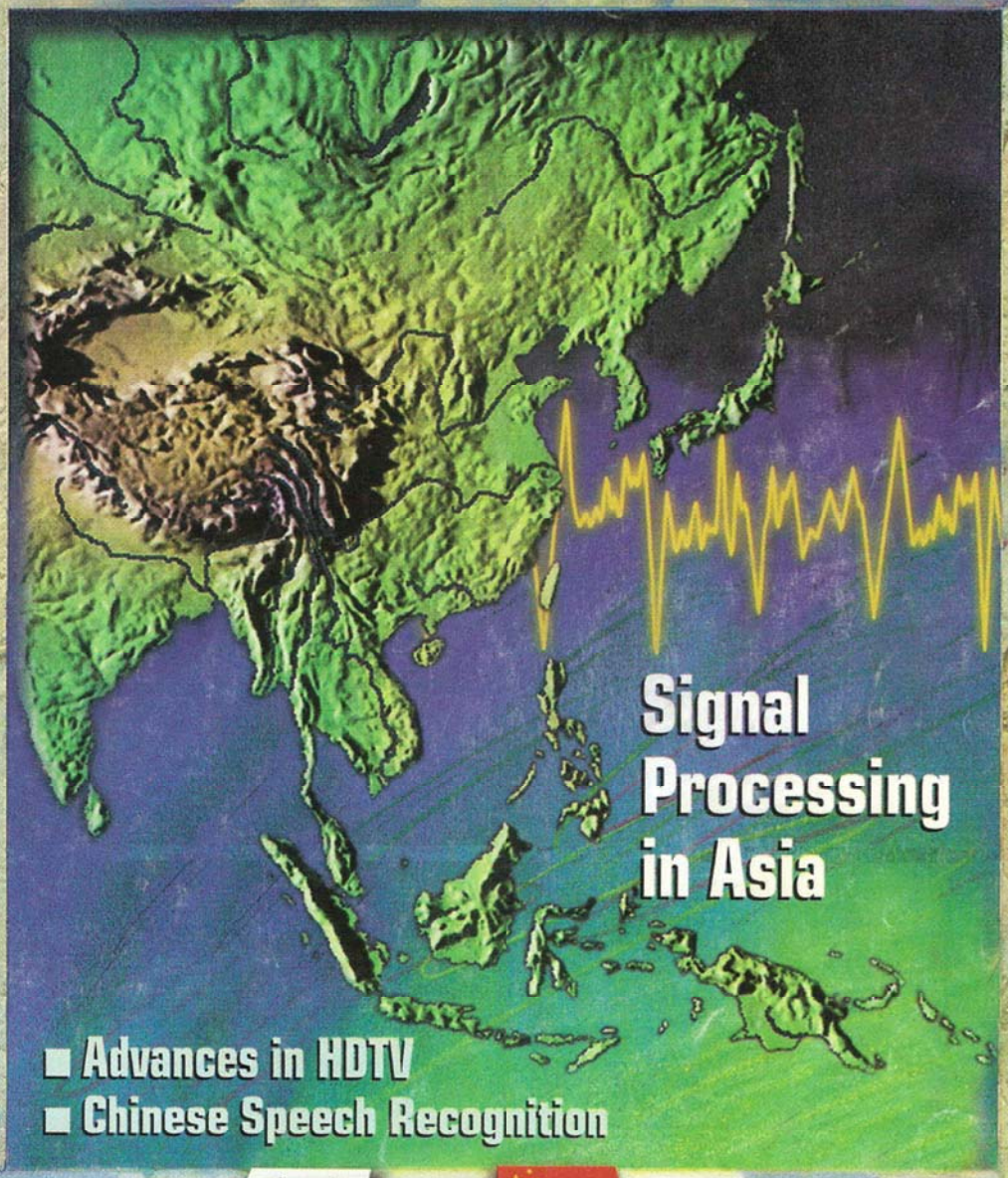
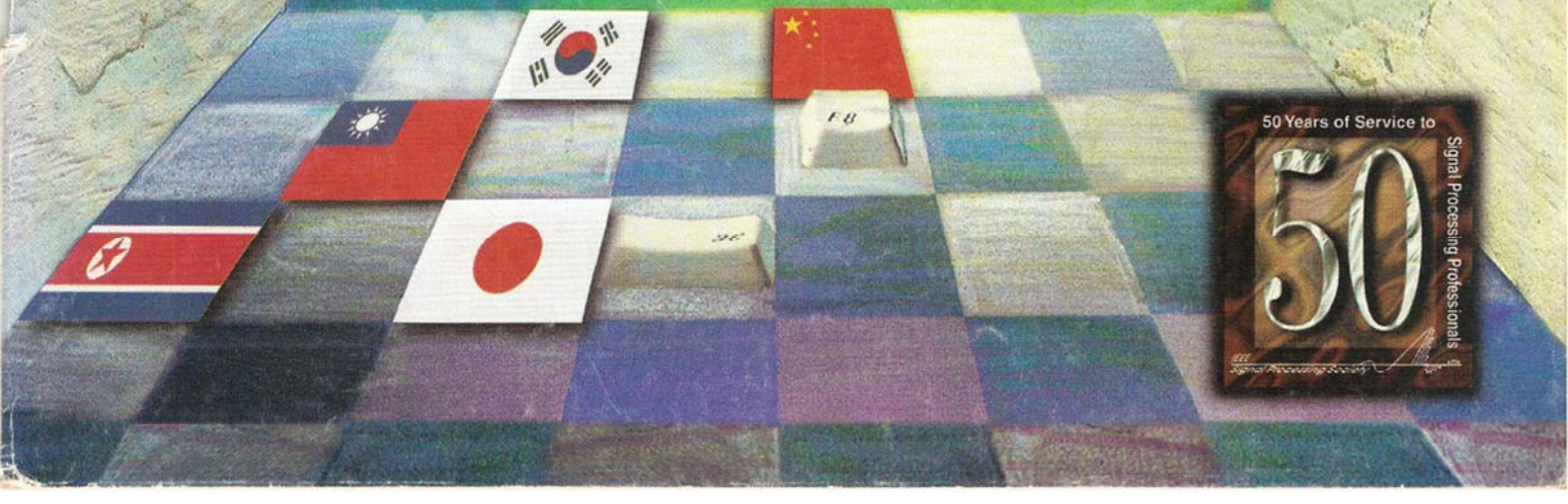
IEEE
Signal Processing
JULY 1997 ○ VOL. 14, NO. 4 ○ ISSN 1053-5888 MAGAZINE



Signal Processing in Asia

- Advances in HDTV
- Chinese Speech Recognition

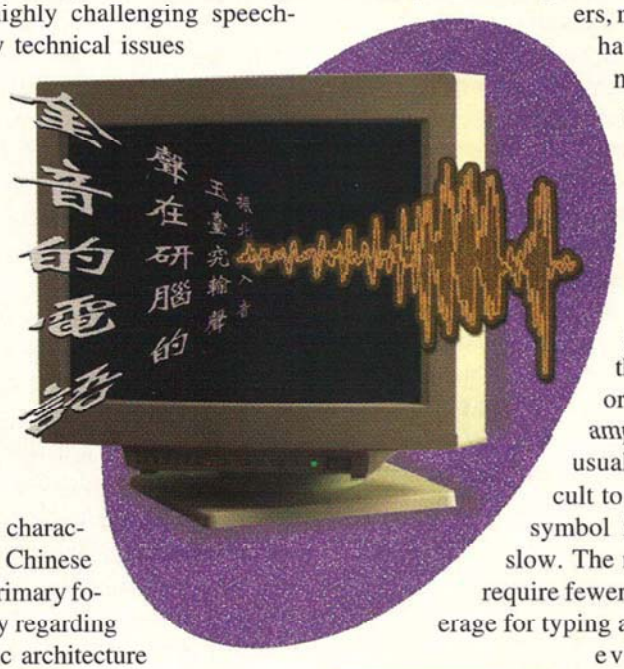


Voice Dictation of Mandarin Chinese

Computer Data Entry Without a Keyboard via Speech Recognition

LIN-SHAN LEE

The Chinese language is not alphabetic, and input of Chinese characters into computers remains a difficult problem even after decades of efforts made by many people to overcome the problem. Voice dictation of Mandarin Chinese with a very large vocabulary is believed to be the perfect solution, but this is highly challenging speech-recognition problem with many technical issues yet unsolved. The characteristics of Mandarin Chinese, significantly different from those of most alphabetic western languages, lead to the fact that many special measures and unique approaches that consider the feature structure of the language are believed to be the key to providing better solutions to the problem. Such special measures and unique approaches are the primary focus of this article.



In this article we analyze the characteristic structure of Mandarin Chinese and discuss related issues. The primary focus is then on the key technology regarding the problem, including the basic architecture for Mandarin dictation, acoustic modeling/processing, and linguistic modeling/processing. Some typical prototype systems, other related applications, and initial industrial efforts and products are finally presented to indicate the feasibility of the key technology discussed.

Chinese Data Entry: Problems and Solutions

More than 25 years after the computer was introduced into the Chinese community, the input of Chinese characters (ideographs) into computers is still a very difficult and unsolved problem. The primary reason is that the Chinese language is not alphabetic. Every Chinese character is a beautiful but complicated square graph, with most of the characters composed of different radicals organized in a highly artistic but irregular manner, and there are at least 10,000 different commonly used Chinese characters [1]. A list of some typical Chinese characters is shown in Fig. 1. These represent good examples of the 10,000 commonly used Chinese characters.

Today, for the input of Chinese characters into computers, more than 200 different methods have been developed based on mapping from keyboards originally designed for alphabetic languages to these Chinese characters. However, almost none of these methods can provide users with a convenient input system as efficient as those for alphabetic languages. These methods are either too slow, too complicated, or require special training. For example, the radical input systems usually have special rules too difficult to memorize, while the phonetic-symbol input systems are usually too slow. The radical input systems generally require fewer keystrokes per character on average for typing and, therefore, are fastest. However, the necessary special mapping rules make them very difficult to learn and very easy to forget if not frequently used. This is why entering Chinese characters quickly into computers using such radical input methods has become a special professional skill in the Chinese community while the majority of people are actually unable to use theses methods in daily work or life.

On the other hand, for a typical phonetic symbol input system, the typing of four to six keystrokes is usually needed to enter a Chinese character, in which two to five keystrokes are for the phonetic symbols, one for the tone, and very often one to two extra are needed to select the desired character from among many homonym characters. This is because Mandarin Chinese is a tonal language, and many homonym characters

1053-5888/97/$10.00©1997IEEE

金聲玉振在中國古典文獻中被用來描述世上最美麗的聲音在臺北的臺灣大學和中央研究院所進行的國語聽寫機的研究已經成功的為中文電腦裝上耳朵期望未來的中文電腦輸入可以完全不用鍵盤對參與這項研究的人而言國語的聲音實為金玉之聲故把研究成果命名為金聲系列

1. A list of typical Chinese character examples.

very often share the same pronunciation, as will be discussed in more detail later on.

Some new techniques (such as Chinese language modeling) have been developed in recent years in which the selection of the desired character among homonym characters can be performed automatically based on the context (with errors to be corrected manually, of course), and some of them have been bundled with efficient software packages such as WINDOWS 95. Many users have found that these techniques provide a much more convenient user environment than before. However, even with such special techniques, the overall input speed is still relatively slow as compared to entering western alphabetic languages; thus, these techniques are not adequate and not widely used. Today, entering Chinese characters into a computer has been a nightmare for many Chinese people trying to use these keyboard input methods, and many have eventually given up.

Taiwan is producing at very low cost a significant portion of the personal computers used worldwide today, but only a relatively small portion of the people in Taiwan are actually using computers in their daily work or life. This is certainly not because the cost of personal computers is too high. It is believed that the major obstacle to popular use of computers in Taiwan is the difficulties in entering Chinese characters into computers.

The situation is very similar on the mainland of China. When the whole world is currently moving toward a fully computerized society at a very fast rate, pushed by ever-developing information technology, the Chinese community, including a quarter of the world's population, still has difficulties in using computers because of its language. The 1.2 billion Chinese people would spend a vast amount of money purchasing computers, peripherals, networks, software packages, and other relevant products to computerize their communities—if their language could be conveniently entered into computers just as western alphabetic languages are. The demand is there, the market will someday be huge, and the potential impact on related areas is almost unlimited.

Handwritten Chinese character input is, of course, a possible solution, but handwriting is generally slow; therefore, it can solve only a small part of the entire character input problem mentioned above. Voice dictation, or speech input, on the other hand, has been suggested as a perfect solution to this problem for some time. Voice input is natural, fast, and convenient. The only problem is that the technology for voice dictation is still not very mature and is highly challenging. Many of the related problems, in particular those peculiar in Mandarin Chinese, are yet to be solved. Today, there are on the order of 100 research groups with thousands of people working on this problem all over the world. Most of them are working on the mainland of China, while the rest are in Taiwan, Hong Kong, Singapore, and other parts of the world. This large number of research groups and researchers indicates very well the potential impact of voice input for Mandarin Chinese.

The problem discussed here is, in general, voice dictation of Mandarin Chinese, or speech recognition for Mandarin Chinese with very large vocabulary and an almost unlimited variety of texts (i.e., with almost all the possible application domains, syntactic structures, and semantic relations existing in the Chinese language, as will be made clear later on). This is because the application is for input for computers, and materials or texts input into computers are generally assumed to be arbitrary without any constraints. Of course, some specific application domains may exist for each user, so the adaptation of the system to the user-specific domain will be discussed later on in this article. Here, we simply assume unlimited domains of texts.

Although the focus here is only transcription of input speech waveforms into corresponding texts without trying to understand the meaning of the text, the goal of very large vocabulary and almost unlimited texts already leads to substantial difficulties for the problem. On the other hand, when we look at the speech-recognition technology available today [2-12], it is now generally realized that speech-recognition technology has matured to a point where the achievable scope of tasks, accuracy, speed, physical size, and the cost of such systems are almost simultaneously crossing the threshold for practically usable systems. Some applications have already been developed and used, and many others are being contemplated today.

In communities with alphabetic languages such as North America and Europe, although speech-dictation technology for very large vocabulary is very well developed [13-22], even with some very convenient products available on the market [23], the most successful products accepted by users are still those for special-domain applications with limited vocabulary such as telephony services. Much of the efforts in product development in the industry probably are also in this direction. A possible reason for this fact is that the input of alphabetic languages into computers is already convenient via keyboards, as compared to the high complexity, cost, and inevitable errors usually associated with the very-large-vocabulary speech-recognition technology.

Such a situation is, however, more or less reversed in the Chinese community. The input of the language into computers is very difficult, but the processing of the language using computers is really necessary. Several preliminary polls indicate that quite a large number of users are ready or even waiting to purchase whatever products become available to them even with relatively high cost and high rate of errors, or cumbersome or unnatural operations, as long as the products work reasonably. This strong and urgent demand is the major drive for the large number of research groups that are focusing their work in this direction. The goal for special-domain applications with limited vocabulary such as telephony services, on the other hand, is of course very important as well, but it seems to have become kind of secondary for the Chinese language.

Although speech-recognition technology for very large vocabulary is very well developed for quite a few languages, the problem for Mandarin Chinese could be very different due to the very special structure of the Chinese language, as will be discussed below. It is believed that many unique measures and special approaches for voice dictation of Mandarin Chinese can be developed based on the characteristics of the Chinese language, which are not only of scientific interest, reflecting the key behavior of the Chinese language, but even of very good reference value for developing the technology for other languages. In this article such key issues and special measures currently seen in this area will be presented. In the Chinese language, there exist hundreds of dialects sounding significantly different from one another although they use the same written characters, but Mandarin is the only official one widely used by all the people for many years. Although there has been some work done on a few dialects [24], we focus on Mandarin only. Also, although there are many research groups working on related problems [25-31], our focus is primarily based on the work done and experiences learned at National Taiwan University and Academia Sinica at Taipei, simply because this is the part of the work the author is most familiar with. Some work done by several other groups will also be mentioned or briefly discussed, but it is not our intention to survey completely all of the work done by the many research groups all over the world.

Characteristic Structure of the Chinese Language

The total number of Chinese characters is unknown, but at least some 10,000 of them are commonly used. A Chinese word is composed of one to several characters. The combination of one to several of such characters gives an almost unlimited number of words, in which at least some 100,000 are commonly used and can be found in different versions of dictionaries and texts on different subjects. Some commonly used words are composed of only one character, and many of these mono-character words appear very frequently in daily language (such as the mono-character words standing for "is," "I," "no," etc.) Since this magazine cannot print Chinese characters in the text, the Chinese characters for these mono-character words are listed in (a) of the box titled "List of Chinese Characters and Words Referred to in the Text," hereafter referred to as "box."

A nice feature of the language is that all the characters are monosyllabic, and the total number of phonologically allowed syllables is only about 1345. In other words, this limited number of syllables represents a much larger number of monosyllabic characters, and the combinations of these many characters in turn provide an almost unlimited number of words. This is why this monosyllable-based structure is usually taken as the first key to Mandarin speech recognition with very large vocabulary, because accurate recognition of these 1345 Mandarin syllables, if achievable, already covers the whole language, including all possible characters and words. In other words, the syllable seems to be a very natural recognition unit in Mandarin speech recognition with very large vocabulary, although this is probably not true in other alphabetic western languages.

Of course, this small number of syllables also implies a large number of homonym characters sharing the same syllable and a high degree of ambiguity. For example, on average every syllable is shared by about 7 to 8 (10,000/1345) characters, each of which can form either a mono-character word or many poly-character words with preceding or following characters in the sentences and so on, as will be discussed in more detail later on. This one-to-many mapping relation from syllables to characters is certainly another key issue in Mandarin speech recognition with very large vocabulary. On the other hand, almost all Chinese characters have their own meanings; i.e., almost each character represents a morpheme or a smallest meaningful unit in the language, and there is almost always a one-to-one mapping relation between characters and morphemes. This is why many of the characters can form mono-character words by themselves, and why the combination of several characters gives an almost unlimited number of poly-character words. In fact, with this property many new words can be easily generated everyday by making new combinations of characters. For example, the combination of the two characters standing for "electricity" and "brain," respectively, becomes a new bi-character word standing for "computer" (see (b1) of the character box).

A very large number of compound words can also be generated by concatenating shorter words; for example, the con-

List of Chinese characters and words referred to in the text

(a) Examples of frequently used mono-character Chinese words

是(is), 我(I), 不(no)

(b) Open vocabulary nature of Chinese language

(b1) 電(electricity)+ 腦(brain) → 電腦(computer)

(b2) 高(high)+ 速(speed) + 公路(highway) → 高速公路(freeway)

(b3) 台灣(Taiwan) + 大學(university) → 台灣大學(Taiwan University)

(b4) 工作(work) + 著(progressive suffix) → 工作著 (is working)

(b5) 台灣大學(Taiwan University) → 台大(Taiwan University)

(c) Different meaning of the words in Fig.3

(c1) 進(advancing) , 近(close to) , 禁(forbidden)

(c2) 記憶(memory) , 技藝(techniques) , 計議(discussions)

(c3) [jii-4][i-4] →記憶(memory) , [i-4][li-4] → 屹立(standing)

(c4) 爭(fighting) , 徵(requesting)

(c5) 競技(competition)

(d) Reasonable choices of words in the sentence of Fig.4(a)

(d1) 電腦(computer) + 科技(technology) → 電腦科技(computer technology)

(d2) 改變(change) + 了(function word) → 改變了(has changed)

(e) An example of Chinese words with multiple linguistic features

架構("construct" as a verb , " structure" as a noun)

(f) Generating rules for Chinese words

(f1) past tense of verbs

吃(eat) + 過 → 吃過(ate)

看(see) + 過 → 看過(saw)

(f2) combining two nouns with specific linguistic categories into a compound noun

豬(pig) +肉(meat) → 豬肉(pork)

(g) The four-character words standing for "sound of gold and jade"

金 聲 玉 振 , 金 玉 之 聲

(h) Simple phrases by concatenating a frequently used word with a preceding or following word

在(in) + 晚上(evening) → 在晚上(in the evening)

到(to) + 台北(Taipei) → 到台北(to Taipei)

美麗(beauty) + 的(function word) → 美麗的(beautiful)

(i) The old Chinese saying

眾志成城 (the integration of great efforts made by many people can build a castle)

catenation of the words standing for "high," "speed," and "highway," respectively, is a new word standing for "freeway" (see (b2) of the box), and the concatenation of the words standing for "Taiwan" and "university" is a new word standing for "Taiwan University," which is also considered a single word (see (b3) of the box).

