Computational Linguistics and Chinese Language Processing
vol. 3 no. 2, August 1998 pp. 93-114

Statistical Analysis of Mandarin Acoustic Units and Automatic Extraction of Phonetically Rich Sentences Based Upon a Very Large Chinese Text Corpus

Hsin-min Wang*

Abstract

Automatic speech recognition by computers can provide humans with the most convenient method to communicate with computers. Because the Chinese language is not alphabetic and input of Chinese characters into computers is very difficult, Mandarin speech recognition is very highly desired. Recently, high performance speech recognition systems have begun to emerge from research institutes. However, it is believed that an adequate speech database for training acoustic models and evaluating performance is certainly critical for successful deployment of such systems in realistic operating environments. Thus, designing a set of phonetically rich sentences to be used in efficiently training and evaluating a speech recognition system has become very important. This paper first presents statistical analysis of various Mandarin acoustic units based upon a very large Chinese text corpus collected from daily newspapers and then presents an algorithm to automatically extract phonetically rich sentences from the text corpus to be used in training and evaluating a Mandarin speech recognition system.

Keywords: Mandarin speech recognition, statistical analysis of acoustic units, phonetically rich sentences, speech database

1. Introduction

Automatic speech recognition by computers can provide the most natural and efficient method of communication between humans and computers. Over the past decades, researchers all over the world have been involved in projects that have aimed to develop automatic speech recognizers, and many high performance systems have begun to emerge from research institutes and laboratories. However, experience has shown that the deployment of such speech recognition systems in realistic operating environments will

*Institute of Infromation Science, Academia Sinica, Taipei, Taiwan, R. O. C. E-mail:whm@iis.sinica.edu.tw

require much better speech data to help us model the inherent variability in speech signals among different speakers and in different environments, and to help us evaluate performance under near realistic conditions. Thus, researchers all over the world have also participated in many efforts devoted to collecting speech databases of their own languages [Akira *et al.*, 1990; Yu and Liu, 1990; Zue *et al.*, 1990; Tseng, 1995; Wang, 1997], in addition to developing robust algorithms for speech recognition.

Because the Chinese language is not alphabetic and keyboard input of Chinese characters into computers requires a considerable amount of effort and training, Mandarin speech recognition is highly desired especially in the Chinese community. In Taiwan, speech recognition systems have been developed for a wide variety of applications, such as small to large vocabulary keyword spotting [Huang, Wang, and Soong, 1994; Bai, Tseng, and Lee, 1997], medium size vocabulary isolated word recognition for voice command and control [Chang *et al*., 1996], large vocabulary speech dictation [Lee *et al.*, 1993a; Lee *et al.*, 1993b; Huang and Wang, 1994; Lyu and Lee *et al.*, 1995; Wang and Lee *et al.*, 1995; Shen, 1996], limited-domain speech understanding [Lin, Wang, and Lee, 1997], and so on. An adequate speech database is certainly critical for successful development of such a system. Recently, the research teams who developed the Golden Mandarin series have designed sets of phonetically balanced training sentences based on different acoustic criteria by considering the statistical distributions of different acoustic units, and these sentences have been shown to be very effective when new users utter them according to a prompt on the computer screen to train their own dictation systems [Shen, 1996]. In this paper, we will first present statistical analysis of various Mandarin acoustic units based upon a very large Chinese text corpus collected from daily newspapers and then present an algorithm to automatically extract phonetically rich sentences from this text corpus to be used in efficiently training and evaluating a Mandarin speech recognition system.

The rest of this paper is organized as follows. The characteristic structure of the Chinese language is briefly introduced in section 2, and statistical analysis of various Mandarin acoustic units based upon a very large Chinese text corpus is discussed in section 3. The basic principles and a detailed description of the two-stage algorithm for automatic extraction of phonetically rich sentences from a text corpus are given in section 4 while two example experiments for extracting phonetically rich Chinese sentences are discussed in section 5. Finally, a few concluding remarks are given in section 6.

2. Characteristic Structure of the Chinese Language

In Mandarin Chinese, the total number of Chinese characters is believed to be unknown, but more than 10,000 characters are commonly used. A Chinese word is composed of from one to several characters, and combinations of these characters in fact give an almost unlimited number of Chinese words, among which at least some 100,000 of them are commonly used. All the Chinese characters are monosyllabic, and the total number of phonologically allowed syllables is only about 1345. Although the majority of Chinese words are composed of two or more syllables or characters, most of the characters can also be considered as monosyllabic words. This is why accurate recognition of all 1345 Mandarin syllables is believed to be the first key problem in Mandarin speech recognition with a very large vocabulary, and this is also why syllables are often chosen as the basic recognition target, very similar to the words used in systems for other alphabetic languages [Lee, Hon, and Reddy, 1990; Ney *et al*., 1994]. Of course, this small number of syllables also implies that a large number of homonym characters share the same syllable, and that there is a high degree of ambiguity. For example, on average, every syllable is shared by about 7-8 (10,000/1345) possible homonym characters. This one-to-many mapping relation from syllables to characters is certainly another key issue in Mandarin speech recognition with a very large vocabulary, and some relevant problems have been discussed in many papers [Lee *et al*., 1993a; Lee *et al*., 1993b; Lyu and Lee *et al*., 1995; Wang and Lee *et al*., 1995; Shen, 1996].

<table>
<tr><td colspan="4">Tonal syllable (1345)</td></tr>
<tr><td colspan="3">Base syllable (416)</td><td rowspan="3">Tone (5)</td></tr>
<tr><td rowspan="2">INITIAL (22)</td><td colspan="2">FINAL(41)</td></tr>
<tr><td>Medial (3)</td><td>Nucleus (9)</td><td>Ending (5)</td></tr>
</table>

Table 1. *The phonological hierarchy of Mandarin syllables, where the number inside each bracket indicates the total number of units of that kind in Mandarin Chinese.*

Another very important feature of Mandarin Chinese is the existence of tones for syllables. Mandarin Chinese is a tonal language, in which each syllable is assigned a tone, and the tones have lexical meaning. There are basically a total of four lexical tones, i.e., the high-level tone (usually referred to as Tone 1), the mid-rising tone (Tone 2), the mid-falling-rising tone (Tone 3), and the high-falling tone (Tone 4) as well as one neutral tone (Tone 5). It has been found that the vocal tract parameters for Mandarin speech are only slightly influenced by the tones, and that the tones can be separately recognized primarily using pitch contour information [Wang and Lee, 1994; Wang and Chen, 1994]. If the differences among the syllables caused by tones are disregarded, then only 416 base

syllables (i.e., syllable structures independent of tones) instead of 1345 different tonal syllables are required to cover the pronunciation of Mandarin Chinese. As a result, every tonal syllable can be considered as a combination of two independent parts, a tone from the five possible choices and a base syllable from the 416 possible candidates disregarding tones. In many large vocabulary Mandarin speech recognition systems [Lee *et al*., 1993a; Lee *et al*., 1993b; Huang and Wang, 1994; Lyu and Lee *et al*., 1995; Wang and Lee *et al*., 1995; Shen, 1996], tones and base syllables are, thus, recognized separately.

