonant (82.91% of all syllables have an initial onset consonant) and about one fourth of the actively used syllables are closed syllables with [n], [ŋ] or the retroflex in the coda position.

Table 6: Percentage of Segments

Onset (%) total: 82.91%							
Consonant	b	c	ch	d	f	g	h
%	4.73	0.85	1.84	10.49	1.24	4.46	6.38
Consonant	j	k	l	m	n	p	q
%	7.07	2.32	3.57	4.52	4.36	0.67	2.92
Consonant	r	s	sh	t	x	z	zh
%	2.00	1.49	8.41	3.41	4.61	3.35	4.23
Nucleus (%) total: 100%							
Vowel	a	ai	ao	e	ei	i	I
%	9.24	4.53	3.17	16.42	2.09	11.92	7.68
Vowel	ia	iao	ie	io	iou	o	ou
%	3.5	2.47	5.4	0.23	5.06	2.27	2.81
Vowel	u	ua	uai	ue	uei	uo	y
%	4.68	2.17	0.4	0.43	5.03	6.91	1.69
Vowel	ye						
%	1.92						
Coda (%) total: 25.86%							
Consonant	n	ng	r				
%	15.04	10.62	0.2				

Summarizing the result of onset consonants by grouping them into natural classes, fricatives are 22.13%, stops 26.07% and affricates are 20.26%. The proportion is similar. Fricatives and stops are similarly frequent in spontaneous Mandarin. Among them, aspirated sounds are 12.01%. Nasal sounds are 8.8%. Nuclei beginning with a glide make up 31.59% (front glide 16.54%, back glide 14.94%), being mid vowels 24.1%, front vowels (single vowels or diphthongs) 30.94%, back vowels (single vowels or diphthongs) 9.76% and rounded vowels 3.61%. Here we see a clear tendency that spontaneous Mandarin production prefers front vowels than back vowels. And rounded vowels are rarely used in spoken Mandarin.

6. DISCUSSION AND CONCLUSION

This paper describes corpus-based results with respect to segmental and lexical features of speech production in spontaneous form. This shows how corpus data can bring new insights into linguistic studies and engineering systems. Pronunciation variations in spontaneous speech, development of a core vocabulary for spoken language and lexical and syntactic particularities of spontaneous speech are research areas to which corpus linguistics can have tremendous contributions. In addition, we are currently using our corpus data to investigate the following phonological issues. Due to reduction of retroflex, the difference between [ɨ] and [ɭ] (e.g. zi1, ci1, si1 vs. zhi1, chi1, shi1, ri1) is not apparent, so how to distinguish them from each other becomes more difficult. And different phonetic values of [a] in different phonological context ([±open syllable, ±front]) need to be determined by their acoustic features (e.g. ta1 vs. tan1 and tang1). Similarly, difference between [ia] and [ie] ([±open syllable, ±close-mid]) needs to be evaluated, too (e.g. jia1 vs. mian4). Moreover, due to influences from the dialect dominantly spoken in Taiwan, Min-Nan, the roundness seems to be dropped in [uŋ] and reduced to [oŋ] in Taiwan Mandarin (e.g. dong1 and xiong1, uŋ->oŋ). The changes ɑ->e/y_n (e.g. fan2 vs. quan2) and ə->ɔ/[+labial]_ŋ are also frequently observed in Taiwan Mandarin (e.g. feng1). These may ground the main segmental differences between Beijing and Taiwan Mandarin. Finally, the author would like to kindly acknowledge the financial support of the National Science Council under grant NSC-92-2411-H-001-075 and the Ministry of Education under grant 91-E-FA06-4-4.

7. REFERENCES

[1] Chen, Keh-Jiann, Huang, Chu-Ren, Chang, L.-P. and Hsu, H.-L., "ACADEMIA SINICA BALANCED CORPUS: design methodology for balanced corpora", *Proceedings of the Eleventh Pacific Asia Conference on Language, Information and Computation*, pp. 167-176, 1996.

[2] Duanmu, San, *The Phonology of Standard Chinese*, Oxford University Press, 2000.

[3] Tseng, Shu-Chuan, "Taxonomy of spontaneous speech phenomena in Mandarin conversation", *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, pp. 23-26, 2003.

[4] Tseng, Shu-Chuan, "Contracted Syllables in Mandarin: Evidence from Spontaneous Conversations", *Language and Linguistics*, 2004. (to appear)