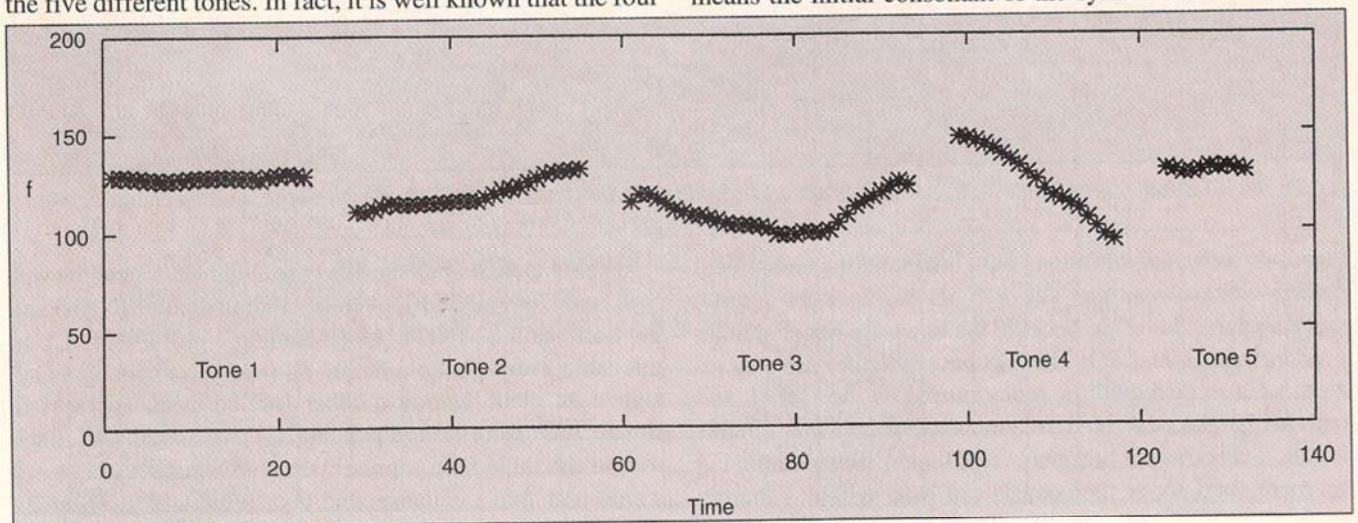
Furthermore, a very large number of word variants can be generated by adding some components such as a suffix; for example, the word standing for "work" becomes a word standing for "is working" by adding the progressive suffix (see (b4) of the box). Moreover, a longer word is very often arbitrarily abbreviated into a shorter new word; for example, the word representing "Taiwan University" using four characters is very often replaced by a short version composed of the first and third characters (see (b5) of the box). As a result, the Chinese language really has a very open vocabulary with an unlimited number of words. This open vocabulary nature is the third key issue for Mandarin speech recognition with very large vocabulary.

Another very important feature of Mandarin Chinese is certainly the existence of tones for syllables. Chinese is a tonal language; in general, every syllable or character is assigned a tone, and the tones have lexical meaning. There are basically 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), the high-falling tone (Tone 4), and one neutral tone (Tone 5). It has been shown [32-35] that the primary difference among the tones is in the pitch contours; there exist standard patterns for the pitch contours for the four lexical tones but not for the neutral tone, and the pitch contours are essentially independent of the vocal tract shape or parameters of the syllables. One example is shown in Fig. 2, where the pitch frequency contours for the syllable [ba] with the four lexical tones and the neutral tone [ba-1], [ba-2], [ba-3], [ba-4], and [ba-5] produced by the same speaker in isolated syllable mode are plotted as functions of time. (The transliteration symbols used in this article are the Mandarin Phonetic Symbols II (MPS II). The number following each syllable denotes the tone of the syllable.)

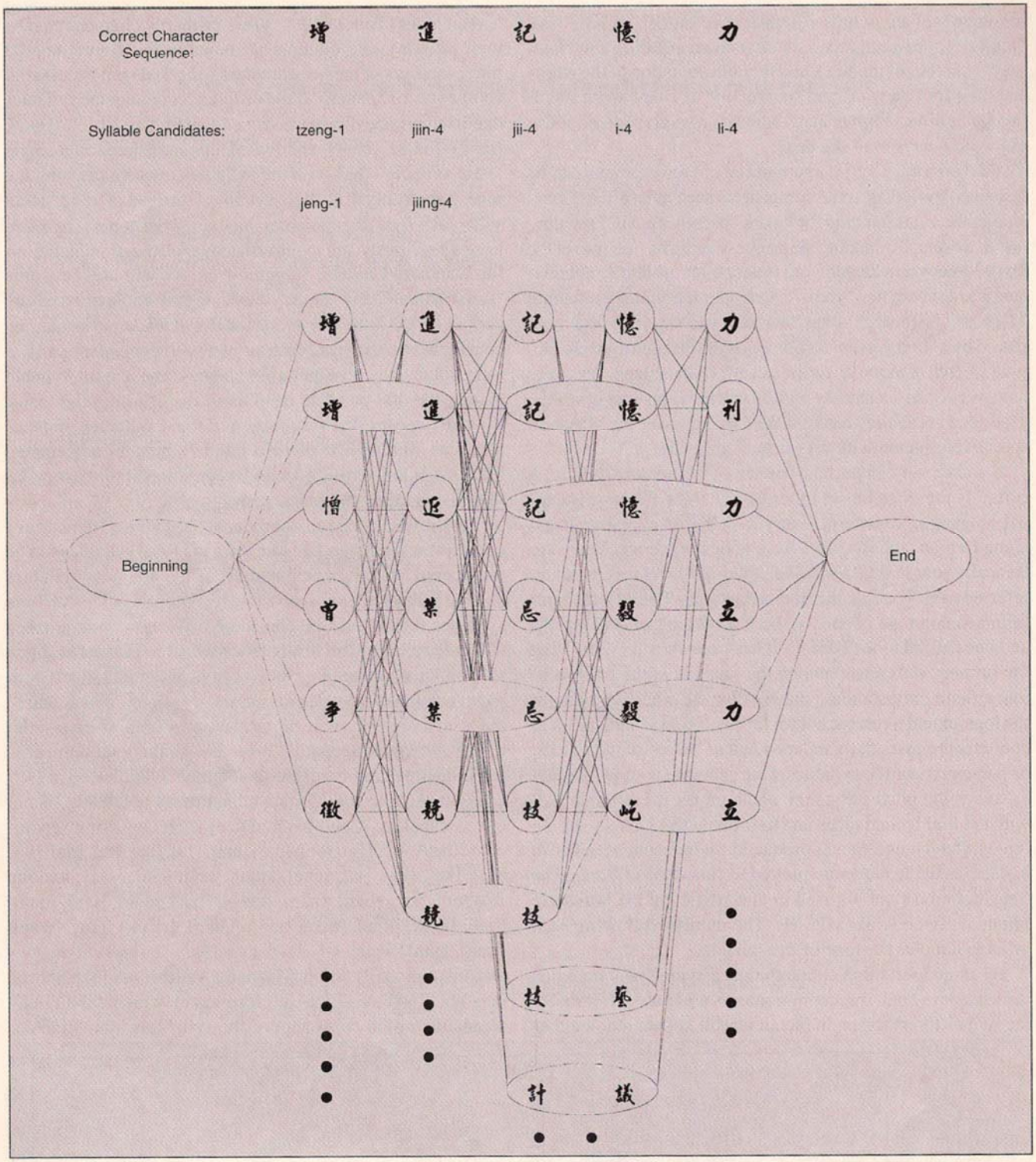
It can be found that although the carrying syllable [ba] is the same, the pitch frequency contours are quite different for the five different tones. In fact, it is well known that the four lexical tones (Tones 1, 2, 3, 4) are primarily characterized by their pitch contour patterns as shown in this figure, but the pitch contours of the neutral tone (Tone 5) do not necessarily form a specific pattern. If the differences among the syllables due to tones are disregarded, i.e., the five syllables, [ba-1], [ba-2], [ba-3], [ba-4], and [ba-5], are considered as a single "base syllable" [ba] (i.e., the syllable structure carrying the tones), then only 408 "base syllables" instead of 1345 "tonal syllables" (i.e., the syllables including the tones), as mentioned previously, are required to cover all the pronunciations for Mandarin Chinese. From now on, we will use the words "tonal syllable" and "base syllable" in this article to avoid any confusion. As a result, every tonal syllable can, in fact, be considered as the combination of two independent parts, a tone among the five possible choices and a base syllable among the 408 possible candidates disregarding the tones. This also means that recognition of tonal syllables, if desirable, can similarly be divided into two parallel procedures, i.e., recognition of the tones and recognition of the base syllables disregarding the tones, respectively.

For the first key issue, the monosyllable-based structure of the Chinese language, because the 1345 tonal syllables can be considered as being combinations of the 408 base syllables and the 5 different tones, possible recognition of the 408 base syllables thus should be considered. We now look at these 408 base syllables. First, all of the 408 base syllables are open syllabic in structure; i.e., they always end with vowels with the exception of vowels plus nasals -n and -ng. This is one of the primary reasons why the total number of these base syllables is not large. Secondly, even though the total number is not large, recognition of these 408 base syllables is, in fact, difficult because there exists a total of 38 confusing sets in this vocabulary. Good examples of such confusing sets include the A-set: {[a], [ba], [pa], [ma], [fa], [da], [ta], [na], [la], [ga], [ka], [ha], [ja], [cha], [sha], [tza], [tsa], [sa]}; and the AN-set: {[an], [ban], [pan], [man], [fan], [dan], [tan], [nan], [lan], [gan], [kan], [han], [jan], [chan], [shan], [ran], [tzan], [tsan], [san]}.

Conventionally, each Mandarin syllable is decomposed into an "INITIAL/FINAL" format, in which "INITIAL" means the initial consonant of the syllable while "FINAL"



2. The pitch-frequency contours as functions of time for the syllable [ba] with the five different tones.



3. A partial list of a simplified example word lattice, where each circle is a mono-character word while each ellipse is a poly-character word.

means the vowel or diphthong part, but including an optional medial or nasal ending. The FINAL is generally a tone-carrying part. However, because the tones are handled independently here, and only the 408 base syllables are considered, we also disregard the tones carried by the FINAL and tentatively take the FINAL as a tone-independent part in this article, except when otherwise mentioned. Each confusing set mentioned above then consists of base syllables sharing the same FINAL but with different INITIALs.

Table 1 is a list of all the 408 base syllables, where the vertical scale lists all 38 FINALs (including a null FINAL) and the horizontal 22 INITIALs (including a null INITIAL). In this table every phonologically allowed base syllable is assigned an identification number located at an appropriate square, indicating the component INITIAL and FINAL. Each row in the table then represents a confusing set (the A-set mentioned above is the second row in the table). These 38 confusing sets for the 38 rows in Table 1 constitute the major

problems to be solved in the acoustic recognition to be discussed below.

A good example is shown in Fig. 3 of the second key issue, which is the large number of homonym characters sharing the same tonal syllable and the high degree of ambiguity caused by all the possible mono-and poly-character words formed by these homonym characters for a sequence of tonal syllables. Here, assume that a sequence of five tonal syllables, [tzeng-1] [jiin-4] [jii-4] [i-4] [li-4], is correctly recognized, but that each of them is shared by many possible homonym characters, which in turn can form many mono-and poly-character words. All these word hypotheses can be used to construct a very complicated graph called a word lattice, as shown in Fig. 3, on which every path is a possible solution for the sequence of tonal syllables.

In Fig. 3, every circle represents a mono-character word while every ellipse represents a poly-character word. As can be seen, not only can a monosyllable like the second one ([jiin-4]) in Fig. 3 represent many mono-character words with different meanings such as "advancing," "close to," "forbidden," etc.(see (c1) in the box), but two adjacent tonal syllables such as the third

Table 1. The 408 base syllables in Mandarin speech. The number indicates the sequence number used in our database.

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

and fourth ones ([jii-4] [i-4]) in Fig. 3 can represent more than one bi-syllabic word such as those standing for "memory," "techniques," and "discussions" (see (c2) in box). A tonal syllable may even combine with other tonal syllables on both sides to represent different bi-syllabic words such as the fourth tonal syllable ([i-4]) in Fig. 3 with [jii-4] on its left to represent a bi-syllabic word, "[jii-4] [i-4]," which means "memory" as mentioned above, but [i-4] with [li-4] on its right to represent another bi-syllabic word, "[i-4] [li-4]," which means "standing" (see (c3) in box), etc. Practically, the problem is even much more complicated than the above because the tonal syllables are highly confusing and difficult to recognize accurately; thus, a set of top n (n=10 or 20, for example) tonal-syllable candidates will usually be considered for each syllable so that the correct tonal syllable is sure to be included in the word lattice.

In this case, each of the n candidates may represent many homonym characters, and the word lattice will be much more complicated than that in Fig. 3. For example, as also shown in Fig. 3, when the first tonal syllable ([tzeng-1]) has a confusing candidate ([jeng-1]), many mono-character words such as those standing for "fighting" and "requesting" are added to the lattice (see (c4) in box), and when the second tonal syllable ([jiin-4]) has a confusing candidate ([jiing-4]), a new bi-syllabic word standing for "competition" will be added when combined with the next tonal syllable ([jii-4]) (see (c5) in box), etc. In fact, Fig. 3 is only a very small partial list, and the real word lattice can be much larger and more complicated. Therefore, it is clear that in order to select the correct character sequence, very powerful linguistic decoding techniques are thus needed to find the best path on this lattice, as will be discussed later on.

Input Mode During Dictation

Using an alphabetic language such as English as an example, right now many very successful continuous-speech dictation prototype systems with very large vocabulary have been developed in laboratories, but most of the products accepted by the market are still primarily in the isolated-word mode [23, 36]. Apparently, for a laboratory prototype system to become a product accepted by the market, there are still many issues to be considered such as the complexity, cost, error rates, robustness with respect to different speakers, different user environmental noise, different text subjects and wording styles, different spontaneous-like speaking modes, etc. All these issues may increase the difficulties in making products for continuous-speech dictation systems with very large vocabulary commercially available. For dictation of Mandarin Chinese, similar issues apparently exist. Certainly, input in the continuous-speech mode is the most convenient, natural, fast, and attractive, and prototype systems for such continuous Mandarin speech recognition with very large vocabulary have also been developed [37-39], but for products practically acceptable to users, other input modes should also be considered.

First, taking into account the monosyllable-based structure of Mandarin Chinese previously discussed, the easiest input mode is certainly via the isolated tonal syllables; i.e., the user can produce a sentence in the form of a sequence of isolated tonal syllables separated by pauses. Although this mode of input is too awkward and almost impossible for alphabetic languages, it is in fact very feasible for a monosyllable-based language like Mandarin Chinese [40]. Because every character (produced as a tonal syllable) is a morpheme with its own meaning, every native speaker of Mandarin Chinese learns these characters produced as isolated tonal syllables one by one in school. It is, therefore, very easy and convenient, though not very natural, for a native speaker to produce a Mandarin sentence as a sequence of isolated tonal syllables. In fact, sequences of isolated tonal syllables appear to be a rather acceptable and an even more enunciated form of pronunciation in the Chinese language. Thus, it was proposed some years ago that an isolated-syllable-based recognition system is the most feasible approach to developing Mandarin dictation systems for very large vocabulary and almost unlimited texts, at least in the early stages [40]. In this way, the difficulty of handling the complicated problem of co-articulation across syllables in continuous-speech recognition can be avoided

There are also some minor reasons for using isolated tonal syllables as recognition units. For example, all Mandarin syllables are open syllabic in structure; i.e., they always end with vowels, with the exception of vowels plus the nasals -n and

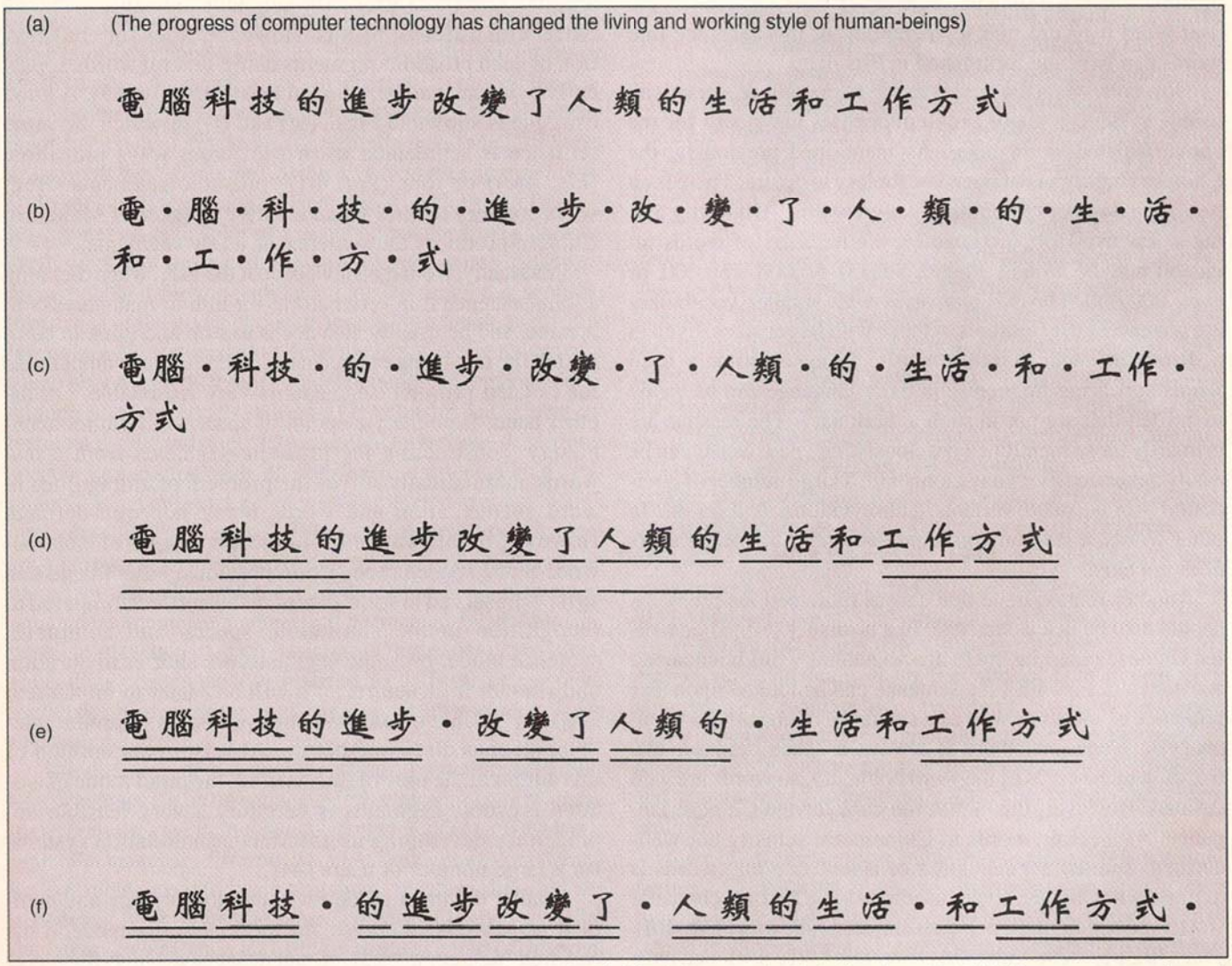
Table 2. Statistics for a lexicon of 50,000 most frequently used words.