	IPA	SPA
Stop(6)	[p] [t] [k] [p'] [t'] [k']	b, d, g, p, t, k
Affricate (6)	[ts] [ts] [t] [ts'] [t s '] [t ']	z, Z, j, c, C, <
Nasal (3)	[m] [n] [ŋ]	m , n, N
Liquid (1)	[l]	l
Fricative (6)	[f] [s] [s] [ ] [x] [z]	f, s, S, T, h, R
Vowel (10)	[a] [o] [ ] [e] [i] [u] [y] [ ] [ ] [ ]	a, o, e, E, i, u, U, Y, y, r
Null phohe*(1)		#

*The null phone is used to represent the null Initial

Table 2(a) *33 PLUs of Mandarin Chinese, where both International Phonetic Alphabet (IPA) and Simplified Phonetic Alphabet (SPA) symbols are listed for reference.*

Conventionally, each of the 416 Mandarin base syllables mentioned above can be decomposed into an INITIAL/FINAL format very similar to the consonant/vowel relations in other languages. There exists a total of 22 INITIALs and 41 FINALs for the 416 Mandarin base syllables, in which the INITIAL is the initial consonant of the base syllable while the FINAL is the vowel or diphthong part of the base syllable but including an optional medial or nasal ending. On the other hand, just as in many other languages, these 63 (22+41) INITIAL/FINALs can also be further decomposed into even smaller acoustic units, for example, phone-like units (PLUs). It has been found that a total of 33 phone-like units (PLUs) is sufficient to transcribe the 416 Mandarin base syllables. The phonological hierarchy of a Mandarin syllable is shown in Table 1, where the relationships among the tonal syllables, base syllables and tones, INITIAL/FINALs, and PLUs are shown. The 33 PLUs of Mandarin Chinese with IPA (International Phonetic Alphabet) representations are listed in Table 2(a), in which the corresponding simplified symbols used in this research (SPA, Simplified Phonetic Alphabet) are also listed for reference. The 22 INITIALs and 41 FINALs are listed in Table 2(b) and (c), respectively, all in the Simplified Phonetic Alphabet (SPA) for simplicity. It can be found that an INITIAL is always a PLU while a FINAL may contain one, two, or three PLUs in

general. That is, a Mandarin base syllable is composed of two to four PLUs. Table 3 lists all 416 base syllables, where the vertical scale lists all 41 FINALs and the horizontal all 22 INITIALs.

#	b	p	m	f	d	t	n	l	g	k
h	j	<	T	Z	C	S	R	z	c	s

Table 2(b) *22 INITIALs of Mandarin Chinese*

Group	Member
1	Y y
2	a ai au an aN
3	o ou
4	e en eN er
5	i ia iE iai iau iou iEn in iaN iN io
6	u ua uo uai uEi uan uen uaN ueN uoN
7	U UE Uan Un UN
8	E Ei

Table 2(c) *41 FINALs of Mandarin Chinese*

			INITIAL																				
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
			#	Z	C	S	R	z	c	s	g	k	h	j	<	T	d	t	n	l	b	p	m	f
FINAL	1	Y		1	2	3	4																	
	2	y						5	6	7														
	3	a	8	9	10	11		12	13	14	15	16	17				18	19	20	21	22	23	24	25
	4	o	26																	414				
	5	e	27	28	29	30	31	32	33	34	35	36	37				38	39	40	41			409	
	6	ai	42	43	44	45		46	47	48	49	50	51				52	53	54	55	56	57	58	
	7	E	59																					
	8	Ei	60	61		62		63		64	65		66				67		68	69	70	71	72	73
	9	au	74	75	76	77	78	79	80	81	82	83	84				85	86	87	88	89	90	91	
	10	ou	92	93	94	95	96	97	98	99	100	101	102				103	104	105	106		107	108	109
	11	en	110	111	112	113	114	115	116	117	118	119	120				121	122	123	124	125	126	127	128
	12	an	129	130	131	132	133	134	135	136	137	138	139				410		140		141	142	143	144
	13	aN	145	146	147	148	149	150	151	152	153	154	155				156	157	158	159	160	161	162	163
	14	eN	164	165	166	167	168	169	170	171	172	173	174				175	176	177	178	179	180	181	182
	15	i	183											184	185	186	187	188	189	190	191	192	193	
	16	u	194	195	196	197	198	199	200	201	202	203	204				205	206	207	208	209	210	211	212
	17	U	213											214	215	216			217	218				
	18	ia	219											220	221	222			412	223				
	19	iE	224											225	226	227	228	229	230	231	232	233	234	
	20	iai	235																					
	21	iau	236											237	238	239	240	241	242	243	244	245	246	
	22	iou	247											248	249	250	251		252	253			254	
	23	ian	255											256	257	258	259	260	261	262	263	264	265	
	24	in	266											267	268	269	411		270	271	272	273	274	
	25	iaN	275											276	277	278			279	280				
	26	iN	281											282	283	284	285	286	287	288	289	290	291	
	27	ua	292	293	294	295					296	297	298											
	28	uo	299	300	301	302	303	304	305	306	307	308	309				310	311	312	313	314	315	316	317
	29	uai	318	319	320	321					322	323	324											
	30	uEi	325	326	327	328	329	330	331	332	333	334	335				336	337						
	31	uan	338	339	340	341	342	343	344	345	346	347	348				349	350	351	352				
	32	uen	353	354	355	356	357	358	359	360	361	362	363				364	356	413	366				
	33	uaN	367	368	369	370					371	372	373											
	34	ueN	374																					
	35	uoN		375	376	377	378	379	380	381	382	383	384				385	386	387	388				
	36	UE	389											390	391	392			393	394				
	37	Uan	395											396	397	398				399				
	38	Un	400											401	402	403		415						
	39	UN	404											405	406	407								
	40	er	408																					
	41	io	416																					

Table 3 *416 Base Syllables of Mandarin Chinese*

3. Statistical Analysis of Mandarin Acoustic Units Based Upon A Very Large Chinese Text Corpus

The Chinese text corpus used here to analyze the statistical distribution of Mandarin acoustic units was collected from daily newspapers. English characters or other special symbols contained in the sentences were simply discarded; then, the remaining sentences were word identified [Chen and Liu 1992] and phonetic spelling indicated using a lexicon

consisting of around 85,000 frequently used Chinese words [CKIP 1993]. All the words in the lexicon are composed of from one to four characters, and the number of words is analyzed in Table 4. Finally, the corpus consisting of a total of 22,660,835 sentences (271,360,277 characters or syllables) was used to analyze the statistical distribution of Mandarin acoustic units.

	1-character word	2-character word	3-character word	4-character word	Total
# of words	14052	48339	11559	10433	84383

Table 4 *The Number of Words Contained in the Chinese Lexicon used in this Research*

First, the analysis was based on the frequency counts of 416 base syllables in the corpus. The results are summarized in Table 5. It can be seen that the top 10 most frequently used base syllables cover more than 19% of base syllables used in everyday newspapers, the top 50 cover more than 50%, and the top 200 cover more than 92%. On the other hand, among the 416 base syllables, more than 100 of them have less than 1% frequency of occurrence, which means that around 25% of the base syllables are barely used in everyday newspapers. The top 30 most frequently used base syllables are listed in Table 6. Although most of the top 30 base syllables most frequently used at the beginning and end of sentences also belong to the overall top 30 most frequently used base syllables except that the order might be slightly different, it was found that some of them show very high frequency only at specific positions; e.g., the base syllables "ta", "tai", "dan", "Ze", and "biN" are very frequently used at the beginning of sentences while "Suo", "dian", and "Cu" are used at the end of sentences.

Group	No. of syllables in the group	Total frequency of occurrence (%)	Accumulated frequency (%)
1-10	10	19.4553	19.4553
11-30	20	18.5154	37.9707
31-50	20	12.3730	50.3437
51-100	50	22.3279	72.6716
101-150	50	12.6151	85.2867
151-200	50	7.2693	92.5560
201-300	100	6.5284	99.0844
301-416	116	0.9156	100.0000
total	416	100.0000	

Table 5 *The Frequency Counts of the 416 Base Syllables.*

	Beginning of sentences		Middle of sentences		End of sentences		Overall	
	base syllable	frequency of occurrence (%)	base syllable	frequency of occurrence (%)	base syllable	frequency of occurrence (%)	base syllable	frequency of occurrence (%)
1	ZuoN	4.3139	SY	3.8267	SY	6.5184	SY	3.9364
2	#i	3.8161	de	2.8260	#i	2.4045	#i	2.8137
3	SY	2.4485	#i	2.7542	Suo	2.3450	de	2.5035
4	ta	2.2782	ji	2.0613	dian	2.2258	ji	1.9116
5	jin	2.0577	ZY	1.6633	Cu	2.1890	ZY	1.6120
6	tai	2.0132	guo	1.5641	#Uan	2.1845	guo	1.4833
7	#iou	1.9962	#uEi	1.3024	li	1.8829	ZuoN	1.4783
8	zai	1.9309	#U	1.2790	huEi	1.7694	#uEi	1.3031
9	dan	1.9144	huEi	1.2252	ren	1.4151	li	1.2191
10	Ze	1.7445	ZuoN	1.2088	ZY	1.3987	huEi	1.1942
11	biN	1.7051	guoN	1.1543	#uEi	1.3348	#U	1.1927
12	#er	1.5319	li	1.1534	ZuoN	1.3311	#Uan	1.1244
13	ZY	1.3131	bu	1.1068	jian	1.2696	jin	1.1161
14	guo	1.3009	#Uan	1.0991	TiN	1.2681	#iou	1.0844
15	#uEi	1.2783	jin	1.0872	ji	1.2566	bu	1.0738
16	bu	1.2563	#u	1.0704	hou	1.2397	guoN	1.0621
17	li	1.2110	#iou	1.0572	de	1.0702	#u	1.0260
18	jiaN	1.2008	ren	1.0380	ti	1.0531	ren	1.0147
19	mEi	1.1709	ZeN	1.0132	hua	0.9774	zai	0.9552
20	<i	1.0862	zai	0.9341	Ti	0.9203	ZeN	0.9489
21	ji	1.0741	<i	0.9309	diN	0.8725	<i	0.9284
22	jiN	1.0522	da	0.8837	TiaN	0.8712	da	0.8547
23	#in	1.0296	jian	0.8064	#u	0.8646	jian	0.8266
24	#U	0.9548	fu	0.8021	guo	0.8597	#er	0.8040
25	da	0.9543	fa	0.7910	#an	0.8545	TiN	0.7724
26	#iE	0.9245	#er	0.7828	CaN	0.8387	fa	0.7652
27	duEi	0.8186	he	0.7624	ZaN	0.8233	tai	0.7629
28	guoN	0.8049	TiN	0.7385	jia	0.7903	fu	0.7482
29	ZeN	0.7585	sy	0.7300	Tian	0.7884	Cu	0.7450
30	#u	0.7446	jia	0.7151	fa	0.7813	jia	0.7097
total		46.6839		38.3677		44.3985		37.9707

Table 6 *The Frequency Counts of the Top 30 Most Frequently Used Base Syllables*

For speech recognition purposes, especially for continuous speech recognition, the co-articulation effects between adjacent syllables are usually significant, so recognition accuracy usually degrades clearly from isolated syllable recognition to continuous speech recognition. Many context-dependent acoustic modeling techniques which specially consider the contextual situation are, therefore, widely used to compensate for the co-articulation effects and, thus, improve recognition accuracy [Lee, Hon, and Reddy, 1990; Ney *et al*., 1994; Lyu and Lee *et al.*, 1995; Wang and Lee *et al.*, 1995]. Just as the frequency counts of the 416 base syllables are very different as discussed above, the concatenation combinations of base syllables are distributed over a wide range. Table 7 lists the frequency counts of tri-base syllables. Although there is a total of 71,991,296 (416^3) possible combinations of tri-base syllables, only 7,927,335 (11.01%) of them were found in the corpus. Among the existing tri-base syllable combinations, 2,671,395 of

them were found only once while only 1,808,572 were found more than 10 times. Furthermore, it is worth noting that 11 of the tri-base syllable combinations appear more than 100,000 times in the corpus; they are "#uEi #Uan huEi" (委員會), "tai bEi SY" (台北市), "ZuoN hua min" (中華民), "hua min guo"(華民國), "TiN ZeN #Uan" (行政院), "li fa #Uan"(立法院), "bai fen ZY"(百分之), and so on. Since it is not feasible to collect sufficient speech data to include all the existing tri-base syllable combinations, even when only combinations which appear 10 or more times are considered, most Mandarin speech recognition systems are, in fact, based on sub-units, such as INITIAL/FINALs [Wang and Lee *et al.*, 1995; Chang *et al.*, 1996; Shen, 1996; Bai, Tseng, and Lee, 1997; Lin, Wang, and Lee, 1997].

Frequency counts	1	>1	>5	>10	>50	>100	>1000	>10000	>100000
# of tri-base syllables	2671395	5255940	2683364	1808572	586974	327852	28085	979	11
% of possible combinations	3.710719	7.300799	3.727345	2.512209	0.815340	0.455405	0.039012	0.001360	0.000015

Table 7: *The Frequency Counts of Tri-base Syllable*s.