Length of Words (number of characters)	Number of Words	Number of Different Tonal-Syllable Strings	Number of Different Base-Syllable Strings
1	4861	1157	402
2	35178	32084	24152
3	5305	5274	5251
4	4278	4267	4262
5	380	380	380
Total	50000	43162	34447

-ng, as mentioned previously. This makes endpoint detection relatively easy for isolated tonal syllables. Of course, on the other hand, technical disadvantages exist in using isolated tonal syllables as units, in addition to the fact that such an input mode is slow and not very natural. For example, the relatively small number of tonal syllables implies a very large number of homonym characters and, therefore, a very high degree of ambiguity in selecting the accurate output sequence of characters, as has been discussed previously and shown in Fig. 3. This was the rationale when several prototype systems were developed based on the isolated-syllable input mode [40-42]. As an example, a Chinese sentence with meaning "the progress of computer technology has changed the living and working style of human-beings" as shown in Fig. 4(a) can be uttered character by character (or syllable by syllable) by the user in this way during dictation, as depicted in Fig. 4(b).

On the other hand, considering the experiences with alphabetic languages, a very straightforward approach for the input mode is in isolated words [36]; i.e., the user can produce a sentence in the form of a sequence of isolated words separated by pauses. From the user's point of view, to utter a poly-syllabic word continuously as a tonal syllable string is much more natural than to do so as several isolated tonal syllables, and fewer pauses implies faster input speed as well. From the speech-recognition technology point of view, there are certainly further advantages. First, although there are many homonym characters sharing the same tonal syllable as discussed previously, the number of homonym poly-character words sharing the same tonal-syllable string is apparently much smaller, and there are many combinations of several tonal syllables that do not even correspond to any word.

To illustrate this, Table 2 presents the statistics of a lexicon composed of the most frequently used 50,000 words in daily Mandarin Chinese, in which the words are categorized according to their length or the number of component characters (or tonal syllables) in the words. The first column of the number of words indicates that about 70% of the most frequently used words are bi-character, and that about 10% are mono-character. The second column is the number of different tonal-syllable strings associated with each category of words. Here it can be found that 4861 mono-character words share 1157 tonal syllables, and 35,178 bi-character words share 32,084 tonal-syllable strings. However, for words with three or more characters, there is almost a one-to-one correspondence between the words and the tonal-syllable strings. The right column is the number of



4. Different input modes: (a) the original sentence, (b) isolated syllables (characters), (c) isolated words, (d) different segmentations of the sentence into words, (e-f) two possible partitions of the sentence into prosodic segments.

different base-syllable strings, in which it can be seen that for words with three or more characters, the tones become almost redundant because of, again, one-to-one mapping between the words and the base-syllable strings. In such a situation, the recognition of words is clearly much easier by matching the input utterances with the words in a lexicon, as long as a lexicon is given. Secondly, the word boundaries as indicated by the pauses between the utterances produced by the user represent very good information for directly finding the correct words within the very complicated word lattice. If we look at the word lattice shown in Fig. 3, when all the words are segmented by pauses between utterances, the large, complicated word lattice will be automatically broken down into a series of smaller, simpler word lattices. The problem will still be difficult but much easier to handle than that in the isolated-syllable mode. Thirdly, the words are the building blocks of the sentences and carry a plurality of syntactic and semantic information, much more than characters do. Such information can be very useful for identifying the correct sentences represented by the utterances if such information can be properly integrated into the speech recognition processes. This was the rationale when several prototype systems were developed based on the isolated-word mode [28-31]. As an example, when the sentence used in Fig. 4(a) is uttered by the user word by word, the segmentation of this sentence into words can have the form listed in Fig. 4(c).

However, when the goal is the development of a practical, usable system, a major problem appears inevitably for the above isolated-word mode. As mentioned previously, the Chinese language is of open vocabulary in nature. There is an unknown number of commonly used words. Arbitrarily taking a few available dictionaries, the numbers of words included may be around 30,000, 50,000, 80,000, 160,000, or even 200,000. Those dictionaries with smaller vocabulary are not necessarily a subset of those with larger sizes. Even in a dictionary with 200,000 words, many commonly used words appearing frequently in daily language can be easily identified that are not in such a dictionary. The reasons are primarily those mentioned previously; i.e., new words can be easily generated everyday along with a large number of compound words, word variants, abbreviations, and so on. In other words, a lexicon commonly accepted by all the users does not exist.

Another related issue that causes more serious problems should also be discussed here. In a normally printed or written Chinese sentence, there are no natural word boundaries. As shown in Fig. 4(a), the sentence can be looked upon as a sequence of words as well as a sequence of characters. In alphabetic languages, there is always a space between two words, which serves as the word boundary, so words are well defined. However, this is not the case for the Chinese language. As a result, words in Chinese are actually not well-defined, and the segmentation of a sentence into words is definitely not unique. Every user may have his own choice of words and segmentation. For example, as shown in Fig. 4(d), every line segment under the characters in the sentence indicates a reasonable choice of word.

For example, the segment of the characters representing "computer" is a word, that representing "technology" is also a word, but that representing "computer technology" is definitely a compound word (see (d1) in box). The segment of characters representing "change" is a word, but that representing "has changed" is a word variant, with the additional character reasonably taken as a function word (see (d2) in box), and so on. Apparently the way shown in Fig. 4(c) is simply one out of the many ways this sentence can be segmented into words. The nonexistence of a commonly acceptable lexicon and a fixed way of segmentation of sentences into words thus actually makes the input mode of isolated words practically useless. The users won't be able to memorize all the words chosen by the designer in the lexicon and stored in the machine and segment the desired sentences into the words listed in that lexicon.

The above problem leads to a different concept, i.e., the input mode of isolated prosodic segments instead of isolated words. A prosodic segment is an utterance easily produced by the user as a breath group, which is usually composed of a few words and is linguistically defined by syntactic or prosodic boundaries in the sentences. The prosodic segmentation of a given sentence is generally not unique. The same user may segment the same sentence twice with different results. However, rules for construction of such prosodic segments using several words apparently exist and can be found at least partially [43]. A good example is shown in Figs. 4(e) and (f), in which the same sentence is segmented in two different ways into three (Fig. 4(e)) or four (Fig. 4(f)) prosodic segments. Both ways are very natural when a native speaker of Mandarin Chinese produces the sentence as a few segments.

Practically, from the viewpoint of the user, when dictating a long sentence it is very natural for him to make breaks to breathe, and he usually also needs to stop and think in composing the text being entered next. Therefore, the input mode for isolated prosodic segments is very reasonable. On the other hand, from the viewpoint of speech-recognition technology, constructing the prosodic segments from a few words automatically solves the problem of ambiguities in word segmentation and words being not well-defined. However, in this way, most of the advantages of isolated-word-mode speech recognition previously mentioned can still be preserved to some extent in this mode. Compared to recognition in the continuous-speech and complete-sentence mode, prosodic segments are shorter in duration and simpler in structure, so it will be easier to implement and may be able to solve various problems in product development as discussed before. At least in the opinion of this author at the time of this writing, the input mode of isolated prosodic segments is certainly a very feasible approach for developing useful Mandarin dictation systems for a large number of users [44].

Finally, the input mode for continuous speech and complete sentence is, of course, the most attractive mode, if the required high complexity of technology and high degree of robustness can be taken care of, and if the cost can be kept

reasonable. There are, however, more problems to be solved in this case.

The Syllable-Based Architecture for Mandarin Dictation

Before going into the detailed technologies, here a special basic architecture for Mandarin dictation will first be presented, considering the special structure and characteristic features of the Chinese language discussed above. All the detailed technologies to be discussed below are primarily based on this architecture. A brief review for the general approaches for voice dictation of western alphabetic languages will be given first for comparison purposes.

Hundreds of different approaches have been proposed for voice dictation of western alphabetic languages with very large vocabulary. Many successful experimental prototype systems or even relevant products have been developed and either tested with satisfactory performance or made commercially available on the market, based on these approaches [13-22]. Although the details of each of these approaches can be quite different from one to another, the basic idea under these approaches may be explained by some simple common concepts. First, a set of basic acoustic units of the target language is defined, usually including phonemes, phones, or other similar phone-like-units (PLUs), subword units, or even smaller or larger units. In order to consider the coarticulation effects across such units in speech signals, some degree of context dependency is usually developed for these units, with the "tri-phone" being a very good example.

In the "tri-phone" approaches, the preceding and following phones on both sides of a given phone need to be considered when modeling the given phone, since different preceding or following phones actually make the acoustic properties of the given phone different. As a result, many different models are needed to describe the acoustic nature of a given phone with different context dependency on both sides. Hidden Markov models (HMMs) [2, 12], which are currently the most useful and popularly accepted models for describing speech signals, for each of such basic units (e.g., the "tri-phone") are then constructed with parameters trained from a large amount of speech data. When an unknown speech utterance is received, searching/matching processes between the unknown utterance and the HMMs of the basic units constructed as mentioned above are performed to identify the possible presence of such basic units in the unknown utterance. This part is usually referred to as acoustic processing/recognition in the area of voice dictation with very large vocabulary.

The outcome of the above acoustic processing/recognition processes for an unknown speech utterance may be much more complicated than a single sequence of identified basic units. For example, different possible segmentation of the unknown utterance into basic units may result in different sequences of basic units, the same segment of speech signal in the unknown utterance may be identified as a part of different basic units, and many basic unit candidates may be identified for the same segment of speech signal in the unknown utterance. In most approaches a lexicon of all possible words is used here. Each word in the lexicon is represented as the concatenation of the basic units chosen.

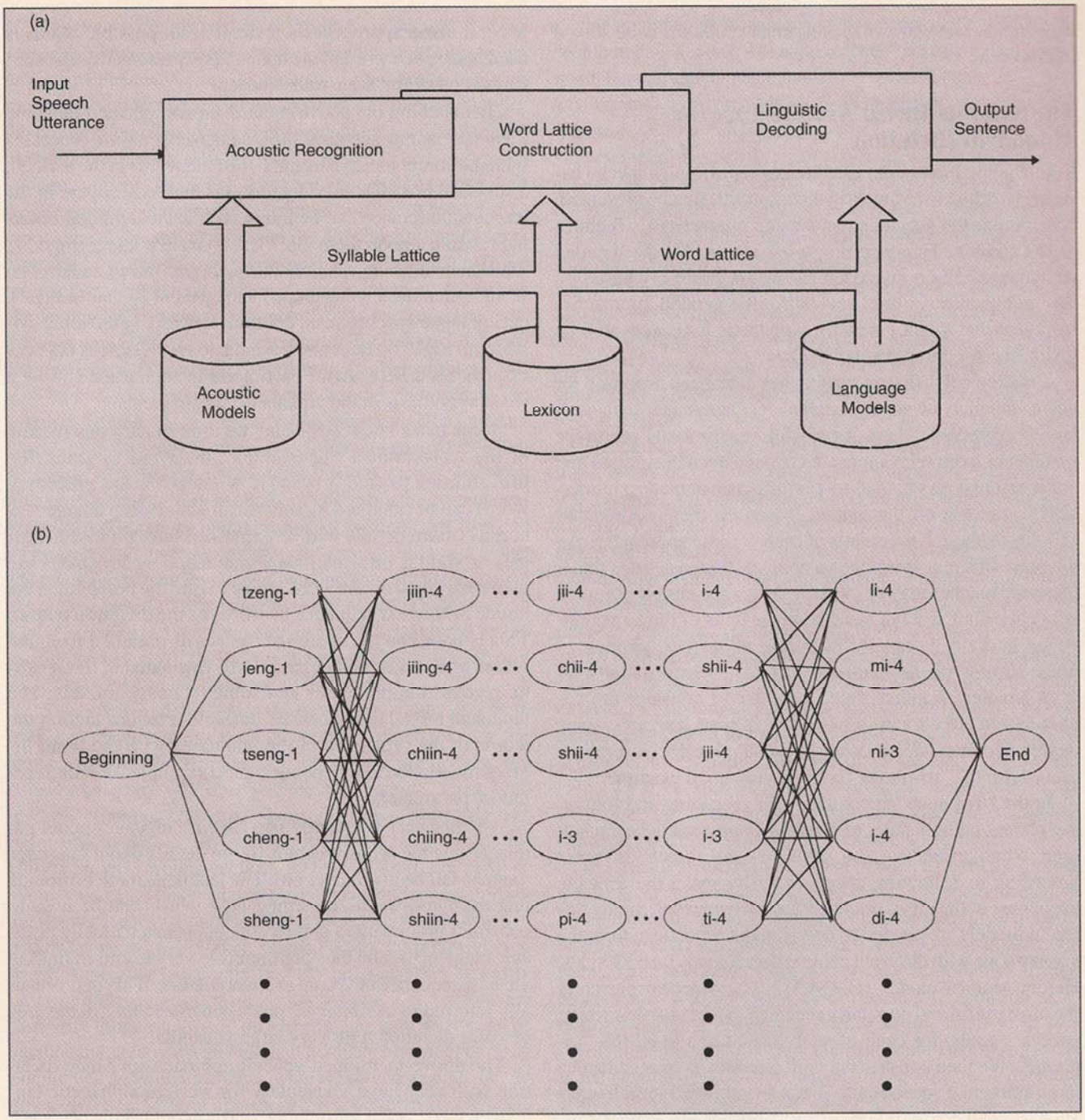
By matching the above-mentioned complicated outcome obtained in the acoustic processing/recognition processes with the component basic units in all the words in the lexicon, a set of all possible word hypotheses that may appear in the unknown utterance can be found with some temporal ordering relation among them. So, these word hypotheses together with their ordering relation can be organized to construct a word lattice (or a word graph) very similar to that in Fig. 3, except those synchronized columns defined by the tonal syllables in Fig. 3 do not exist since the tonal syllable is not necessarily used here. Also, each node in the lattice is now a word hypothesis in the alphabetic language.

Some extra knowledge may be used in the construction of the word lattice. For example, the word bi-gram (the probabilities to find a specific word given the preceding word) can be obtained by statistically analyzing a large text corpus (more details will be explained later on in this article) or similar information. Some decoding processes are then performed on the word lattice to find the best path in the lattice or the best sequence of words as the dictation output. This is based on the scores of the words obtained from the scores of the component basic units evaluated in the acoustic recognition processes and some language models. The language model describes the possible relations among the words in the sentences of the language, with the word bi-gram mentioned above being a good example for language-model parameters.

Another very frequently used set of language-model parameters is the word tri-gram (more details about language models will be given later on). The language models provide linguistic constraints regarding how words can be used to construct grammatical and reasonable (or statistically probable) sentences, and they are therefore very useful in finding the best sequence of words in a word lattice. This part is usually referred to as linguistic processing/decoding in the area of voice dictation with very large vocabulary.

The above-mentioned general approach for voice dictation with very large vocabulary for western alphabetic languages looks quite reasonable, and it would probably work equally well with Mandarin Chinese, too. However, it is believed that with the special structure and various characteristic features of Mandarin Chinese, better recognition architecture, different from the above, considering such structure and features may be found. Systems developed based on such a special recognition architecture may perform much better in various aspects including computational load and accuracy and robustness with respect to different variabilities.

First of all, the monosyllable-based structure is a unique feature of Mandarin Chinese. As mentioned previously, each tonal syllable represents many homonym characters, each of which is almost always a morpheme with its own meaning, and combinations of several of these characters give an open vocabulary of almost unlimited words. As a result, the tonal



5. (a) The syllable-based architecture for voice dictation of Mandarin Chinese. (b) The simplified time-aligned syllable lattice obtained in the acoustic recognition processes. In some cases the tonal-syllable candidates are not time-aligned as shown here.

syllable in Mandarin Chinese carries a plurality of linguistic information, which is never true in other alphabetic languages. In other words, in alphabetic languages a single syllable alone usually doesn't mean anything except for special cases, but in Mandarin Chinese each single tonal syllable alone is the pronunciation of many characters, and each character has its own meaning.

Secondly, the tonal syllable in Mandarin Chinese is of some kind of "equal distance" to the other two important linguistic units, the phone (or other PLUs) and the word. As will be described in more detail later on, a Mandarin tonal syllable is most frequently composed of two to four phones, while, as discussed previously, most frequently used Chinese words are composed of two to four characters (or syllables). Some special tonal syllables include only one phone, while some special words include only one character (or tonal syllable). Because the phone (or other PLUs) is the basic unit for acoustic processing/recognition and the word is the basic unit for linguistic processing/decoding, an intermediate unit of tonal syllable carrying plurality of linguistic information makes great sense in voice dictation with very large vocabulary. For example, the construction of word lattices based on possible candidates of tonal syllables can be much more reliable with much less ambiguity than on possible candidates of phones.

Thirdly, also mentioned previously, all Mandarin tonal syllables are open syllabic with a very simple structure, i.e., an INITIAL followed by a FINAL. Also, the total number of phonologically allowed Mandarin tonal syllables is relatively limited (only 1345). Both of these make the recognition of Mandarin tonal syllables a practical, reasonable, and feasible task, although the confusing sets in the base syllables also make the recognition of base syllables relatively difficult.

Finally, since each character has its own meaning, word boundaries within Chinese sentences do not exist, and since the words are not well defined in Chinese language, the character itself becomes a very good unit to develop language models for Chinese language. In other words, other than using word bi-grams or word tri-grams and so on, the character bi-gram or character tri-gram (i.e. the probabilities to find a specific character given the preceding one or two characters) and so on are found very useful as well. In this sense a tonal syllable (or character) in Mandarin Chinese sometimes really corresponds to a word in alphabetic languages. All these discussions lead to the special recognition architecture presented below, which is a syllable-based architecture significantly different from those for alphabetic languages.

The syllable-based architecture for voice dictation of Mandarin Chinese is shown in Fig. 5(a). The primary difference is that here the purpose of the acoustic recognition processes is to identify the presence of the tonal syllables, instead of the phones (or other basic units), in the input speech utterance, since the tonal syllable in Mandarin Chinese carries so much linguistic information and all further processing should be based on these tonal syllables. Of course, on the other hand, the acoustic recognition processes can still be based on HMMs of phones (or other basic units) just as other alphabetic languages. The only difference is that the intermediate unit of a tonal syllable is produced because these tonal syllables make great sense due to the special structure of the language. Due to the high degree of confusion among the base syllables as mentioned previously, each tonal syllable is recognized as a set of possible tonal-syllable candidates and each candidate can have some acoustic recognition score. A tonal-syllable lattice can then be constructed using these tonal-syllable candidates as shown in Fig. 5(b), in which every column of nodes are the candidates for a tonal syllable. Therefore each path on the lattice represents a possible tonal-syllable sequence for the input speech utterance. There can be many different situations for the tonal-syllable lattice here depending on different input modes and the exact techniques used in the acoustic-recognition processes. For example, if the input mode is in continuous speech, some recognition techniques may result in insertion/deletion of some tonal syllables, so all the different paths in the lattice may not include the same number of tonal syllables as shown in Fig. 5(b) and the tonal-syllable lattice may be more complicated. The tonal-syllable lattice can then be transformed into a word lattice primarily based on the words in the lexicon.

Linguistic decoding processes primarily based on the language models are finally performed on the word lattice. The latter parts seem very similar to those for alphabetic languages, but can in fact be quite different as well. As has been discussed previously, since the tonal syllables directly correspond to characters with meaning, not only the word lattice construction from the tonal-syllable lattice is straightforward, but different versions of language models based on words and characters can be easily applied and integrated, as will be clear later on. Also, as shown in Fig. 5(a), although the primary knowledge sources for acoustic recognition, word-lattice construction, and linguistic decoding are, respectively, the acoustic models, the lexicon, and the language models, they can apparently be cross-referenced to achieve better results. For example, the extra knowledge from the lexicon and the language models can definitely help in the acoustic-recognition processes, etc. Moreover, even if the architecture shown in Fig. 5(a) seems to have three distinct stages, it doesn't have to be implemented as consecutive stages. For example, some processes may be overlaid together with corresponding knowledge sources properly integrated. In the extreme case it is even possible that all the processes shown in Fig. 5(a) can be performed in a single stage using all available knowledge sources including the acoustic models, the lexicon, and the language models, as long as a good design for such a single-stage recognition mechanism can be developed.

The syllable-based architecture for Mandarin dictation had been proposed very early [45]. Today this architecture still serves as the common concepts under the many different approaches used by almost all research groups working on this problem (with their works known to the public), although probably with completely different detailed techniques to implement these common concepts. This architecture is applicable regardless of the input mode chosen and the advances in various speech-recognition technologies because the special structure of the Chinese language never changes. In the following sections we discuss the acoustic recognition processes to identify the tonal syllables in the input unknown speech utterance followed by word lattice construction and linguistic decoding.

The Recognition of Tones

The recognition of the tones for the tonal syllables will be discussed first in this section. This is always a very special part of Mandarin speech recognition because in most other languages tones don't have lexical meaning and are not necessarily considered. Because there are only five different tones (four lexical tones plus a neutral tone), recognition of the tones is generally not too difficult although very high accuracy is not easy to achieve. Substantial work on this problem was started quite early [40, 46-49], and only a few examples will be summarized here. In general, both hidden HMMs [12] and neural networks [31] are almost equally successful with very similar performance, and some other approaches also work well. The four lexical tones are usually easier to recognize while the neutral tone introduces most of the ambiguities.

As mentioned previously, unlike the four lexical tones, the neutral tone does not have a specific pitch pattern; thus it is easily confused with the other four lexical tones. Because syl-

lables with the neutral tone usually are shorter in duration and lower in energy, short-time energy is found useful in many studies in addition to the apparent key features derived from pitch-frequency contours. Many different feature vectors have been used in various studies, and the following is simply a typical example:

$$v = [p_{t+1} + p_t, p_{t+1} - p_t, e_{t+1} + e_t, e_{t+1} - e_t] \quad (1)$$

where p_t is the logarithmic value of the pitch frequency at frame t, and e_t is the logarithmic value of the short-time energy at frame t. Apparently, the first component represents the level of the pitch frequency while the second is the local slope in the pitch frequency. These two features alone already provide a certain degree of recognition accuracy if a reasonable approach is taken, while the last two components about short-time energy are very helpful for improving performance.

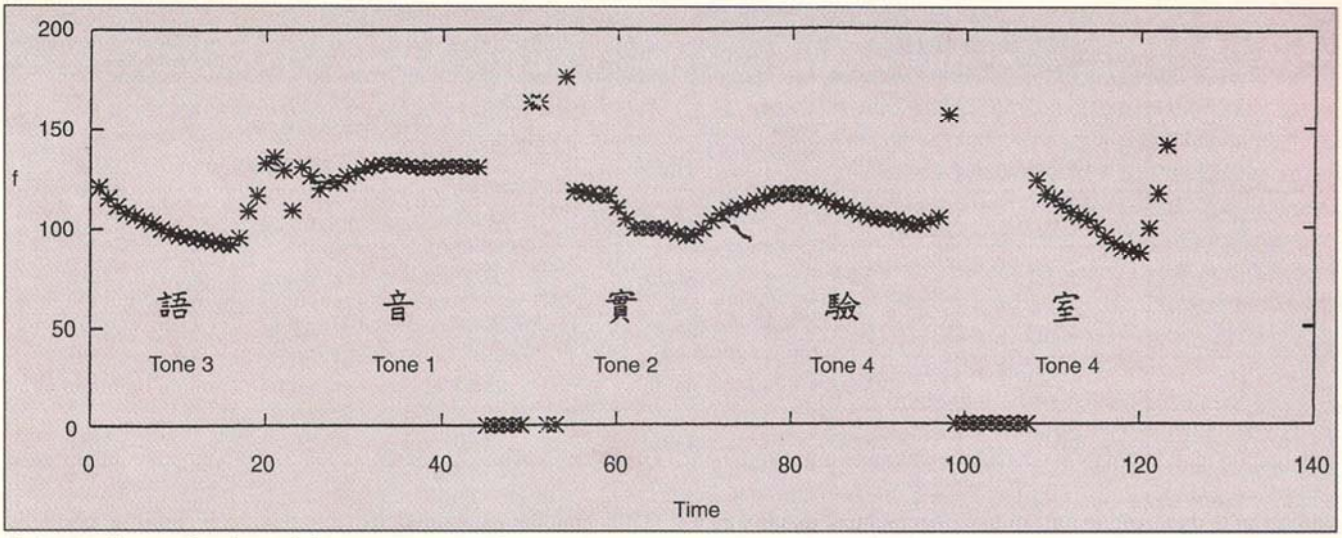
If the input is in the isolated syllable mode, both HMMs and neural networks give very good results. The exact accuracy depends on the training/testing speech database and the detailed approaches. As a typical example, using training speech of less than 1 minute, the recognition accuracy of 97.3% or higher [49] can be achieved using standard HMM techniques without special tuning in speaker-dependent tests, i.e., the machine is trained by the voice of a single speaker and tested with different voices produced by the same speaker. The results are then averaged over a group of speakers. If the input is in the isolated-word mode, the situation is still relatively simple. As mentioned previously and shown in Table 2, for words with three or more characters, the tones are actually redundant, and there is really no need for tone recognition. For mono-character words, tone recognition is simply the same as for isolated syllables. The only different situation is for bi-character words, in which the complicated tone-sandhi and co-articulation effects do make the tone behavior relatively complicated. However, because there are only five different tones and two syllables, 552 context-dependent models will be sufficient at most, if HMMs will be used for example.

If the input is in the isolated-prosodic-segment (i.e., a concatenation of a few words) mode or continuous-speech and complete-sentence mode, the situation becomes complicated, as is well known, by the tone sandhi and co-articulation effects and so on. A good example is shown in Fig. 6, in which the pitch frequency contour of a prosodic segment is plotted. Compared with the pitch contours of isolated syllables in Fig. 2, apparently, the tone behavior is quite dependent on the right and left contexts as well as on the prosody and intonation of the whole utterance. If we try to consider all the context dependency and assume that each possible tone concatenation combination needs a context-dependent model, then a total of 175 models will be needed, i.e., 5^3 (for syllables in the middle of a sentence) + 5^2 (for syllables at the end of a sentence) + 45 (at the beginning of a sentence, because the neutral tone never appears at the beginning of a sentence) + 5 (isolated models). However, practically, this number can be significantly reduced if the special characteristics of tone behavior can be carefully considered. For example, both Tones 1 and 2 end high with similar levels, and both Tones 3 and 4 end low with similar levels. This makes the influence of Tones 1 and 2 on the following tones very similar, which has in fact been observed empirically, and so on. With the aid of such human knowledge obtained from empirical observations, a hybrid approach integrating human knowledge into a statistical algorithm to automatically merge and tie the HMMs [50, 51] can be used, so some of the feature distributions can be shared. In a typical example out of many similar good results, a total of 23 context-dependent tone models was found to give very good recognition performance [37, 49]. This set of 23 tonal models is listed in Table 3 as an example result. It can be seen that, in this case, four models will be needed for Tone 1, each of which represents a typical pattern of tone concatenation with right and left neighbors, etc. The actual recognition rate again depends on the training/testing speech database and the detailed approaches used. In a typical example, using training speech of roughly 6.4 minutes, the recognition accuracy of the tones in continuous-speech

Table 3. The 23 context-dependent tone models.

Tones	Tone 1	Tone 2	Tone 3	Tone 4	Neutral Tone
Number of Models	4	6	6	4	3
Typical Tone Concatenation Combinations	1 1-(2) (3)-1 (3)-1-(2)	2 2-(2) (1)-2 (1)-2-(2) (3)-2 (3)-2-(2)	3 3-(1) (1)-3 (1)-3-(1) (3)-3 (3)-3-(1)	4 4-(1) (3)-4 (3)-4-(1)	5 (1)-5 (3)-5

For example, in the first column there are 4 different context-dependent tone models for Tone 1, where 1 represents the standard pattern for Tone 1, 1-(2) represents the Tone 1 model at the sentence beginning followed by Tone 2, (3)-1 the Tone 1 model at the sentence end preceded by Tone 3, and (3)-1-(2) the Tone 1 model in the middle of a sentence preceded by Tone 3 and followed by Tone 2; in the second column there are 6 different context-dependent tone models for Tone 2, etc.



6. A typical example of a pitch-frequency contour for a continuous utterance in a prosodic segment.

and complete-sentence mode can reach 89.8% or higher in speaker-dependent tests [49].

Recognition of Base Syllables in Isolated-Syllable Mode

Because isolated syllables may be a feasible input mode for dictation and since this is, in fact, a very special feature of Mandarin Chinese, recognition of the 408 base syllables in the isolated-syllable mode is discussed here first. This problem has been investigated by many research groups, but here we simply choose two typical examples. In the first example, very carefully trained, delicate, continuous HMMs (CHMMs) are used for all 408 base syllables. As mentioned previously, the primary difficulty here is caused by the existence of the 38 confusing sets as represented by the 38 rows in Table 1. In each of these confusing sets of base syllables having the same FINAL but different INITIALs, the INITIAL parts are usually very short compared to the FINAL parts in the base syllables. Therefore, any important differences among the INITIAL parts of different base syllables can very often be easily swamped by the irrelevant differences among the FINAL parts when the computation of the HMMs goes through the FINAL parts.

An example approach to this problem is then to train the INITIAL models and FINAL models separately, and then to cascade them together into the 408 base-syllable models. One may further make the INITIAL HMMs right-context dependent based on the beginning phoneme of the following FINALs, considering the fact that the acoustic properties of an INITIAL are highly dependent on the beginning phoneme of the following FINAL, but make the FINAL HMMs context independent. The resulting 408 base syllables are shown in Fig. 7, in which 113 right-context-dependent INITIAL models extended from the 22 INITIALs are cascaded with 38 context-independent FINAL models to form the 408 base-syllable models [40, 52-54]. In this way, the INITIAL and FINAL models can be separately trained and optimized and the very short INITIAL parts can be assigned a larger number of states. Also, the base-syllable models for the base syllables in a given confusing set can have exactly identical parameters in the last few states. As a result, the effect of the FINALs in the recognition phase can be minimized while the difference in the INITIALs can be emphasized to better distinguish these base syllables. The actual performance of this approach depends on detailed modeling methods and the training/testing speech database. As a typical example, using training speech of roughly 10 minutes in speaker dependent tests, a high top-1 accuracy on the order of 93.8% or higher and a top-5 inclusion rate (the probability that the correct base syllable is within the top 5 candidates selected in the recognition process) on the order of at least 98.5% can be obtained in isolated syllable mode with carefully tuned CHMMs [54].

Although the CHMM-based approaches mentioned above have achieved very successful recognition rates, they suffer from not only a highly intensive computation load in both the training and recognition phases, but also a time-consuming process of human-aided segmentation of the training data in order to emphasize the discriminative INITIAL part of each base syllable. To meet the low-cost, real-time implementation requirements for a practically useful Mandarin dictation system, and in order to make on-line adaptation for different users practically feasible, it is highly desirable to have some other approaches that can reduce the computation load significantly and make the training process easier, without sacrificing recognition accuracy. A typical approach in this direction specially developed for isolated Mandarin base-syllable recognition, referred to as the segmental probability model (SPM) [41, 55], will be presented here as a good example of several similarly successful approaches. This model can be viewed as a modified version of a CHMM with the state transition probability matrix abandoned and the utterances for the isolated base syllables simply equally segmented by the states. Considering that isolated Mandarin base syllables have relatively simple phonetic structures, and that the primary problem in such a recognition task is to distinguish each base syllable from the others instead of decoding it into a few phonemes with their boundaries, it is therefore reasonable to assume that an optimal sequence de-

Table 4. The average recognition rates for isolated syllables in speaker-dependent tests when CHMMs and SPMs were used with different model configurations.						
M	1		2		3	
N	CHMM	SPM	CHMM	SPM	CHMM	SPM
3	70.63	70.03	80.15	80.41	82.27	83.53
4	72.27	72.27	81.55	83.27	84.31	84.91
5	75.65	75.28	82.90	84.09	84.31	85.28
6	77.70	77.29	84.24	83.57	83.98	84.68
7	80.07	78.62	84.68	84.20	84.37	84.91

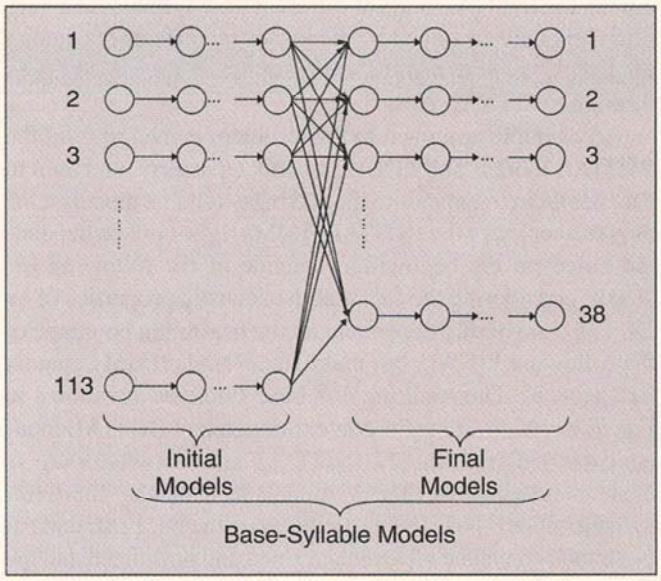
coding and dynamic programming procedure, usually performed in traditional CHMM approaches, is in fact not necessary. A simple deterministic state sequence defined by equal-length segments will be adequate for recognition of isolated Mandarin base syllables. The computation load can thus be reduced significantly, and the training process also made much easier. The basic concept of the SPM and a comparison with a CHMM is illustrated in Fig. 8.

Considering the fact that SPM can be viewed as a simplified version of CHMM, a very intuitive guess for the performance comparison is that the accuracy of SPM will be more or less degraded as a natural price paid for reduced complexity. However, extensive experiments, in which different model configurations for both CHMM and SPM were tested under exactly the same conditions, show that this is not true. Some typical results are presented here, which indicate that there is really no meaningful difference between the performance of the SPM and CHMM as far as recognition of isolated Mandarin base syllables is concerned.

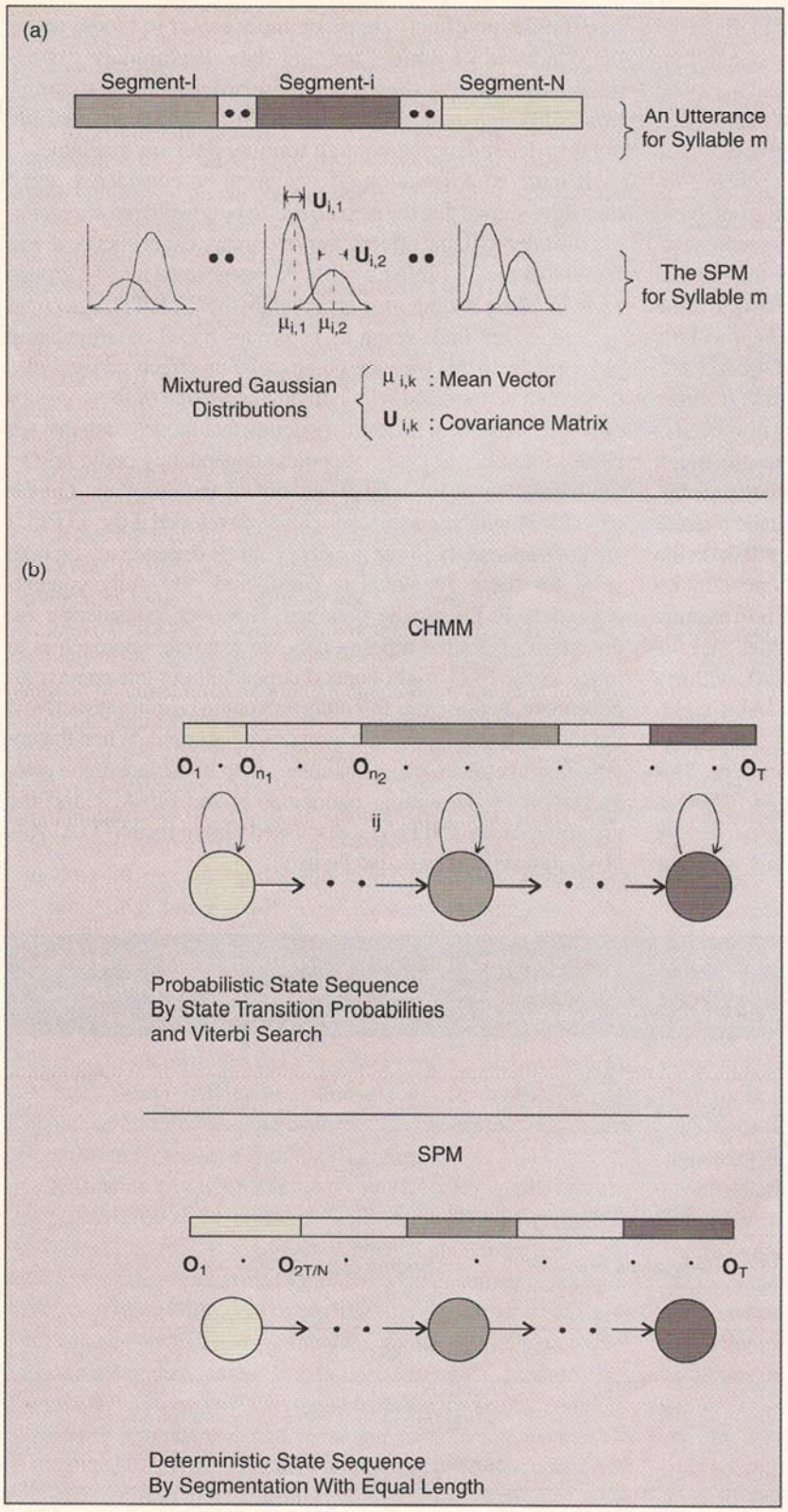
These experiments were performed in the speaker-dependent mode, with roughly 18 minutes of training speech for each speaker. The state number, N, was changed from 3 to 7, and the mixture number, M, was changed from 1 to 3. The average top-1 recognition rates for all the speakers are listed in Table 4. Note that the numbers for the CHMM in Table 4 are much lower than the number of 93.8% mentioned in the first paragraph of this section because a much smaller number of mixtures $1 \le M \le 3$ was used here to reduce the computation requirements, and no special approaches or fine tuning on the models such as those previously discussed were used. From the table, it can be seen that the achievable recognition rates for both the CHMM and SPM are in fact very close to each other if the same model configurations are used. For example, in the simplest case (N=3, M=1) the CHMM can achieve 70.63% and the SPM can achieve 70.03%, with the latter being only very slightly worse; and in the most complex case (N=7, M=3), the CHMM can achieve 84.37% and the SPM can achieve 84.91%, with the latter being slightly better.

These results verify the concept that a deterministic but properly specified state sequence can perform as well as the optimal state sequence found by the Viterbi search algorithm. This can be explained by the relatively simple phonetic structure of the target vocabulary of isolated Mandarin base syllables. Each Mandarin base syllable is composed of at most three to four phonemes, and the phonetic structure is simply an INITIAL/FINAL format as mentioned previously. This conclusion is, of course, limited to the tested vocabulary of isolated Mandarin base syllables only and is not necessarily extendible to other vocabulary with more complicated phonetic structures.

Although the rates achievable here, say 80.41% for N=3 and M=2, are not satisfactory, apparently many specially designed approaches can be applied to fine-tune the SPM models and improve the accuracy, just as for the case of the CHMM. Because the most discriminating parts of these base syllables are in the INITIAL parts, the analysis frame window can be shifted much more slowly in the beginning parts of each utterance than in the remaining parts, and the likelihood values obtained from the beginning parts of each utterance can be further weighted to emphasize the INITIALs [55]. Other helpful techniques such as the discriminative training approach



7. The specially trained CHMMs for the 408 base syllables by cascading 113 INITIAL models and 38 FINAL models.



8. The segmental probability model (SPM) for isolated Mandarin syllable recognition: (a) basic concept, (b) comparison with the CHMM.

based on the generalized probabilistic descent (GPD) method [56-58] can be used as well.

On the other hand, the speed of SPM recognition can also be improved significantly with an approach called Fast-SPM. For M=1 (i.e., 1 mixture for each segment), the evaluation of the log likelihood function can be simplified significantly, but in that case, top-1 accuracy will be low, though the top-10 inclusion rate is almost 100%. Fast-SPM is, therefore, a two-stage architecture in which M=1 models are used to select the top-10 candidates out of 408 in the first stage for further consideration in the second stage; in the second stage the delicate models are used to choose the output syllable from the top 10 selected in the first stage. With all these approaches properly designed and applied, extensive experiments indicate that with the same training data (i.e., roughly 18 minutes of training speech for each speaker) and testing conditions (i.e., isolated syllables and speaker-dependent mode) as before, the recognition accuracy can be raised to as high as 95.4%, but at a speed roughly 45 times faster than CHMMs [55]. Still another nice feature of the SPM is that it is also fast in training and flexible in adapting to various conditions due to its very simple structure. This makes on-line learning and fast speaker adaptation possible.