Although there are 902 (22*41) possible INITIAL-FINAL combinations, only 416 of them are phonologically allowed, and they comprise the 416 base syllables of Mandarin Chinese. Furthermore, there are 9,152 (416*22) possible INITIAL-FINAL-INITIAL combinations (equal to the number of possible combinations of base syllables with the INITIALs of their following base syllables), among which 8,722 (95.30%) were found in the corpus. The most frequently used INITIAL-FINAL-INITIAL is "S-Y-#", which has 0.83% frequency of occurrence. The remaining combinations are distributed quite flatly such that, from the second most frequently used combination to the 100-th most frequently used combination, the frequency of occurrence decreases gradually from 0.42% to 0.12%. The top 100 most frequently used combinations cover 18.77% of the occurrence of all the INITIAL-FINAL-INITIAL combinations. On the other hand, there are 17,056 (416*41) possible FINAL-INITIAL-FINAL combinations (equal to the number of possible combinations of base syllables with the FINALs of their preceding base syllables), of which 14,321 (83.96%) were found in the corpus. Again, the distribution is quite flat such that, from the most frequently used combination to the 100-th most frequently used combination, the frequency of occurrence decreases gradually from 0.35% to 0.10% and the accumulated frequency of the top 100 most frequently used combinations reaches 15.75%. Table 8 lists the frequency counts of the INITIALs that appear at the beginning of sentences. It can be found that the top 6 most frequently used INITIALs, such as "#" (null INITIAL), "Z", "j", "d", "t", and "b", out of the 22 INITIALs account for more than

50% of the frequency of occurrence at the beginning of sentences. On the other hand, for the FINALs as the end of sentences, as shown in Table 9, the top 9 most frequently used FINALs, such as "i", "Y", "u", "iEn", "uo", "iN", "an", "uEi", and "e", out of the 41 FINALs also account for more than 50% of the frequency of occurrence in everyday newspapers.

INITIAL	#	Z	j	d	t	b	g	S	T	z	l
Frequency of occurrence (%)	16.9662	10.0083	9.0221	7.0685	6.0899	5.7940	4.7310	4.6953	4.6872	4.3168	4.0258
Accumulated frequency (%)	16.9662	26.9745	35.9966	43.0651	49.1550	54.9491	59.6801	64.3754	69.0626	73.3794	77.4053
INITIAL	m	h	<	C	R	f	s	n	c	k	p
Frequency of occurrence (%)	3.4716	3.0636	2.9544	2.1851	2.1048	2.0557	1.8689	1.5026	1.4073	1.2850	0.6957
Accumulated frequency (%)	80.8769	83.9405	86.8949	89.0801	91.1849	93.2405	95.1094	96.6120	98.0193	99.3043	100.00

Table 8 *The Frequency Counts of INITIALs at the Beginning of Sentences (in the Order of Frequency of Occurrence)*

FINAL	i	Y	u	iEn	uo	iN	an	uEi	e	uoN	else
Frequency of occurrence (%)	9.1773	8.5746	7.6424	6.8212	4.8508	4.5254	4.0051	3.9879	3.9704	3.4420	43.0029
Accumulated frequency (%)	9.1773	17.7519	25.3943	32.2155	37.0663	41.5917	45.5968	49.5847	53.5551	56.9971	100.00

Table 9 *The Frequency Counts of FINALs at the End of Sentences (in the Order of Frequency of Occurrence).*

	Beginning of sentences	Middle of sentences	End of sentences	Overall
Tone 1	26.4100	20.7775	18.3910	21.0327
Tone 2	25.1454	23.9060	22.0296	23.8369
Tone 3	17.0535	17.8385	13.9836	17.4352
Tone 4	31.3911	34.3228	44.2315	34.9039
Tone 5	0.0000	3.1552	1.3643	2.7913

Table 10 *The Frequency Counts of the 5 Tones*

Then, the analysis was based on the frequency counts of the 5 different tones in the corpus. The results are summarized in Table 10. It was found that these 5 tones are in the order of Tone 4, Tone 2, Tone 1, Tone 3, and Tone5, according to the frequency of occurrence, no matter whether they occur at the beginning, middle, or end of the sentences, except that Tone 1 is more frequently used than Tone 2 at the beginning of sentences. The frequency counts of Tone 5 should be smaller if phonetic labeling errors, such as " 請著 (Zuo2) 便服… " was phonetic spelling indicated as " 請著 (Ze5)...", " 求個

(ge4) 股表現… " as " 求個 (ge5) … ", " 能了 (liau3) 卻多年心願… " as " 能了 (le5) … ", and so on, which often occurred at the words or characters with more than one allowed pronunciation, were taken into account. Furthermore, Tone 5 syllables are barely used at the beginning of sentences, except for some exclamation sentences that contain only a single character, such as " 啊 (#a5)!", " 呀 (#ia5)!", " 嘿 (hEi5)!", " 哇 (#ua5)!", and so on. The total number of possible tri-tones is 125 (5^3). The frequency counts of these tri-tones according to the frequency of occurrence are shown in Figure 1 while the details for the top 20 most frequently used tri-tones are listed in Table 11. The accumulated frequency counts of the 5 tones according to 125 tri-tones (these 125 tri-tones are also in the order of frequency of occurrence) are further shown in Figure 2.

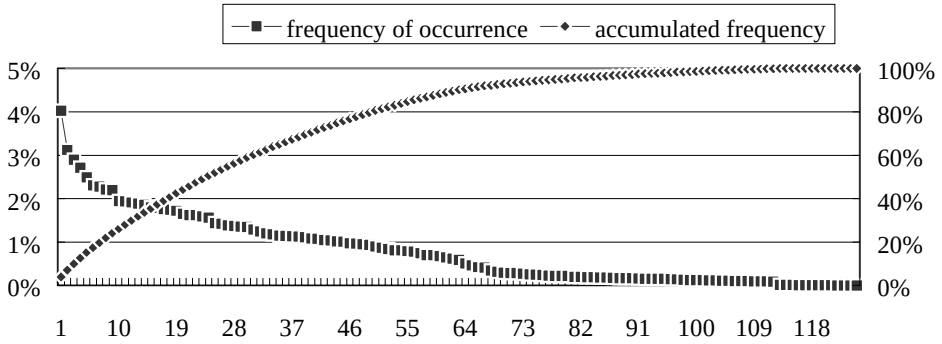

Figure 1 *The Frequency Counts of 125 Tri-tones in Their Order of Frequency of Occurrence*