Recognition of Base Syllables in Continuous Utterances

For input modes other than isolated syllables, i.e., modes with isolated words, prosodic segments, or complete sentences, recognition of base syllables will be essentially the same because in all these cases, the base syllables are simply carried in a continuous utterance of a base-syllable string. The only difference is that the continuous utterance can be shorter or longer. So, here, we will present recognition of base-syllable strings in such continuous utterances, regardless of whether the utterances are poly-character words, prosodic segments, or complete sentences. Similar approaches have been studied by many research groups with similar results obtained, but here we simply summarize some typical examples [28-31, 37, 44].

As discussed before, Mandarin syllables are traditionally decomposed into INITIALs and FINALs, with a total of 22 INITIALs and 38 FINALs. Furthermore, these INITIAL/FINALs can be further decomposed into smaller PLUs. It has been found that a total of 33 PLUs will be enough to transcribe the 408 Mandarin base syllables. The phonological hierarchy of a Mandarin syllable discussed here is illustrated in Table 5(a), where the relationships among tonal syllables, base syllables and tones, INITIAL/FINALs, and PLUs are shown. The 33 PLUs for Mandarin Chinese with the IPA (International Phonetic Alphabet) representations are listed in Table 5(b). It can be found that an INITIAL is always a PLU while a

FINAL generally may contain one, two, or three PLUs. That is, a Mandarin base syllable is composed of from one to four PLUs, including the situation of a null INITIAL.

With the above information, special efforts have been made in selecting the most appropriate sub-syllabic acoustic units for base-syllable recognition in continuous speech, considering the condition of limited training data on the one hand (currently a commonly accepted, large enough speech database for Mandarin speech does not exist) and the special mono-syllabic characteristics of Mandarin Chinese on the other hand. First of all, the INITIAL/FINALs seem to be a good choice considering the basic structure of Mandarin syllables. In order to consider the condition of limited training data, a general observation on continuous Mandarin speech is that the co-articulation effects within a syllable are much more significant than those across syllables due to the mono-syllabic structure. Also, within a syllable the acoustic characteristics of the INITIAL are certainly highly dependent on the FINAL, but those of the FINAL are much less dependent on the INITIAL. With this in mind, a good approach is to assume that both the co-articulation effects across syllables and the dependence of the FINAL on the preceding INITIAL within a syllable are negligible, and to make the INITIALs right-context dependent on the beginning phoneme of the following FINAL and make the FINALs context independent. This gives something like a set of 113 INITIALs and 38 FINALs, very similar to what was done in the CHMM approaches for isolated syllables as mentioned previously. In this way, the co-articulation effects are much more easier to model under the condition of limited training data. Preliminary experimental results have shown that these are reasonable assumptions although it is always better to consider all possible context dependency if enough training data are available.

If more co-articulation effects are to be considered, some tests have shown that the next effects to be considered are probably the inter-syllabic effects. In preliminary experiments, it was found that the 38 FINALs could be categorized into 12 groups based on their ending phonemes while the 22 INITIALs could be categorized into seven to 11 groups based on articulation phenomena. In this way, the co-articulation effects across syllables could be modeled with human knowledge of these groups plus some statistical acoustic modeling techniques such that the reduced number of cases of context dependency could relieve the requirements for a large amount of training data. On the other hand, similar approaches can be developed if the 33 PLUs are used instead. If all the possible context dependency on both sides for these 33 PLUs is considered, 511 fully context-dependent PLUs can be obtained. However, considering the condition of limited training data, an example approach is to make all the PLUs right-context dependent but left-context independent. In that case, 149 units turn out to be quite attractive if the inter-syllabic effects are temporarily ignored. When the co-articulation effects across syllables are to be included, the categorization of the ending phonemes of the FINALs and the grouping of the INITIALs discussed above for INITIAL/FINAL approaches can also be used.

Table 5. (a) The phonological hierarchy of Mandarin syllables, where the number inside every bracket indicates the total number of that kind of unit in Mandarin Chinese. (b) The 33 PLUs for Mandarin Chinese represented in the International Phonetic Alphabet (IPA).

(a)

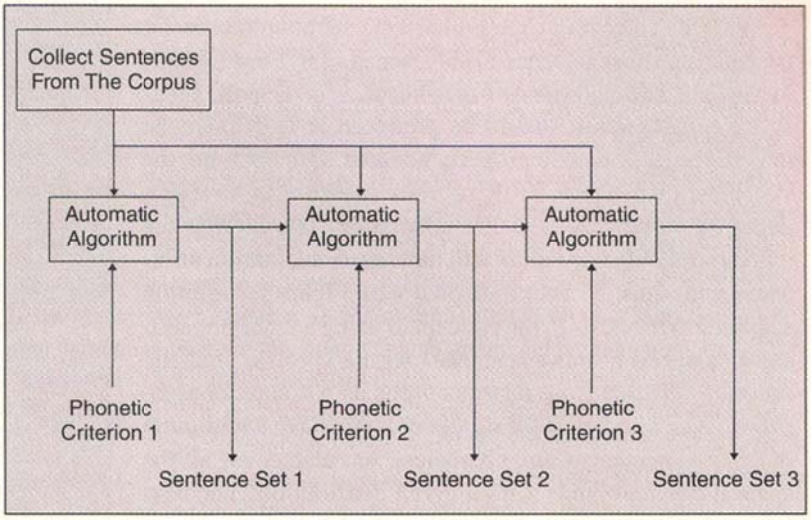
<table>
<tr><td colspan="4">Tonal syllable (1345)</td></tr>
<tr><td colspan="3">Base Syllable (408)</td><td rowspan="2">Tone (5)</td></tr>
<tr><td rowspan="2">INITIAL (22)</td><td colspan="2">FINAL (38)</td></tr>
<tr><td>Medial (3)</td><td>Nucleus (9)</td><td>Ending (2)</td></tr>
</table>

(b)

	IPA
Stop (6)	[p] [t] [k] [p′] [t′] [k′]
Affricate (6)	[ts] [tʂ] [tɕ] [ts′] [tʂ′] [tɕ′]
Nasal (3)	[m] [n] [ŋ]
Liquid (1)	[l]
Fricative (6)	[f] [s] [ʂ] [ɕ] [x] [ʐ]
Vowel (10)	[a] [o] [ɤ] [e] [i] [u] [y] [ɪ] [ɿ] [ɚ]
Null Phone* (1)	

*The Null phone is used for representation of the null INITIAL

When the basic sub-syllabic units have been chosen, the rest of the work is not that much different from that done for alphabetic languages. For example, the standard segmental k-means algorithm [59, 60], including some minor modifications, can be used in the training processes, with some approaches applied to merge or tie the mixtures/states/models for better acoustic modeling considering human knowledge as discussed above. On the other hand, in the recognition processes many algorithms are available to search through all the possible paths and find out the best base-syllable sequences; the frame synchronous network search and the tree-trellis search are two typical examples [61-64]. In general, all these approaches are equally as applicable to Mandarin Chinese as they are to other alphabetic languages, as long as the appropriate sub-syllabic acoustic units are chosen and the acoustic models are well trained. The recognition results very much depend on the degree of context dependency defined on the sub-syllabic units and the extent to which the model parameters are well trained using the available data. For example, if the INITIAL/FINALs are to be used, the set of 113 right-context-dependent INITIALs and 38 context-independent FINALs mentioned above give very good results with limited training data. More context dependency can always be included in modeling, but the achieved recognition accuracy then generally reflects the tradeoff between two factors: improved accuracy due to finer modeling and degraded performance due to inaccurate estimates of the model parameters when the training data are not sufficient. This is because more units are used and more model parameters are needed. When sufficient training data are available, the former dominates and the accuracy can be improved; otherwise the performance may be degraded.



9. The three-stage incremental sets of phonetically balanced training sentences automatically selected from a large corpus.

A similar situation exists if the PLUs are used. In general, the 149 right-context-dependent but left-context-independent PLUs mentioned above give very good results with limited training data. More context dependency can be included in modeling and the number of units can range from 149 to 511 as also mentioned above, with recognition performance again reflecting the tradeoff between the above two factors. If the context dependency can be properly defined with respect to the available training data, there is really no meaningful difference in the recognition performance between the choice of INITIAL/FINALs or PLUs. Not only can the achievable accuracy be almost the same, but the model size, total state number, and search speed can all be almost identical. The recognition accuracy actually achieved again depends on the detailed techniques chosen and the database used for training and testing. As a typical example, using training speech of roughly 16.4 minutes in speaker-dependent tests for base-syllable recognition in continuous-speech and complete-sentence mode, an accuracy of 88.3% or higher can be achieved using 113 INITIALs and 38 FINALs as the basic unit [37].

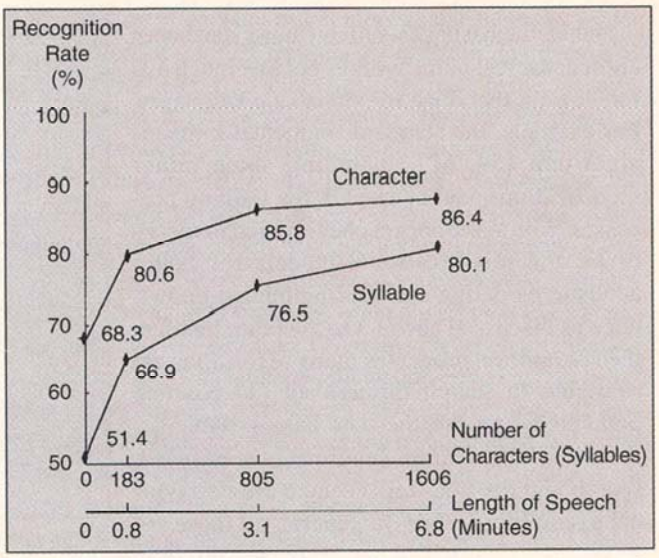
Incremental Speaker Adaptation by Phonetically Balanced Sentences

In order for Mandarin dictation systems to be widely used, efficient speaker adaptation functions [65-67] are certainly necessary because most users will not be ready to spend a very long time producing enough training data for a speaker-dependent system (i.e., the system is trained by the voice produced by the user himself only). In fact, this is also one of the most difficult barriers in developing marketable dictation systems. Because the dictation system has to accept a very large vocabulary and almost unlimited texts instead of a special application or a finite set of vocabulary, speaker independence (i.e., the user doesn't produce any voice to train the system and the system is trained by the voice produced by a group of other people) is achievable only with relatively low accuracy, while the amount of training data required for a speaker-dependent system will generally be very large. When a system requires too long a time for users to produce adequate training data, users may simply decide to give up on using the system. Therefore, speaker adaptation is necessary in which the speaker-independent system trained by a large group of people with relatively low accuracy for a new user is adapted to a new user with much higher accuracy using only a limited amount of training data produced by the new user. A good idea for this approach is to develop incremental adaptation processes such that accuracy can be improved step by step. In this way, the user can find that the system is learning his voice gradually, or he can start to use the system earlier at the price of tolerating a relatively higher rate of error. With this purpose in mind, the training data for a new user needed in the adaptation processes can be organized into a few stages, each bearing some special phonetic features, so that accuracy can be significantly improved after each stage of training data is produced [68].

With the above concept, multi-stage adaptation texts for the training speech need to be developed first. For example, in the first stage, all the sub-syllabic units used in the base-syllable recognition should be produced at least once, so that all the units can be properly adapted. Furthermore, the statistical distribution of these units should also be reproduced to some extent in the adaptation data so that the more-frequently used units will have more adaptation utterances and, thus, be better adapted with higher recognition accuracy. This leads to the concept of using a set of phonetically balanced training sentences with a specific phonetic criterion selected from a large corpus by a computer algorithm. Such a sentence set should not only have a minimum number of sentences and characters, but also cover all the desired phonetic units with a given distribution. The next few stages can then be developed with a similar concept but with different phonetic criteria. For example, the second stage may cover all the 408 base syllables with a desired statistical distribution. The third stage may cover the top 600 most frequently used tonal syllables out of 1345 with a desired distribution, etc.

A computer algorithm was therefore developed to select incremental sets of phonetically balanced sentences with different chosen phonetic criteria from a large Chinese text corpus. As an example, a total of three phonetically balanced sentence sets were chosen to form a three-stage adaptation procedure as listed in Table 6 and shown in Fig. 9. The corpus used here from which the sentence sets were selected consisted of a total of 124,845 sentences (1,374,182 characters) collected from daily newspapers.

In the first stage, a phonetically balanced sentence set was obtained that covered all the necessary INITIAL/FINAL sub-syllable units with a desired distribution. This set consisted of 24 sentences or 183 characters (or syllables). The total length of the speech signal produced for these 24 sentences in continuous speech mode was roughly 50 sec only. In the second stage, 76 additional sentences or 622 additional characters (or syllables) were added together with the sentences in the first stage to form a phonetically balanced sentence set covering all the 408 base syllables with a desired distribution. The total length of speech signal produced for these 100 (24+76) sentences was roughly 3.1 min.



10. The improvements in accuracy in the three-stage incremental speaker-adaptation procedure.

The third sentence set can be similarly selected as shown in the last row of Table 6. In this way, the speech data produced for these phonetically balanced sentence sets could be used as very good adaptation data for a new speaker. Furthermore, since these sentence sets also reproduced (to a very good approximation) some desired statistical distribution of the selected phonetic units, the more frequently used units could be better trained and recognized more accurately. With these sets of phonetically balanced sentences produced by the new speakers, the system could adapt to a new speaker stage by stage.

Figure 10 is an example experimental result for the above three-stage incremental adaptation procedure averaged for a number of outside speakers (i.e., speakers who didn't produce any voice to train the initial speaker independent system) based on the continuous-speech input mode using the context-dependent INITAL/FINAL units mentioned previously. Only the lower curve in Fig. 10 for tonal-syllable accuracy is discussed here, while the upper curve for character accuracy will be discussed later on. The average tonal-syllable accuracy for the initial speaker-independent models trained by the data produced by many other speakers was only 51.4%, as can be seen in Fig. 10. This number is signifi-

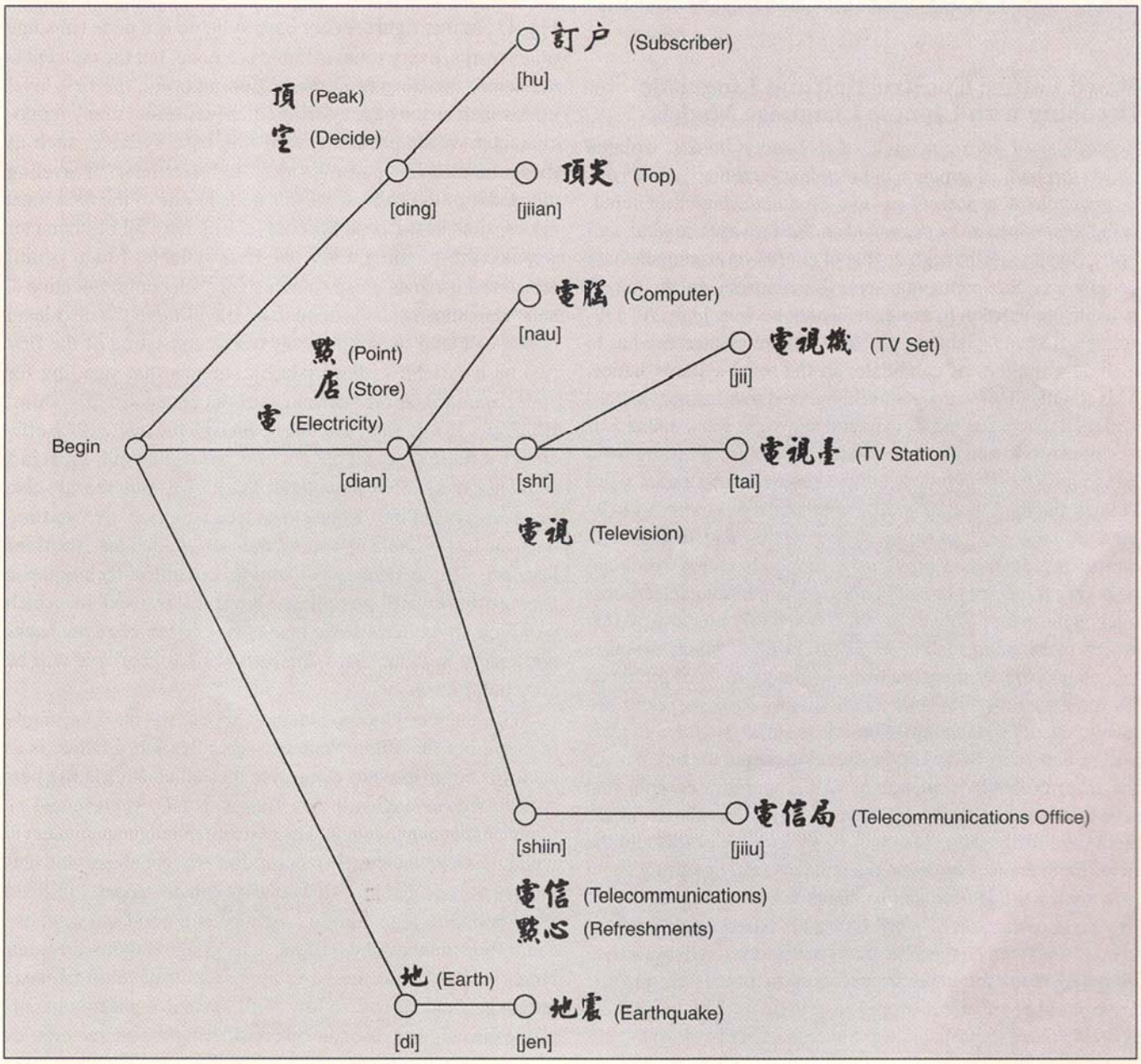
Table 6. The example phonetically balanced training sentence sets for incremental speaker adaptation.

Training Sentence Set	Accumulated Number of Sentences (Characters or Syllables)	Accumulated Speech Length in Continuous Speech Mode	Phonetic Criteria Used to Select the Set
1	24 (183)	50 sec.	Covering all INITIAL/FINAL units with a desired distribution
2	100 (805)	3.1 min.	Covering all 408 base syllables with a desired distribution
3	200 (1606)	6.8 min.	Covering top 600 most frequently used tonal syllables with a desired distribution

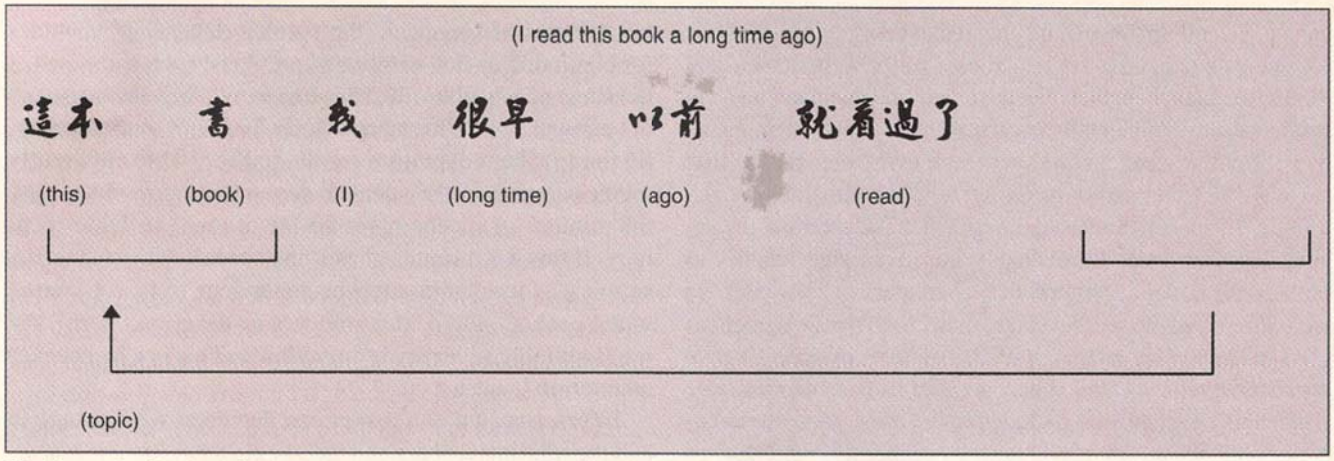
cantly lower than those that appeared in the previous sections for speaker-dependent cases, such as 89.8% for tones and 88.3% for base syllables, because here the speakers had not produced any voice of their own to train the system. However, after the speakers produced their own voice for the first stage of 183 characters or 24 sentences (with roughly 0.8 minutes of speech) and used the data in the adaptation, the average accuracy was immediately improved significantly to 66.9%. When an additional 622 characters or 76 sentences were uttered in the second stage, these 24+76=100 sentences (183+622=805 characters, with 3.1 minutes of speech) gave an accuracy of 76.5%. When another 100 sentences (801 characters) were further included in the third stage, the accuracy could be improved to 80.1%. Apparently, with this incremental adaptation procedure the recognition rates can be improved very fast stage by stage.