Top	Tone combination	Frequency of occurrence (%)	Top	Tone combination	Frequency of occurrence (%)
1	4 4 4	4.0207	11	3 4 4	1.9423
2	2 4 4	3.1205	12	2 4 2	1.9109
3	4 2 4	2.8940	13	1 2 4	1.8858
4	4 4 2	2.7095	14	2 1 4	1.8579
5	1 4 4	2.4949	15	1 1 4	1.7971
6	4 4 3	2.3062	16	4 2 2	1.7906
7	4 1 4	2.2778	17	4 2 1	1.7653
8	4 3 4	2.2139	18	4 3 2	1.7466
9	4 4 1	2.1988	19	1 4 2	1.7214
10	2 2 4	1.9492	20	3 2 4	1.6506
total (top10)		26.1855	total (top20)		44.2540

Table 11 *The Frequency Counts of the Top 20 Most Frequently Used Tri-tones*

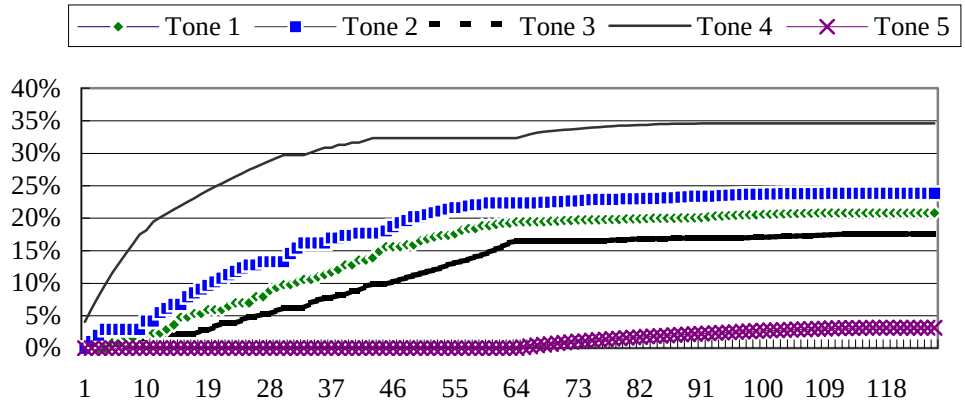

Figure 2 *The Accumulated Frequency Counts of the 5 Tones According to the 125 Tri-tones. (The 125 Tri-tones are shown in their order of frequency of occurrence.)*

Note that all the top 64 (4^3) most frequently used tri-tones are composed of 4 lexical tones, and that these tri-tones account for 90.62% of the frequency of occurrence while the rest are tri-tones with at least one Tone 5 syllable, and such tri-tones account for only 9.38% of the frequency of occurrence. Furthermore, among the tri-tones with at least one Tone 5 syllable, the 65-th to the 112-nd most frequently used tri-tones are tri-tones with only one Tone 5 syllable, the 113-rd to the 124-th are tri-tones with two Tone 5 syllables, and the 125-th (the least frequently used tri-tone) is a tri-tone composed of three Tone 5 syllables. The tri-Tone 5 combination has only 0.0061% frequency of occurrence; some examples are "zy5 men5 de5" (from 孩 " 子們的 " …), "men5 de5 le5" (from 看他 " 們的了 "), "men5 de5 ba5" (from 看我 " 們的吧 "), "ge5 ge5 de5" (from 一 " 個個的 "), etc. From Table 11, it is worth noting that all of the top 20 most frequently used tri-tones consist of at least one Tone 4 syllable. In fact, from Figure 2, we can further find that all of the top 30 most frequently used tri-tones consist of at least one Tone 4 syllable.

Though the above statistical analysis was performed based upon the text corpus collected from daily newspapers, in which the verbiage and the writing style might be slightly different from that of colloquial language in other specific domains, such as novels, magazines, and so on, it is believed that, except for some domain specific proper nouns, most of the frequently used words or characters are very similar across different domains. The statistical results obtained here based upon the text corpus collected from daily newspapers, therefore, provide valuable information which is certainly referable. Moreover, newspapers provide a reliable channel for collecting a very large-scale text corpus since they are generated day after day with the most up-to-date contents. This is

the major reason why the statistical analysis was performed based upon a text corpus collected from daily newspapers in this study, and why, for many speech recognition systems, the language models are trained primarily based upon a text corpus collected from daily newspapers.

4. An Algorithm for Automatic Extraction of Phonetically Rich Sentences From a Text Corpus

For speech recognition purposes, so-called phonetically rich sentences consist of an almost smallest set of grammatically valid sentences, which not only include all necessary recognition units, but all these units should appear in some desired statistical distribution. Such a set of phonetically rich sentences will, then, be very useful in training and evaluating a speech recognition system. Because the recognition tasks (application domains, vocabulary, recognition units, such as phones, diphones, triphones, as well as other sub-word units, which are context-dependent or independent, etc.) are different for different recognition systems, trying to manually generate for each task such a set of phonetically rich sentences to be used in training and evaluating the system as was done in the past [Akira et al., 1990; Yu and Liu, 1990; Zue et al., 1990] will not be cost-effective. Furthermore, it's even very difficult for human experts to reproduce the statistical distribution of the recognition units in the recognition task while they are manually generating or selecting the training and testing sentences. Apparently, automatically generating such a phonetically rich sentence set from a text corpus which defines the task is highly desired. Here, a two-stage algorithm is, therefore, proposed.