These three stages of phonetically balanced sentences consist of 1606 characters or 200 sentences with only about 6.8 minutes of speech. In the speaker-dependent examples for continuos speech mentioned previously, 16.4 minutes of training speech gives 88.3% of base-syllable accuracy and 6.4 minutes of training speech gives 89.8% of tone accuracy. So the speaker adaptation results in Fig. 10 are apparently much better than the speaker-dependent case, since only 6.8 minutes of speech gives 80.1% of tonal-syllable accuracy. If this 6.8 minutes of training speech with 1606 characters was used in a speaker-dependent test (not started with speaker-independent models as described here), the tonal-syllable accuracy is only 72.4%. This is why speaker adaptation is attractive.

In practice, a new speaker can decide at which stage to end the training process and then begin to use the system directly. After he begins to use the system, further adaptation can be performed on-line during real applications as well, as long as corrections can be made and the user can tolerate the errors. A nice feature of the learning curve in Fig. 10 is that



11. A typical partial listing for a tree structure of the lexicon.



12. The topicalization of a sentence as a typical example of long-distance movement.

the slopes in the first two stages are really high; i.e., the improvements in accuracy are very fast in the first two stages after the user spends very limited time producing the necessary adaptation data.

Word Lattice Construction and Linguistic Decoding with Chinese Language Models

Regardless of an input mode of isolated syllables, isolated words, prosodic segments, or complete sentences, the acoustic recognition processes always produce a lattice of tonal-syllable candidates as presented in the previous several sections. Because of the high degree of confusion among the base syllables in the confusing sets, the accuracy in the tonal-syllable recognition in any case cannot be very high. As a result, each tonal syllable in the input unknown utterance has to include a number of candidates in the tonal-syllable lattice. This tonal syllable lattice should be used to construct a word lattice as discussed previously and shown in Figs. 3 and 5. If the input mode is in isolated words, the very complicated word lattice may be broken down into a sequence of smaller word lattices, but the situation is still complicated due to the high degree of ambiguity existing in the mono- and bi-character words. As mentioned previously, mono-character words appear very frequently in daily language, and bi-character words occupy more than 70% of the most frequently used top 50,000 words in Mandarin Chinese. From Table 2 which was discussed previously, there are serious homonym word problems for these two categories of words. In any case, powerful linguistic decoding techniques based on some language models will be necessary to find as the dictation output the best path or the most probable sequence of words within a complicated word lattice or a sequence of word lattices, as shown in the right part of Fig. 5(a). This will be the subject of this and the next few sections. However, we will first say something about how such word lattices can be constructed.

The construction of word lattices is based on a matching process between all possible paths in the tonal-syllable lattice obtained from the acoustic recognition processes and the large number of words stored in a lexicon. This allows all possible word hypotheses to be included in the word lattice. In order to make such a matching process efficient, especially for the very large number of words for a dictation task, the words in the lexicon are usually stored in a tree structure. A good example of a partial list of a tree structure is shown in Fig. 11. In this figure, every base syllable is a node (in some other works, every tonal syllable is a node, but the concept is the same). Starting from the beginning point, the first-level nodes such as the base syllable [dian] represent many mono-character words produced with this base syllable, such as those standing for "point," "store," or "electricity." Traveling along one of the paths from this node to one of the next level nodes, such as the base syllable [shiin], here all bi-character words corresponding to the two base syllables [dian] [shiin] are stored, such as those standing for "telecommunications" and "refreshments." Continuously traveling to the next level nodes will lead to tri-character words consisting of the first two base syllables [dian] [shiin], such as that standing for "telecommunications office," produced as [dian] [shiin] [jiiu], etc. In this way, searching through the tree will be efficient for finding out all the possible words that may exist in a sequence of syllable candidates. Such a lexicon tree can also be re-organized into a backward tree structure, i.e., starting with the last syllable in a word, then the second last, the third last, etc. This is because in some recognition techniques a forward-backward searching algorithm is used in which word matching in the lexicon may be performed in the backward path. In those cases, the backward lexicon tree will be very helpful.

With the word lattice obtained above, the basic principle in obtaining the output sentence from this word lattice is to perform linguistic decoding over the lattice to find the best path based on linguistic constraints usually represented as Chinese language models. The concept of language models is well known in the speech-recognition area for alphabetic languages as well, but how this concept can be properly utilized and applied to the Chinese language is a good question because the structure of the Chinese language is quite different. Here, we start with the basic approach to develop Chinese language models and follow with special measures to improve the language models by considering the characteristics of the Chinese language.

In order to find the most probable output Chinese sentence, $\overline{X}$, from a word lattice, L, a natural approach is to search through the entire lattice from the beginning to the end and find a single path with the maximum likelihood. The output Chinese sentence can then be obtained by concatenating all the words on this path. Let Z be the set of all possible paths in L, X be an arbitrary element of Z, and $\overline{X}$ be the maximum-likelihood path to be found. The desired path, $\overline{X}$, can then be defined as

$$\overline{X} = \arg\max_{X \in Z} P(X|S), \tag{2}$$

which specifies the maximum-likelihood condition, where S is the input speech signal that is probably a sequence of isolated syllables, words, prosodic segments, or a complete sentence, depending on the input mode. Using Bayes' theorem, the above equation can be rewritten as

$$\overline{X} = \arg\max_{X \in Z} \{P(S|X) * P(X)\} \tag{3}$$

because $P(S)$ is identical for all paths, X, and is therefore deleted. In this equation, there are two probabilities: $P(S|X)$ can be computed from the acoustic scores obtained for the tonal-syllable candidates during the acoustic recognition processes discussed previously, while the probability $P(X)$ is to be estimated by some language models [69-74].

Here a relatively simple Markov Chinese language model is first defined in the following as an initial example. The probability $P(X)$ for a given path, $X = \{X_1, X_2, X_3 \ldots, X_R\}$, where X_r is the rth word on the path, can first be decomposed into

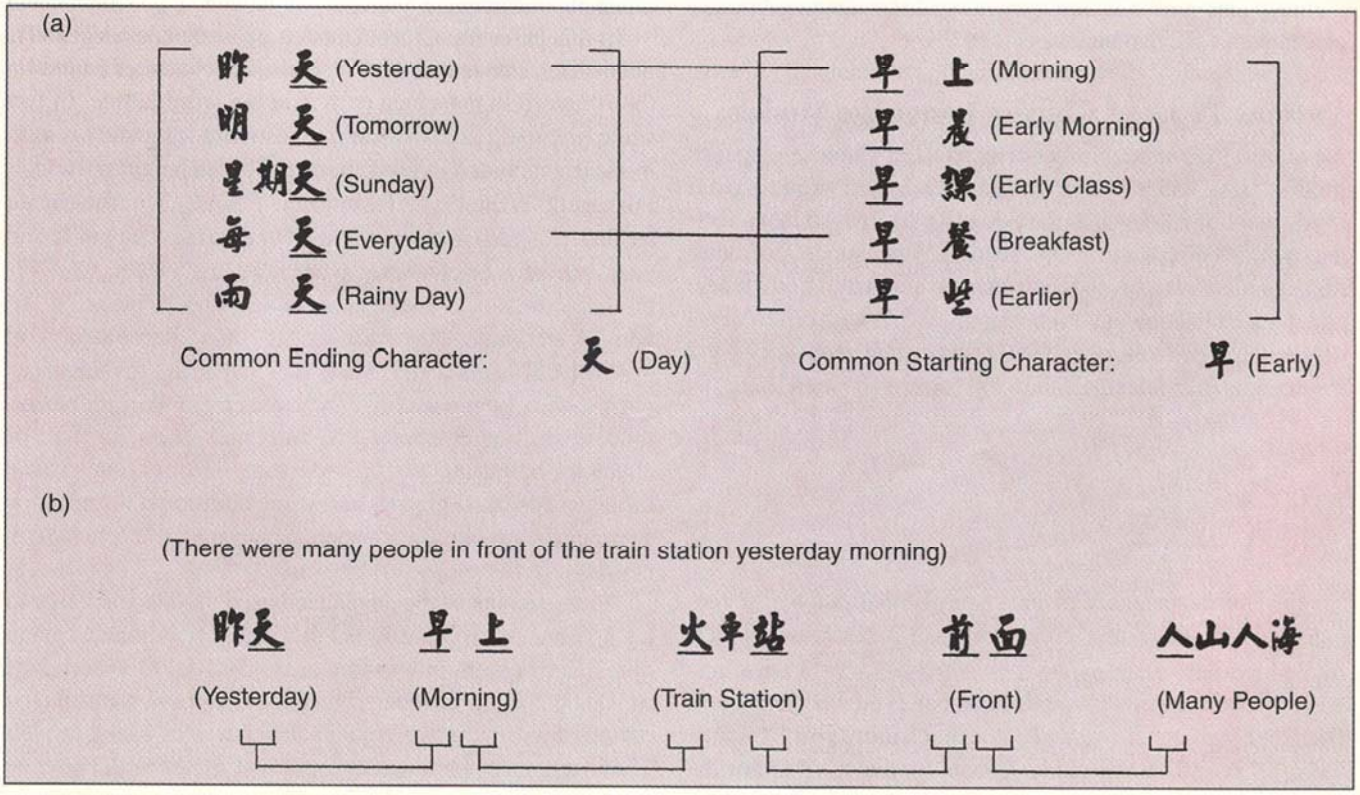
$$P(X) = P(X_1) \cdot \left[\prod_{2 \le r \le R} P(X_r | X_1, X_2 \ldots, X_{r-1}) \right]. \tag{4}$$

If we assume that every word, X_r, of X satisfies the Markovian property [75], then the above equation can be simplified as

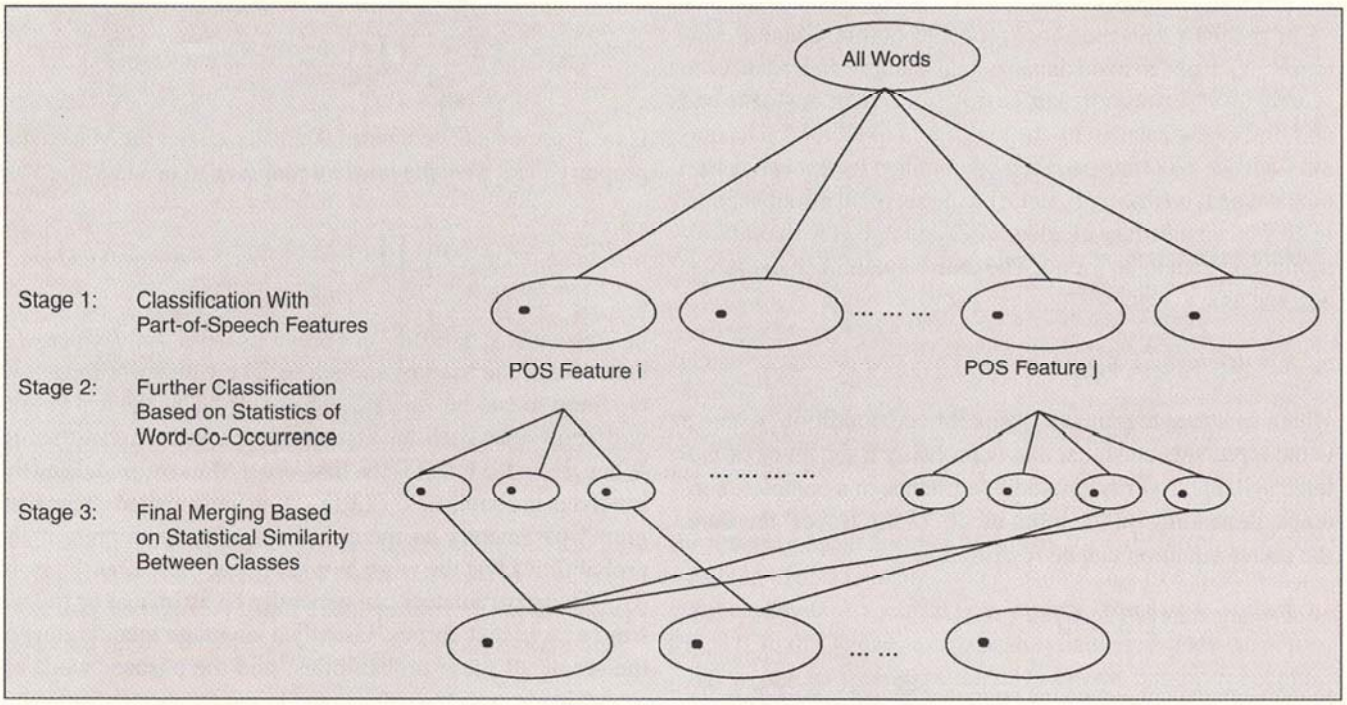
$$P(X) = P(X_1) \cdot \left[\prod_{2 \le r \le R} P(X_r | X_{r-d}, X_{r-d+1} \ldots, X_{r-1}) \right]; \tag{5}$$

i.e., the word X_r depends on d previous words only, where d is the order of the Markov modeling. The validity of the above assumption can be verified empirically; i.e., the test results will show that such an assumption is, in fact, reasonable. When d is set to 1, this is the first-order Markov model and the involved probabilities, $P(X_r|X_{r-1})$, usually called "word bi-gram" parameters as mentioned previously, represent the probabilities that the word X_r will appear right after the word X_{r-1}. These parameters can generally be estimated or trained from a large text corpus. Usually a language model refers to the set of all these probabilities, and the phrase "word bi-gram" is also used to refer to the language model consisting of word bi-gram probabilities, etc.

A stronger language model may be that for d=2, or the so-called word tri-gram with probabilities $P(X_r|X_{r-1}, X_{r-2})$, or even some other higher-order models (the so-called word N-grams with d=N-1, etc.). While the capabilities of these language models are yet to be verified by experiments, their performance really also depends on whether a large enough training text corpus is available to obtain good estimates of such large numbers of parameters. For example, if a lexicon



13. (a) Chinese Word classes with common ending/starting characters. (b) The probability evaluation for a sentence using language models based on such word classes.



14. The three-stage word classification algorithm to obtain Chinese word classes considering both grammatical and statistical knowledge.

of 50,000 words is to be used, the word bi-gram requires $(5\text{x}10^5)^2$ probabilities, and the word tri-gram requires $(5\text{x}10^5)^3$ probabilities. Not only will training such a large number of parameters be very challenging even if a huge text corpus is available (for example, it is important whether the corpus is well balanced on the desired domains and subjects), but storage and retrieval of these parameters will also be difficult. In order to implement a dictation system with reasonable memory size and cost, compactness of the language model parameters will also be a key issue.

Various Types of Chinese Language Models

An almost unlimited number of variants of Chinese language models exist, with the word bi-gram (d=1) and word tri-gram (d=2) mentioned above as the two basic forms. All these variants will be discussed in this section. First, models of other higher orders also provide very useful linguistic knowledge, and the combination of parameters of models with different orders with proper weighting factors very often gives very good results. For example, the following formula is often used:

$$P(X_r|X_{r-1},X_{r-2}) = q_3 \cdot P'(X_r|X_{r-1},X_{r-2}) + q_2 \cdot P(X_r|X_{r-1}) + q_1 \cdot P(X_r); \quad (6)$$

in other words, the word tri-gram may be difficult to train because a very huge text corpus will be needed, but under-trained word tri-gram parameters, $P'(X_r|X_{r-1}, X_{r-2})$, appropriately interpolated with word bi-gram or even word uni-gram (i.e., d=0) parameters may become much more powerful than word bi- or uni-grams alone. Secondly, the word is not the only basis for Chinese language modeling. Another possible candidate for Chinese language modeling is certainly the character [76]. As mentioned before, there are no natural boundaries between words in Chinese sentences; therefore, a sentence can be viewed as a sequence of words or as a sequence of characters. Also, almost every character is a morpheme with its own meaning and linguistic features; apparently, a character appearing after another set of characters also represents important linguistic constraints. This leads to the concept of Chinese language modeling based on characters rather than words.

In this case, Eq. (5) remains completely unchanged. The only difference is that now X_r is the rth character instead of the rth word in the given path X in the word lattice. In fact, some linguistic information such as word frequency is automatically included in such character-based language models. For example, the character bi-gram $P(X_r|X_{r-1})$ for the character pair (X_{r-1}, X_r) is closely related to the frequency of the appearance of a bi-character word, "$X_{r-1}X_r$," if "$X_{r-1}X_r$" is a bi-character word. On the other hand, if both characters X_{r-1} and X_r are mono-character words, then the character bi-gram $P(X_r|X_{r-1})$ is actually the word bi-gram for these two words. Similar reasoning can be extended to higher-order models such as character tri-grams and so on. In fact, the character-based models provide some information existing in the word-based models and some additional information, so proper interpolation between the two is also helpful when a good interpolation algorithm is used.

A nice feature of the character-based models is that the total number of commonly used characters is much smaller than that of commonly used words, say 10,000 as compared to 100,000, so the number of bi-gram, tri-gram, or similar parameters will be much smaller for character-based models. Therefore, these parameters can be estimated with better accuracy as compared to those for word-based models if the training corpus is not unlimited in reality. The smaller

number of parameters also makes it possible to obtain model parameters with higher order, such as the so-called N-grams, and makes storage, retrieval, and implementation of such language models easier.

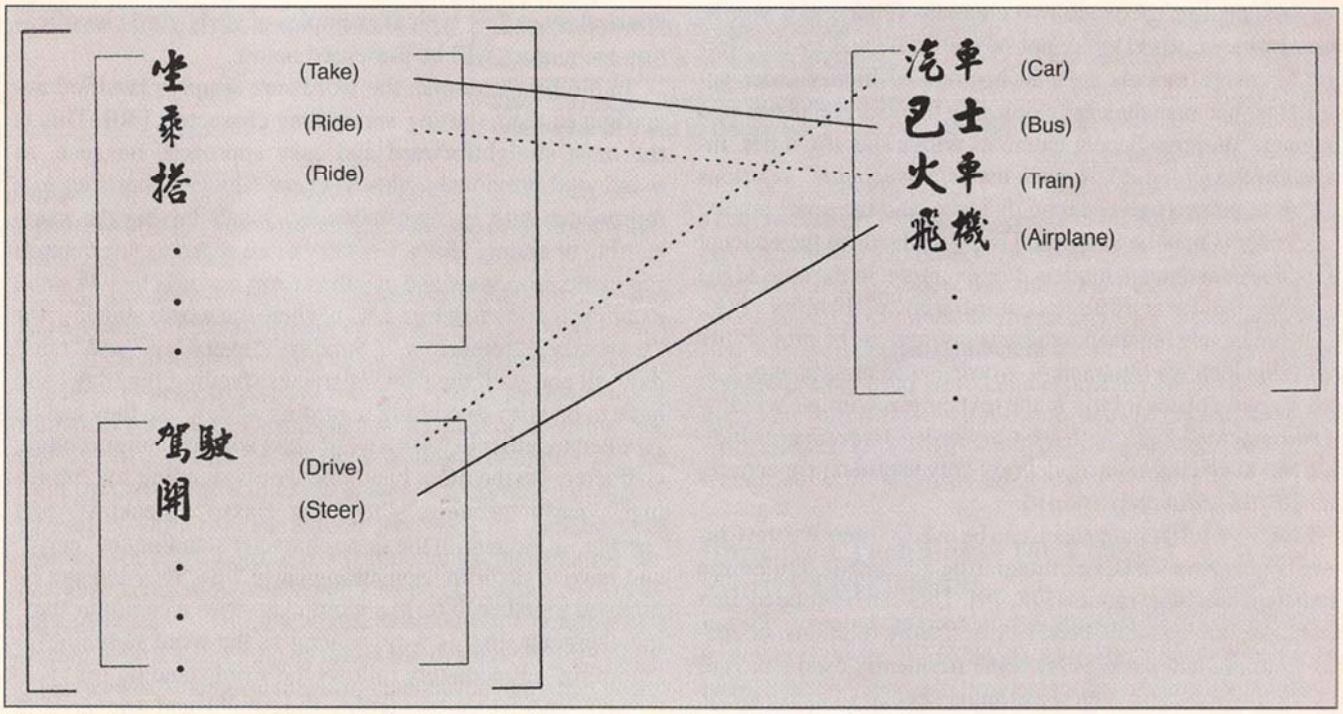
On the other hand, as mentioned previously, the words in a Chinese sentence are not well defined, and the segmentation of a sentence into words is not unique; therefore, training word-based language models requires at least a large enough training corpus that is consistently segmented into words. This is certainly not easy to obtain. Therefore, another good feature of the character-based language models is that the characters in a sentence are straightforward, so the problem of consistently segmenting the training corpus into words can be directly bypassed. However, experiments indicate that character-based models apparently possess very good capabilities in linguistic decoding, but that word-based models are certainly better in every case, if both of them are used alone.

For example, a word bi-gram is more effective than a character bi-gram, and a word tri-gram is more effective than a character tri-gram. On the other hand, successful interpolation of the two is certainly better. Also, very interesting experiences were obtained in comparison between the language models for Chinese and English languages. For English the word N-grams are useful, but the character N-grams are probably not since the characters do not generally bear any meaning. The experiences are such that, though this is difficult to describe in quantitative measures, the word bi-gram is apparently more effective for linguistic decoding in Chinese than in English, and so is the word tri-gram. The possible reason is that in Chinese every word is composed of one to several (say most frequently two) characters that also have their own meaning. So, for example, the word bi-gram probability for a Chinese word appearing after another is very close to the probability for four characters (if each of these two words has two characters) appearing in a sequence, which is similar to a character 4-gram. Since almost each character has its own meaning, this word bi-gram probability, if also considered as something like a character 4-gram probability, certainly provides stronger linguistic constraints. Such interesting phenomena probably do not exist in word N-grams for alphabetic languages such as English.

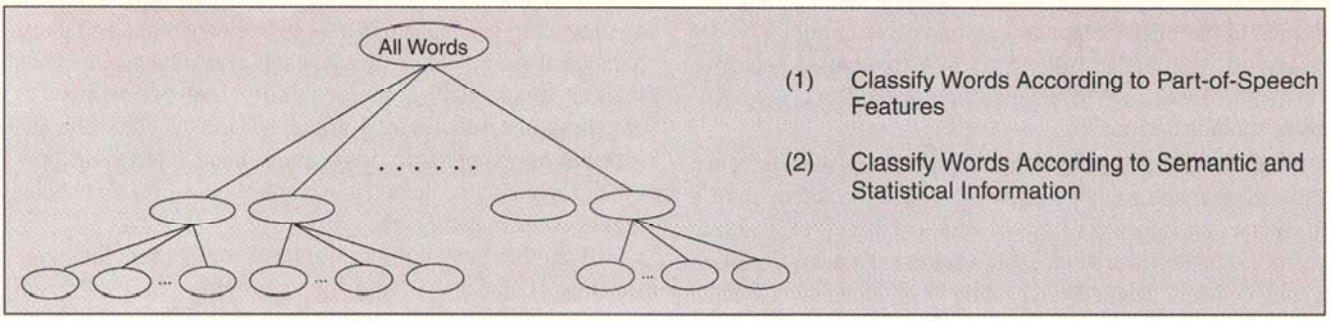
Still another even more useful basis for Chinese language modeling is the word class, i.e., grouping many different words with similar linguistic properties together as a class, so that a very large number of words can be categorized into a much smaller number of word classes. The Chinese language model can then be constructed based on these word classes [77]. In such a situation, for example, the bi-gram probability $P(X_r|X_{r-1})$ can be replaced by

$$P\left(X_r|X_{r-1}\right) = P\left(C\left(X_r\right)|C\left(X_{r-1}\right)\right)P\left(X_r|C\left(X_r\right)\right), \quad (7)$$

where $C(X_r)$ and $C(X_{r-1})$ are, respectively, the word classes containing the words X_r and X_{r-1}. In this way, not only can the number of word classes be much smaller, but this number can even be adjusted by the designer based on various considerations such as the desired accuracy, acceptable memory size, and the estimation ability of the parameters. Also, because of the much smaller number of word classes, training of higher-order language models becomes possible. Another nice feature of such word-class-based language models is the automatic smoothing effect for many words in the same word class. For example, both words X_r^1 and X_r^2 belong to the same word class, C_r, and the words X_{r-1}^1 and X_{r-1}^2 both belong to the same word class, C_{r-1}. As long as X_r^1 appears frequently right after X_{r-1}^1 in the training corpus, the bi-gram parameter $P(C_r|C_{r-1})$ will be adequately trained. So even if the word X_r^2 doesn't appear right after X_{r-1}^2 frequently in the



15. Typical example word classes obtained from the three-stage word-classification algorithm.



16. The algorithm for clustering words into overlapping classes integrating semantic information.

training text corpus, the bi-gram parameter $P(C_r|C_{r-1})$ will indicate an appropriate probability for X_r^2 appearing right after X_{r-1}^2. This solves the difficult problem of requiring that all the word pairs (X_r, X_{r-1}) be adequately trained and assigned an appropriate probability value, which is usually difficult even if a huge text corpus is used for training. Finally, just as before, different orders of word-class-based language models can be combined together with proper weighting factors, and they can also be interpolated with word-based or character-based models because each of them bears somehow different linguistic information. The key problem here is how to group the words into effective word classes, which will be discussed later on.

All the above types of Chinese language models can be further extended. A special feature of the Chinese language is that word order is quite free and the long-distance relation or long-distance movement appears very frequently. A good example is shown in Fig. 12, in which the sentence that means "I read this book a long time ago" is used as an example. In this sentence, the object element standing for "this book" does not follow the verb element standing for "read," but instead is moved to the beginning of the sentence as the topic of the sentence (the so-called topicalization effect). This implies that the "association" between two elements (characters, words, word class, or others) may not be well reflected by the Markov language models such as bi-gram or higher-order parameters, because they only describe the local behavior of a language among adjacent elements with a specific order. Instead, in the Chinese language many "association" relations can be separated by very long distance, and the order among the elements may be very free [78]. This leads to the concept of extended language models. For example, in the case of the bi-gram parameter $P(X_r|X_{r-1})$, a different probability, $P(X_r, X_s)$, may be used instead, which is defined as the probability that two elements (characters, words, word classes, etc.), X_r and X_s, will appear jointly in the text corpus within a window of some given length with arbitrary order. Experiments indicate that such language models are very useful if properly designed and adequately trained.

Finally, a further approach can be used to improve the Chinese language models, i.e., integrating grammatical rules into statistical language models [78, 79]. This can be done by first analyzing the syntactic behavior, semantic relations, or special grammatical patterns for some frequently used words or function words, and then developing special rules for them. These rules can then be properly integrated with the statistical language models such as bi- or tri-grams. Because the grammatical rules very often provide extra information orthogonal to the statistics obtained from the text corpus, experimental results indicate that such a hybrid approach is very attractive since significant improvements in accuracy can usually be obtained at almost no extra cost in computation or memory size.

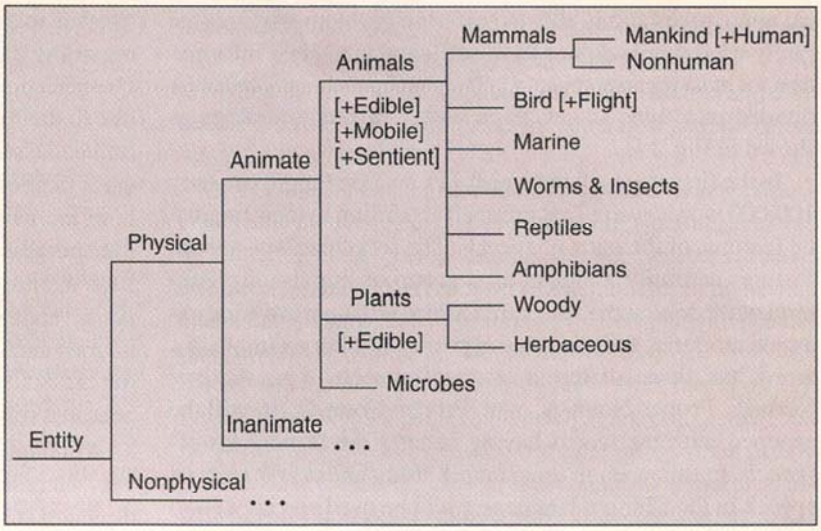
Typical Chinese Word-Classification Techniques

Although there exist so many different types of Chinese language models as mentioned above, it has been found in almost all the experiments that word-class-based language models are always very useful. They have actually been practically used in various prototype systems very successfully. The key issue, however, is how to group the words into appropriate word classes on which the language models can be constructed. Word-classification techniques have been discussed extensively for western languages [77] in the natural language analysis area, but here we will discuss some experiences in the classification of Chinese words in terms of dictation applications for Mandarin speech with very large vocabulary. A few typical examples of such word classification techniques will be presented below.

In the first example, the words are simply classified according to their starting and ending characters [80]. This is the most straightforward and easy approach. Because, as mentioned previously, almost every Chinese character is a morpheme with its own meaning, words having the same starting or ending characters very often share some common linguistic properties and can thus form a word class. A good example is shown in Fig. 13(a), where the words standing for "yesterday," "tomorrow," "Sunday," "everyday," and "rainy day" all end with the same character standing for "day" and have to do with something regarding a "day," so they can be grouped together to form a word class with a common ending character. On the other hand, the words standing for "morning," "early morning," "morning class," "breakfast" and "earlier" all start with the same character standing for "early" and have to do with "something early." So, they too can be grouped together to form a word class with a common starting character. In this way, as long as the word standing for "morning" immediately follows the word standing for "yesterday" (indicating "yesterday morning") and appears fre-

quently enough in the training text corpus, other combinations such as the word standing for "breakfast" immediately following the word standing for "everyday" (indicating "every breakfast") will be automatically trained even if they do not appear frequently enough in the training text corpus.

The way in which the probability of a sentence, $P(X)$, is evaluated is shown in Fig. 13(b), in which the sentence "There were many people in front of the train station yesterday morning" is composed of five words. If the word-class-based bi-gram is used, the bi-gram for every two concatenated words will be evaluated—for example, the word standing for "yesterday" followed by the word standing for "morning" and the word standing for "morning" followed by the word standing for "train station," etc. Note that here, the second word standing for "morning" is considered to be a word in a word class with a common starting character in the former case, but as a word in a word class with a common ending character in the latter case. In this way, every word generally belongs to two word classes: one with a common starting character and the other with a common ending character. So generally, the total number of word classes will be the total number of characters times two because every character can be a common starting character and a common ending character of a word class.



17. A simplified partial list of the conceptual structure of Chinese words.

The nice feature of this technique is that the categorization is very simple, so any new word added to the lexicon can be automatically categorized into its corresponding class without any problem. The weak point of this approach, however, is also clear. Of course, not all words having the same starting or ending characters always have identical linguistic properties. Very often some words will be inappropriately assigned to some classes irrelevant to their linguistic features. Experiments indicate that the performance of language models based on these word classes is quite satisfactory though, because such inappropriately assigned words only constitute a very small portion for the statistical models based on a large number of words and a huge text corpus. However, better word classification techniques are highly desired, and the following is one example.

This example approach is based completely on statistics from a large corpus. Because vector quantization has been a very efficient technique for clustering a large number of vectors into classes, efforts have been made to try to represent the statistical behavior of words in a large text corpus using feature vectors so that vector quantization techniques can be used [81]. Assume a total of N_T words is considered; one or several N_T-dimensional feature vectors can then be constructed for each word. Each component represents the appearance frequency of a certain word out of the N_T words, appearing jointly with the given word within a window of a certain length, preceding or following or on both sides of the given words. Vector quantization techniques can then be applied to cluster these vectors into the desired number of classes, as long as a good distance measure between two vectors can be properly defined.

Although this concept sounds reasonable, for a very large number of Chinese words (for example, on the order of 100,000) the very large number of super-long vectors is not easy to quantize. Quite a few special techniques have therefore been developed and applied to simplify the above problem in order to make word classification quantization practically feasible. In a preliminary study, a lexicon of roughly 100,000 words were used, and about 1,000 word classes were eventually obtained. By observing the 1,000 word classes carefully, it was found that most words that were clustered into the same word class had relatively similar linguistic behavior and that the language models constructed with these word classes actually performed significantly better than those with the word classes discussed previously based on the beginning and ending characters of the words [81].

It was also very interesting to observe some linguistic behavior of Chinese words from the results; for example, adverbs were classified better if the classification was based on the following words to their right only; while quantity words were classified better if based on the preceding words to their left only. This is certainly because words modified by adverbs usually follow the adverbs, but quantity words are usually followed by numbers, etc. However, the primary weak point of this approach is that the classification is completely based on statistics; thus, all frequently used words are classified very well due to a sufficient amount of training data, but less frequently used words and especially rarely used words may very often be clustered into some very inappropriate word classes. This is why some improved word classification techniques will be discussed below.

Improved Techniques for Chinese Word Classification

In order to solve the problems discussed above, an improved three-stage hierarchical word classification algorithm has been developed and will be presented here. This algorithm integrates the advantages of both the grammatical and statisti-

cal approaches and is able to solve the problem where some rarely used words do not have sufficient statistical information for classification [37, 44]. The classification process was divided into three stages, each with a different strategy as shown in Fig. 14.

In the first stage, all the words in a lexicon (approximately 100,000 words) were first grouped according to their linguistic features of the parts of speech. The set of parts-of-speech features carefully assigned by a group of linguists in long-term work done at the Academia Sinica at Taipei is a good example reference for this stage of work [82]. For example, if a word has three different parts-of-speech, e.g., Active-Verb-B, Proper-Noun-A, and Proper-Noun-C, it will be grouped with the words having exactly the same parts-of-speech. In this way, in an example study about 200 parts of speech in the Chinese language could be used and about 950 initial classes could be obtained. Of course, in this way, every word was assigned to a single class, but a class may have several different parts of speech.

In the second stage, the words in the same class, which were believed to have similar syntactic behavior, were further grouped into smaller classes based on their statistical behavior [83]. That is, words having similar word-co-occurrence feature vectors were further grouped into even smaller classes based on the criterion that the similarity measure between them exceeded a given threshold. It is believed that some implicit semantic information is integrated in the second stage. In the third stage, however, in order to avoid classification that is too restrictive, some classes obtained in the second stage with different parts-of-speech features can be further merged together according to the statistical similarity between them, even if they have been separated in the first stage.

In general, with the above procedure, words clustered into the same class have very similar syntactic and semantic behavior because both their parts-of-speech features and co-occurrence relations with adjacent words in the corpus are very similar. Also, the number of finally obtained classes is adjustable for different application tasks, considering factors such as accuracy and memory size. In the experiments, it was found that roughly 950 to 2,000 classes were very good choices for the Chinese lexicon of about 100,000 words used. Furthermore, for those rarely used words with insufficient statistical information, the classifications were still satisfactory because the parts-of-speech features had been carefully used. This is why the language models based on this word-classification algorithm were found in a series of experiments to be significantly more powerful in linguistic decoding with a smaller number of model parameters and much more robust with respect to a smaller training text corpus.

A good example is shown in Fig. 15 in which the words standing for "car," "bus," "train," and "airplane" automatically belong to a word class, i.e., transportation vehicles. Those words standing for "take" and "ride" are categorized into an initial class of verbs describing some kind of "state" with regard to these transportation vehicles, while other words standing for "drive" and "steer" are categorized into another initial class of verbs describing some "operations" regarding the vehicles. In other words, after the first stage of classification, the two initial classes of verbs are separated due to their different semantic features. However, these two initial classes will eventually be merged in the third stage of classification to become a single word class because they both are usually followed by the same class of nouns, i.e., transportation vehicles. In this way, as long as combinations such as those standing for "take the bus" or "steer the airplane" appear in the training text corpus, all the other combinations such as those standing for "ride the train" or "drive the car" will be automatically covered, even if they do not appear in a relatively small training text corpus.

Although word classes obtained as described above are very successful, further improvements are still possible. The direction is to try to include even more semantic information in the word classification. In natural language processing, semantic information is so crucial that it has been intensively utilized for years. However, not too many studies in speech recognition that sufficiently integrate semantic information have been reported. For the Chinese language, word order is very free in sentences, the syntactic constraints are relatively loose, and the meaning of a sentence is usually primarily determined by the semantic features of the component words. It is thus believed that better integration of semantic information should be able to further improve the accuracy in speech recognition. The only question is how to do this properly. In the three-stage word classification technique discussed above, every word is assigned to a single word class, and only the syntactic knowledge and statistical information are primarily considered when clustering the words.

In a further-improved approach to be presented here, however, not only does the clustering of words consider both syntactic and semantic information, but each word is allowed to be present in more than one class. Such a concept with overlapping word classes is crucial when considering both syntactic and semantic information for Chinese language modeling. Almost every Chinese word can have multiple linguistic features, and its corresponding meaning is strongly influenced by the context. For example, the Chinese word in (e) of the character box is frequently used both as a verb meaning "construct" and as a noun meaning "structure." It should thus be assigned into at least two different classes. The basic algorithm for clustering words into such overlapping classes is conceptually depicted in Fig. 16.

First, each word in the vocabulary is classified based on all of its possible parts-of-speech features according to human knowledge [82]. Here, each word now can be assigned to more than one category, as long as its linguistic features so indicate. This step results in a set of overlapping grammatical categories. In this way, words in the same category should have very similar grammatical behavior. Secondly, the words in each of the categories are further partitioned into smaller classes according to the similarities between each two of them. This similarity, $S(w_i, w_j)$, between words w_i and w_j considers both seman-

tic information from human knowledge and statistics from a large text corpus simultaneously:

$$S(w_i, w_j) = a \times S_s(w_i, w_j) + (1-\alpha) \times S_c(w_i, w_j), \quad (8)$$

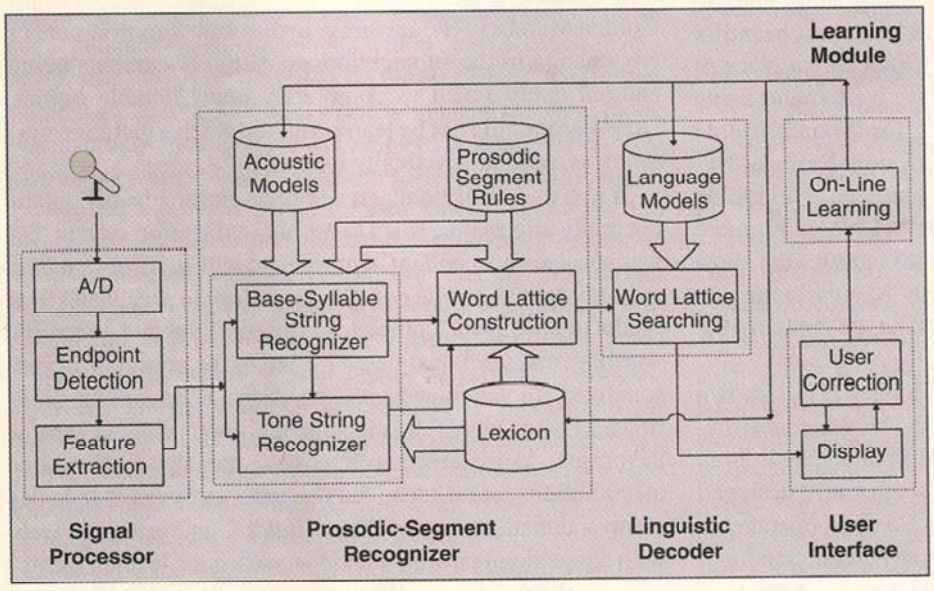
where $S_c(w_i, w_j)$ is the similarity evaluated from the context vectors acquired from a large text corpus [83] while $S_s(w_i, w_j)$ is determined by the conceptual structure [84], which is a hierarchical knowledge representation for the semantics of Chinese words. Figure 17 is a simplified partial list of the conceptual structure for Chinese words. With this conceptual structure, not only can the semantics of a word be extracted from its location within the hierarchy, but the relative distance between two given words can be measured. In this approach, $S_s(w_i, w_j)$ may depend on the relative distance and the least common ancestor for w_i and w_j in the hierarchy, and all necessary parameters can be automatically trained. It was found in the experiments that Chinese language models developed based on word classes obtained in this way had even better performance.