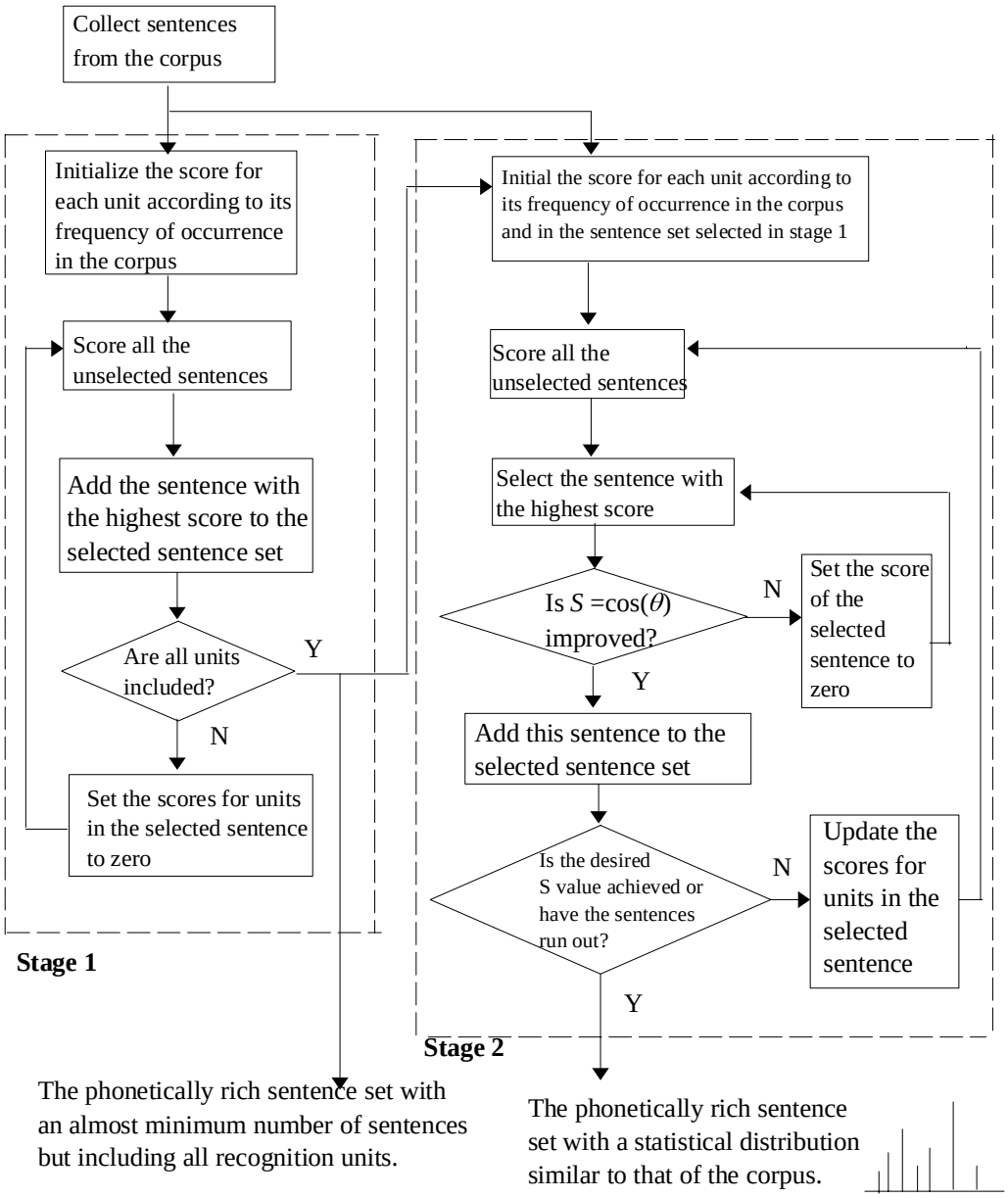

Figure 3 *The Flow Chart of the Two-stage Sentence Selection Algorithm*

The flow chart of our two-stage algorithm is shown in Figure 3, and the basic principles used in designing this algorithm can be described by the following rules:

(1) All recognition units used in the corpus should be included.
(2) Those sentences with a larger number of different recognition units should be

selected with higher priority.

(3) In the first stage, those sentences consisting of units with lower frequency of occurrence in the corpus should be selected with higher priority, so that the total number of sentences to cover all the units can be as small as possible.

(4) In the second stage, on the other hand, those sentences consisting of units with higher frequency of occurrence in the corpus should be selected with higher priority, so that the desired statistical distribution can be achieved as soon as possible.

In the first stage, the input is the whole text corpus, and the desired output is an almost smallest set of sentences, including all the necessary recognition units plus co-articulation effects. To achieve this goal, a score is first assigned to each unit (co-articulation effects should be included when defining context-dependent units), which is initialized as the reciprocal of its frequency of occurrence in the text corpus, so that rare units have higher priority for selection. A score is also defined for each sentence, which is calculated as the average of the scores of its component units but is modified using two weights. The first weight, defined as,

$$w_1 = \frac{\text{number of distinct units in a given sentence}}{\text{number of units in a given sentence}}, \qquad (1)$$

is higher for sentences with a larger number of distinct recognition units because such sentences should have higher scores and be selected with higher priority. The second weight is used to confine the selected sentences to the desired sentence length. Certainly, a long sentence can contain much richer contextual information than a short sentence. However, in general, it is difficult for people to utter long sentences with clear pronunciation. That is, the desired sentences should be neither too long nor too short. The second weight is, therefore, defined as,

$$w2 = \begin{cases} 1.0 & min\_length \le L \le max\_length \\ 0.5 & otherwise \end{cases} \qquad (2)$$

where L is the length (number of units) of a given sentence while *min_length* and *max_length* are the minimum and maximum constraints for the sentence length, respectively. In this paper, 6 for *min_length* and 12 for *max_length* are adopted. Once a sentence is selected, the scores of all the units contained in this sentence are immediately set to zero to avoid these units being selected again. The first stage of the algorithm thus

recursively updates the scores of the units and of all the left unselected sentences and selects additional sentences with the highest score, until all the recognition units are included. In this way, an almost minimum number of sentences which includes all the recognition units can be obtained.

In the second stage of the algorithm, the input is the left unselected sentences in the text corpus and the set of sentences obtained in the first stage, and the desired output is a set of phonetically rich sentences with a statistical distribution for the units very similar to that of the original text corpus. In this stage, the score of each unit is re-defined in a different way. An additional down factor is first defined for each unit, which is proportional to the reciprocal of the number of times this unit appears in the original text corpus. This down factor is used to reduce the priority of a unit to be selected again after each selection. The initial score for each unit in the second stage is then defined as a constant subtracted by its down factor multiplied by the number of times it has been selected previously in the first stage. In this way, the units with higher frequency of occurrence in the original text corpus and lower frequency of occurrence in the set of sentences obtained in the first stage will have higher priority for selection. The rest of the algorithm is very similar to the first stage part. However, in this stage, a similarity measure *S*, as defined in equation (3), is used to estimate the degree to which the statistical distribution of the units in the selected phonetically rich sentence set is similar to that in the original text corpus:

$$S = \frac{\vec{V}_c \cdot \vec{V}_b}{\left|\vec{V}_c\right|\left|\vec{V}_b\right|} = \cos(\theta) \qquad (3)$$

where $\vec{V}_c = \left[n_c(1), \ \ldots, \ n_c(i), \ \ldots n_c(N) \right]$,

$\vec{V}_b = [n_b(1), \ \ldots, \ n_b(i), \ \ldots n_b(N)]$, $n_c(i)$ is the number of times the i-th unit appears in the corpus, $n_b(i)$ is the number of times the i-th unit has been included in the currently selected phonetically rich sentence set, and N is the total number of different recognition units. Apparently, $\vec{V}_c$, $\vec{V}_b$ represent the statistical distribution of the units in the corpus and in the selected sentence set, respectively, S is the normalized inner product of $\vec{V}_c$ and $\vec{V}_b$, and θ is the angle between $\vec{V}_c$ and $\vec{V}_b$. When $S = 1$, i.e., $\vec{V}_c = k \cdot \vec{V}_b$, the statistical distributions will be exactly identical. Now, the sentence

with the highest score and on the same time can improve the similarity measure S is first added to the phonetically rich sentence set. Once a sentence is selected, the scores for all its component units are immediately subtracted by its down factor. By recursively selecting additional sentences one by one as described above until the desired similarity measure S is achieved, one can obtain a set of phonetically rich sentences with a statistical distribution similar to that of the text corpus to be used as a good training and evaluating set.

4.1 Detailed Description of the Two-stage Algorithm

Further details for each stage are given here. Except for N, $n_c(i)$, $n_b(i)$, and L, which have been defined above, all the symbols that will be used in the following are given first.

N_c:the total number of the recognition units in the corpus;

$s[i]$:the score for the i-th unit;

$d_s[i]$:the score down factor for the i-th unit.

4.1.1 The First Stage

The procedure in stage 1 is :

(1) Collect all the sentences from the corpus.
(2) Initialize the score for each unit:

$$s[i] = \frac{1}{n_c[i]}, \quad i = 1, \ \ldots, \ N. \tag{4}$$

(3) Score all the unselected sentences.
(4) Add the sentence with the highest score (*SENT*) to the selected sentence set.
(5) If all the units contained in the corpus are included, end stage 1 and go to stage 2
(6) Set the scores for units contained in *SENT* to zero;

$s[i_k]=0$, $k=1, ...,L$, i_k is the k-th unit of *SENT*. (5)

Then, go to step 3.

4.1.2 The Second Stage

The procedure in stage 2 is :

(1) Continue from stage 1.
(2) Initialize the score for each unit:

for $i=1, \ldots, N$

$$s[i]=const$$

$$d_s[i]= \frac{s[i]}{n_c[i]}$$

$$s[i] = s[i] - d_s[i] \times n_b[i] \qquad (6)$$

(3) Score all the unselected sentences.
(4) If the sentence with the highest score (*SENT*) can improve the similarity, add *SENT* to the selected sentence set. Otherwise, go to step 7.
(5) If the constraint for S is satisfied or if all the sentences in the corpus have run out, end stage 2.
(6) Update the scores for the units contained in *SENT*;

$$s[i_k] = s[i_k] - d_s[i_k], \quad k = 1, \ \ldots L, \quad i_k \text{ is the k-th unit of SENT.} \qquad (7)$$

Then, go to step 3.

(7) Set the score of the highest score sentence to zero and go to step 4.

5. Two Example Experiments for Extraction of Phonetically Rich Chinese Sentences

Two example experiments were performed to test the proposed algorithm. Both were designed to select a set of phonetically rich Chinese sentences to be used for continuous Mandarin speech recognition. Context-independent tonal syllables were chosen as the recognition units in the first experiment while context-dependent INITIALs and context-independent FINALs were chosen in the second experiment. Both examples were chosen simply due to their simplicity. The same algorithm can certainly be used if some other more complicated context-dependent units are needed, as long as the target units are defined. The Chinese text corpus used here consists of a total of 124,845 sentences (1,374,182 syllables), which is a subset of the corpus described in section 3.

In the first experiment, the recognition units chosen were the 1345 phonologically allowed context-independent tonal syllables (i.e., assuming inter-syllabic co-articulation is negligible) in Mandarin due to the monosyllabic structure of the Chinese language. The results are summarized in Table 12. It can be found that at the end of stage 1, only 366 sentences (2790 syllables) were sufficient to include all the recognition units (1345 tonal syllables), in which each tonal syllable appeared only about 2.5 times on average. These numbers correspond to very small percentages of the whole corpus (0.29% of sentences and 0.20% of syllables, respectively). Note that, if a larger text corpus is used, these

percentages can be even smaller. A nice feature here is that the statistical distribution of the tonal syllables included in these 366 sentences obtained in stage 1 is already quite similar to that of the corpus ($S = \cos(\theta) = 0.9064$) because the tonal syllables with higher frequency of occurrence are very naturally carried over into the set selected in the first stage of the algorithm although in this stage, tonal syllables with lower frequency of occurrence have higher priority for selection. When the second stage was performed, on the other hand, the similarity measure $S = \cos(\theta)$ improved very quickly as more sentences were included. When 650 to 750 sentences were included, the statistical distribution was really very close to that of the corpus (S = 0.9931 and 0.9959, respectively) although still much less sentences (0.52% to 0.60%) were needed as compared to the whole corpus. Though the co-articulation effects were not considered in this example, it is obvious that the phonetically rich sentences with co-articulation effects can also be obtained if the context-dependent units are defined.

Stage	Selected sentences		Selected syllables		S	θ
	total number	% in corpus	total number	% in corpus	$\cos(\theta)$	(degrees)
1	366	0.293	2790	0.203	0.9064	24.990
2	400	0.320	3022	0.220	0.9410	19.777
	450	0.360	3377	0.246	0.9681	14.515
	500	0.400	3723	0.271	0.9802	11.433
	550	0.441	4067	0.296	0.9869	9.295
	600	0.481	4405	0.321	0.9907	7.808
	650	0.521	4744	0.345	0.9931	6.742
	700	0.561	5093	0.371	0.9949	5.817
	750	0.601	5477	0.399	0.9959	5.171

Table 12 *The Simulation Results for Selecting Phonetically Rich Chinese Sentences from a Corpus Using 1345 Phonologically Allowed Context-independent Tonal Syllables as the Recognition Units*

In the second experiment, the recognition units chosen were 113 context-dependent INITIALs and 41 context-independent FINALs due to the monosyllabic nature and the INITIAL/FINAL structure of Mandarin Chinese. As shown in Table 2(c), the 41 FINALs can be divided into 8 groups according to their beginning phonemes, and the FINALs in the same group can be assumed to have the same influence on their preceding INITIALs because they all have the same beginning phoneme. For example, the /s/ in the base syllables /sai/, /sau/, /san/, etc. is assumed to be the same in each case but different from the /s/ in the base syllables /su/, /suo/, etc. In this way, the 22 INITIALs can be expanded to 113 context-dependent INITIALs. In other words, the 113 context-dependent INITIALs are kind of "generalized diphones"; i.e., they depend on the group of following FINALs, but co-articulation with the FINALs of the previous syllables is assumed to be

negligible. FINALs, on the other hand, are just assumed to be context-independent because the co-articulation effect on both sides of a FINAL is not significant. Therefore, this example only considered "intra-syllable" co-articulation with a "right-to-left" direction but not "inter-syllable" co-articulation. The results are listed in Table 13. It can be found that only 28 sentences (191 syllables) were sufficient to cover all the recognition units (113 context-dependent INITIALs and 41 context-independent FINALs) after the first stage was completed, in which each INITIAL appeared about 1.7 times and each FINAL about 4.7 times. Though the similarity measure $S = \cos(\theta)$ was not very high ($S = 0.8068$, $\theta = 36.211^{o}$) after the first stage was performed, the second stage could improve S even more quickly than in the previous example. When 80 to 100 sentences (515 to 639 syllables) were included, the statistical distribution was really very close to that of the corpus although still much less sentences (0.064% to 0.080%) were needed as compared to the whole corpus. Other phonetically rich sentences for more complicated context-dependent units (e.g., context-dependent INITIALs and FINALs considering both left and right contextual effects, or other context-dependent phone-like units) can certainly be selected using the same algorithm.

Stage	Selected sentences		Selected syllables		S	θ
	total number	% in corpus	total number	% in corpus	$\cos(\theta)$	(degrees)
1	28	0.022	191	0.0139	0.8068	36.211
2	30	0.024	203	0.0148	0.8469	32.120
	40	0.032	264	0.0192	0.9452	19.048
	50	0.040	327	0.0238	0.9735	13.232
	60	0.048	388	0.0282	0.9862	9.522
	70	0.056	450	0.0327	0.9919	7.279
	80	0.064	515	0.0375	0.9955	5.408
	90	0.072	577	0.0420	0.9971	4.383
	100	0.080	639	0.0465	0.9979	3.682

Table 13 *The Simulation Results for Selecting Phonetically Rich Chinese Sentences from a Corpus Using 113 Context-dependent INITIALs and 41 Context-independent FINALs as the Recognition Units.*

6. Conclusions

How to design a set of phonetically rich sentences to be used in efficiently training and evaluating a speech recognition system has become a very important issue in speech recognition research. In this paper, we have presented statistical analysis of various Mandarin acoustic units, such as syllables, tones, and INITIAL/FINALs, which have been widely adopted as the basic recognition units in many Mandarin speech recognition systems, based upon a very large Chinese text corpus collected from daily newspapers. Furthermore, we have proposed a two-stage algorithm to automatically extract pho-

netically rich sentences from a text corpus to be used in training and evaluating a speech recognition system. We have also proved the efficiency of this algorithm through two example experiments on selecting phonetically rich Chinese sentence sets from a Chinese text corpus. This algorithm can be applied to any language, any recognition task, and any pre-defined recognition units with co-articulation effects, as long as the text corpus defining the task is given.

Acknowledgments

The author would like to thank Mr. Yu-hsueh Chou and Mr. Yuan-cheng Chang for their contributions in programming. Special thanks are due to Prof. Lin-shan Lee for many helpful comments. Thanks are also due to the three anonymous reviewers for their valuable suggestions.

References

Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, "ATR Japanese Speech Database As A Tool of Speech Recognition and Synthesis", *Speech Communication*, No. 9, 1990, pp. 365-374.

Bai B.-R., Tseng C.-Y., and Lee L.-S., "A Multi-phase Approach for Fast Spotting of Large Vocabulary Chinese Keywords from Mandarin Speech Using Prosodic Information", *ICASSP97*, Vol. 2, pp. 903-906.

Chang H.-Y., B. Chen Chou C.-S., and Liu C.-M., "Speaker-independent Mandarin Polysyllabic Word Recognition", *Int. Symp. on Signal Processing and Its Applications*, 1996.

Chen K.-J. and Liu S.-H., "Word Identification for Mandarin Chinese Sentences", *COLING92*, pp. 101-107.

CKIP group, "Analysis of Syntactic Categories for Chinese", *CKIP Technical Report, No. 93-05*, Institute of Information Science, Academia Sinica, Taipei, 1993.

Huang C.-C. and Wang J.-F., "A Mandarin Speech Dictation System Based on Neural Network and Language Processing Model", *IEEE Trans. on Consumer Electronics*, Vol. 40, No. 3, 1994, pp. 437-445.

Huang E.-F, Wang H.-C., and Soong F.-K., "A Fast Algorithm for Large Vocabulary Keyword Spotting Application", *IEEE Trans. on Speech and Audio Processing*, Vol. SAP-2, No. 3, 1994, pp. 449-452.

Lee K.-F., Hon H.-W., and R. Reddy, "An Overview of the SPHINX Speech Rcognition System", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 38, No. 1, Jan. 1990, pp. 35-45.

Lee L.-S. et al., "Golden Mandarin (I) - A Real-time Mandarin Speech Dictation Machine for Chinese Language with very Large Vocabulary", *IEEE Trans. on Speech and Audio Pro-*

cessing, Vol. 1, No. 2, April 1993, pp. 158-179.

Lee L.-S. et al., "Golden Mandarin (II) - An Improved Single-chip Real-time Mandarin Dictation Machine for Chinese Language with very Large Vocabulary", *ICASSP93*, pp. 503-506.

Lin B.-S., Wang H.-M., and Lee L.-S., "Key-phrase Understanding Framework Integrating Real World Knowledge with Speech Recognition with Initial Application in Voice Memo Systems for Mandarin Chinese", *IEEE TENCON97*, pp. 157-160.

Lyu Renyuan, Lee L.-S. et al., "Golden Mandarin (III) - A User-adaptive Prosodic-segment-based Mandarin Dictation Machine for Chinese Language with very Large Vocabulary", *CASSP95*, pp. 57-60.

Ney H., V. Steinbiss, R. Haeb-Umbach, B.-H. Tran, and U. Essen, "An Overview of the Philips Research System for Large-Vocabulary Continuous Speech Recognition", *Int. J. Pattern Recognition and Artificial Intelligence*, Vol. 8, No. 1, Feb. 1994, pp. 33-70.

Shen J.-L., *Improved Mandarin Dictation: New Technologies and Golden Mandarin (III) Windows 95 Version*, Ph. D. dissertation, National Taiwan University, Dec. 1996.

Tseng C.-Y., "A Phonetically Oriented Speech Database for Mandarin Chinese", *proc. International Congress of Phonetic Sciences*, Vol. 3, 1995, pp. 326-329.

Wang H.-C., "MAT - a project to collect Mandarin speech data through telephone networks in Taiwan", *Int. J. of Computational Linguistic & Chinese Language Processing*, Vol. 2, No. 1, February 1997, pp. 73-90.

Wang H.-M. and Lee L.-S., "Tone Recognition for Continuous Mandarin Speech with Limited Training Data Using Selected Context-dependent Hidden Markov Models", *J. of The Chinese Institute of Engineers*, Vol. 17, No. 6, 1994, pp. 775-784.

Wang H.-M., Lee L.-S. et al., "Complete Recognition of Continuous Mandarin Speech for Chinese Language with very Large Vocabulary but Limited Training Data", *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 2, March 1997, pp. 195-200.

Wang Y.-R. and Chen S.-H., "Tone Recognition of Continuous Mandarin Speech Assited with Prosodic Information", *J. Acoustic Society of America*, Vol. 96, No. 5, 1994, pp. 2637-2645.

Yu S.-M. and Liu C.-S., "The Construction of Phonetically Balanced Chinese Sentences", *Telecommunication Laboratories Technical Journal*, R.O.C., Vol. 28, No. 1, Jan. 1990, pp. 84-91.

Zue Victor, Stephanie Seneff, and James Glass, "Speech Database Development at MIT: TIMIT and Beyond", *Speech Communication*, No. 9, 1990, pp. 365-374.