Further Issues in Chinese Linguistic Processing

As mentioned previously, the open vocabulary and the almost unlimited number of words in the Chinese language are two of the major problems in developing Mandarin speech-recognition systems with very large vocabulary. One approach is, of course, to collect as many words as possible from different dictionaries, text corpora, etc., to construct a lexicon that is as complete as possible. In this case, due to the large number of words in the lexicon, in order to have a language model with a small enough number of parameters capable of selecting correct words for utterances, the word-class-based language models constitute one of the major solutions. Since there are almost an unlimited number of ways of grouping word classes, various concepts of word-classification techniques have been presented in the above. Of course, there also exist other important approaches to this problem, one of which is to try to store only all the "basic words" in the lexicon, and to develop a set of word-generating rules such that most of the word variants and compound words can be automatically generated and, therefore, need not be stored.

For example, a verb followed by a given character simply represents the past tense; such as the verb standing for "eat" becomes "ate," the verb standing for "see" becomes "saw," etc. (see (f1) in box). Also, a very large number of compound nouns can be formed by combining two nouns with specific categories. For example, the nouns standing for "pig" and "meat" form a new noun standing for "pork" based on very simple general rules (see (f2) in box). A lexicon of such "basic words" and a set of such word-generating rules usually have to be obtained via careful investigation by linguists and will be very helpful in significantly reducing the necessary number of words while maintaining the same power of the language models. Of course, in this case, the word-classification techniques and resulting language models also have to be modified.

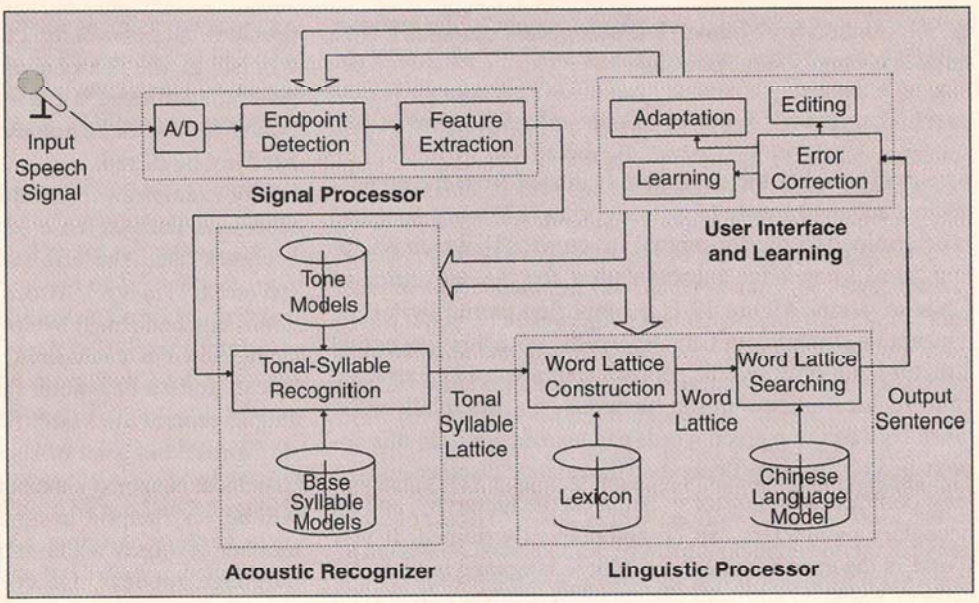
The next important issue is the learning of the new words. It is simply impossible to include all the words in a lexicon; thus, an important problem is how new words can be easily included in a lexicon and integrated into the language model. This process is usually performed when errors occur and the user makes corrections. The way in which the new words are integrated into the language model depends on the design of the language model. If the language model is primarily based on characters or words, integration will be much easier. However, if the language model is based on word classes, how each new word can be assigned to an appropriate word class becomes critical and, in turn, depends on the principles on which the word classification is based. For example, if word classification is based on beginning or ending characters as shown in Fig. 13, the assignment of a new word to the correct word class is straightforward; however, this is much more difficult if the word classification is based on statistics, syntax or semantics.



18. The system block diagram of Golden Mandarin (IIIa).

Another challenging problem is making possible the learning of new words on-line in real time, because when a new word is entered during the dictation of a document on a certain subject, the same new word may be repeatedly used in the same document. In such cases, the learning of new words will be useless if it cannot be performed on-line in real time. This leads to another important concept, that of having two types of learning, i.e., short-term and long-term. Short-term learning is for those new words that will be frequently used while the same document on a certain subject is being entered, but should not be included in the main lexicon or integrated into the main

language model. This avoids increasing the size of the lexicon by too much and disturbing the functions of the language model, since these words may be very rarely used after the current document is finished. Long-term learning, on the other hand, is for the new words that should be included in the main lexicon and integrated into the main language model, since these new words represent the user's special wording and speaking/writing style. These concepts of learning new words lead to a very challenging related problem, i.e., adaptation of the language model to different users as well as different application domains and subject areas. This would allow very high accuracy to always be maintained for each user, even when the domains and subjects of documents being entered are switched from one to the other [85]. This problem is difficult, and substantial efforts are currently being made even for alphabetic languages. Equal efforts should certainly be focused on the Chinese language.



19. The system block diagram of Golden Mandarin (IIIb).

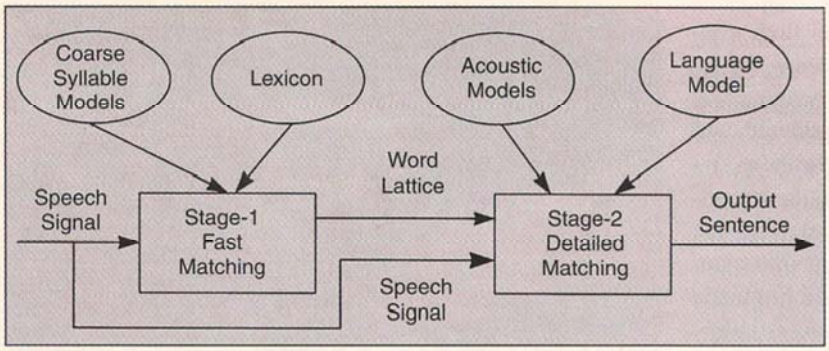
Typical Prototype Systems

With the basic problems and core technology for voice dictation of Mandarin Chinese summarized as above, several typical prototype systems will now be presented to demonstrate the feasibility of the concepts and the stages of developments. Although several prototype systems have been developed by different research groups, the Golden Mandarin Series of prototype systems developed at National Taiwan University and Academia Sinica at Taipei are very good typical examples and are the ones the author is the most familiar with. In classical Chinese literature, it is said that the most beautiful sound in the world is that produced by knocking a piece of gold with a piece of jade. Such a sound is given a name using four-character Chinese words standing for "sound of gold and jade" (see (g) in box). The research group at National Taiwan University and Academia Sinica has been working on the problem of voice dictation of Mandarin Chinese for more than 12 years. For the researchers in this group, the most beautiful sound in the world is certainly Mandarin speech. This is why they have called their prototype systems the Golden Mandarin Series.

Golden Mandarin (I), completed in March 1991, is known to be the first successfully developed real-time Mandarin dictation system in the world [40]. It was implemented on an IBM PC/AT, connected to three sets of specially designed hardware boards on which 10 TMS 320C25 chips operated in parallel. It is speaker dependent without any adaptation functions. The input mode is in isolated syllables. The acoustic recognition is primarily based on a set of specially trained, delicate CHMMs for the 408 base syllables, as shown in Fig. 7. The language model is only a very straightforward character-based bi-gram without any learning functions.

Golden Mandarin (II) was completed in September 1993 [41, 42]. It was implemented on a digital signal processor (DSP) card with a single-chip Motorola DSP 96002D and can be installed on any personal computer. The input mode is still in isolated syllables. The acoustic recognition is primarily based on the SPMs shown in Fig. 8, while the language model is based on word classes. The major achievements of Golden Mandarin (II) as compared to Golden Mandarin (I) were two-fold: the reduced hardware requirements from 10 DSP chips to one DSP chip due to the computation efficiency of the SPM, and the various adaptation/learning functions. Golden Mandarin (I) was speaker dependent and the user had to spend a very lengthy period of time training the system before using it. Moreover, for Golden Mandarin (I), any noise in the user's environment or any change in the subject domain of the documents being entered could result in completely unpredictable output. New words could not be learned either. Such a dictation system was, in fact, practically useless.

It was then realized that it would be extremely difficult, if not really impossible, to try to develop a dictation system that was speaker independent, that worked with all different user requirements and noise conditions, and could accept all kinds of documents with completely different subject domains, wording, and writing styles. This led to the concept of "personalized" dictation systems with adaptation/learning functions. In other words, it is technically impossible, at least at the present, to implement a dictation system that works under all conditions for all users, but it is practically feasible to develop a dictation system that is flexible in various aspects with adaptation/learning functions, so that it can learn the user's conditions, adapt to the user's requirements, and eventu-



20. An improved architecture for dictation of continuous Mandarin speech properly integrating acoustic and linguistic knowledge.

ally become "personalized" for each user. This concept was partially realized in Golden Mandarin (II) [42].

Golden Mandarin (II) was speaker adaptive. A learning curve for speaker adaptation very similar to that shown in Fig. 10 based on incremental sets of phonetically balanced sentences very similar to those shown in Fig. 9 was implemented on Golden Mandarin (II). The only difference is that here the speaker adaptation was developed on an SPM in isolated-syllable mode, while the curve in Fig. 10 is for an HMM in continuous-speech input mode. The system could also adapt to environmental noise to some extent, as long as the noise was stable and of reasonable density. At the linguistic level, on the other hand, new word learning could be performed, and the language model could slightly adapt to the wording and writing style of the user although the learning function was not adequate in many cases. All these adaptation/learning functions, from the acoustic level to the linguistic level, could be performed on-line in real time, so performance could be improved continuously.

The earliest versions of Golden Mandarin (III) were completed in March 1995. They covered a vocabulary of roughly 100,000 words and had two different versions. Version(IIIa) was implemented on the same DSP card as was used for Golden Mandarin (II) with a single-chip Motorola DSP 96002D and could be installed on any personal computer, such as a 486, but the input mode was now in isolated prosodic segments, i.e., continuous within short sequences of a few words [44]. Version (IIIb), on the other hand, was implemented on a Sun SPARC 20 workstation, and the input mode was in continuous speech and complete sentences [37]. Version (IIIa) was primarily based on the recognition of words except that a few words could be concatenated to construct prosodic segments to solve the difficult problem of undefined word boundaries mentioned previously. Recognition of base syllables was primarily based on the context-dependent version of the PLUs listed in Table 5(b) so that the co-articulated nature of continuous speech could be properly modeled. Tone recognition was performed only for mono- and bi-character words because tones for words with three or more characters are actually redundant, as mentioned before. The language model was primarily based on word classes with some syntactic and semantic knowledge.

The complete system block diagram for Golden Mandarin (IIIa) is shown in Fig. 18, which is composed of five modules. The signal processor for front-end processing of endpoint detection and feature extraction and the user interface for display and user correction are straightforward. The prosodic segment recognizer is a core module. In this module the base-syllable strings and tone strings are first recognized with the aid of a lexicon and a set of prosodic segment rules describing how different words with given syntactic/semantic features can be constructed into prosodic segments. A lattice of words or prosodic segments is then constructed to be used by the linguistic decoder based on the Chinese language model. The learning module is, in general, very similar to that of Golden Mandarin (II), including incremental speaker adaptation with four stages, learning of environmental noise, new word learning, and adaptation of the language model as well. The major difference is that, here, because Golden Mandarin (IIIa) is primarily word-based using PLU models, the incremental speaker adaptation is based on sets of training words instead of sentences, which are phonetically balanced with respect to the chosen PLUs.

Golden Mandarin (IIIb) was implemented on a Sun Sparc 20 workstation with the input mode in continuous speech and complete sentences. The time needed to recognize a sentence is, on average, 1.27 times the length of the speech. The recognition of the base syllables is primarily based on the context-dependent version of the INITIAL/FINALs listed in Table 1, while tone recognition is performed with the context dependent tone models listed in Table 3, both with CHMM modeling. The language model is primarily based on word classes with some syntactic and semantic knowledge. The complete system block diagram is shown in Fig. 19, which is generally very similar to Fig. 18, except that the prosodic segment recognizer in Fig. 18 is replaced by a continuous speech acoustic recognizer. Most of the adaptation/learning functions of version (IIIa) are also present here in version (IIIb).

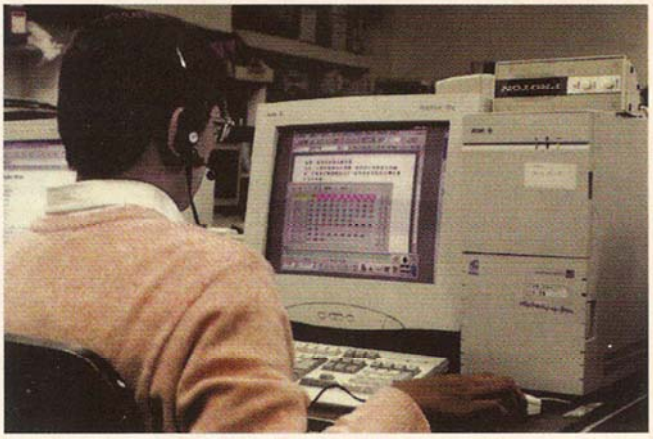
Although the Golden Mandarin (IIIa) and (IIIb) versions look reasonable and performance is satisfactory, further improved versions have been developed [38, 39], one of which will be presented here. In all the above systems, acoustic matching is performed first for the construction of the tonal-syllable lattices and linguistic decoding with Chinese language models is performed next to obtain output sentences. This is the way in which the overall architecture of the system can be simplified. However, in this way the acoustic and linguistic processors are essentially separated; thus, the search will very likely only give a local optimum for both processors. A new and different system architecture has therefore been developed whose two-stage simplified block diagram is shown in Figure 20. In the first stage of this architecture, to utilize the monosyllable-based structure of the Chinese language and to deal with the extremely large search space in continuous speech recognition with very large vocabulary, a fast matching module based on relatively coarse context-independent acoustic models is used to select enough tonal-

syllable candidates quickly. A word lattice is then constructed in the subsequent module from these selected tonal-syllable candidates based on the lexicon. In the second stage, only those words within the word lattice are considered, and the detailed matching module uses a time-synchronous dynamic programming algorithm. In this algorithm the knowledge and information from the acoustic models and the Chinese language model are naturally combined in a carefully designed process, such that the acoustic and linguistic processors are actually integrated into a single processor. Experiments indicate that this prototype system preserves all the features and functions of Golden Mandarin (IIIb) but with higher accuracy and speed.

The most recent prototype system, the Golden Mandarin (III) Windows 95 version, was completed in September 1996. It preserves all the features of Golden Mandarin (IIIb) and the advantages of the following, but it has been downsized from the Sparc 20 to an ordinary Pentium PC using its standard sound card and was implemented on MS Windows 95 (Chinese Version). The system is multi-modal in the sense that the user can easily switch between the voice input scheme and any other Chinese character input scheme provided by MS Windows 95, such as those based on phonetic symbols or radicals, either during the dictation or for correction purposes.

When the phonetic-symbol input scheme is used, the output sentence will be automatically generated by the linguistic decoding processes. In addition, new approaches to extract robust speech recognition feature parameters have been applied, and the acoustic models for tones and base syllables are actually adapted by including the noise characteristics in real-time [86-88]. As a result, this system becomes much more robust with respect to acoustic and environmental variabilities, including different characteristics and conditions for the microphone, and different types and levels of noise and interference.

Fast incremental speaker adaptation processes have been implemented with a specially developed user training interface as well. The adaptation results shown in Fig. 10 are in fact for this prototype system. The lower curve in Fig. 10 shows that with 6.8 minutes of speech produced by the new user the accuracy of tonal syllables can achieve 80.1% on average. Although the tonal-syllable accuracy on the lower curve in Fig. 10 doesn't look very high, it is for the top-1 candidates only. With more candidates for the tonal syllables included in the syllable/word lattice construction and linguistic decoding processes, the percentage of the correct tonal syllables that are included and considered in the linguistic decoding processes can be much higher. Powerful enough language models can then select the correct characters or words, even if the top-1 tonal syllable is not correct. This is why the upper curve in Fig. 10 for character accuracy is always significantly higher than the tonal-syllable accuracy. This is where the Chinese language models play the role. In Fig. 10 the character accuracy is averaged over a variety of different texts addressing different subject domains, and the accuracy of 86.4% with 6.8 minutes of training speech for the new user is believed to be good enough for a new user to start testing/using the system. Further improvements are definitely achievable with more training data. Also, since the user usually dictates his texts focusing on a much narrower subject domain, higher accuracy can very often be achieved if the language model can be adapted to that subject domain. Such a function for the user to enter his own "personal" text file to obtain his own "personal" language model is also included. The memory requirement of the system is roughly 8.5 MB, and the time needed for the complete process of recognizing an utterance on the Pentium 133 is roughly 1.55 times the length of the speech utterance. In other words, after the user produces the speech utterance, the recognized sentence should appear on the screen after a short waiting period of roughly 0.55 times the length of the utterance.

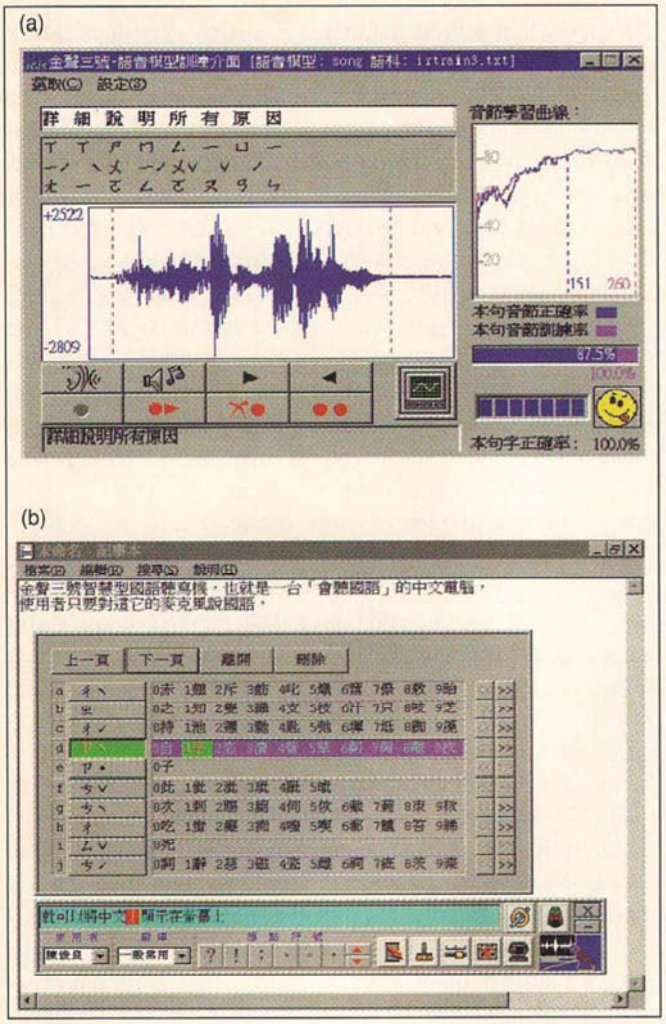


21. The Golden Mandarin (III) Windows 95 version prototype system.

A photo of the complete prototype system is shown in Fig. 21, while the training and dictation phases of the user interface of the system is shown in Fig.22(a) and (b). On the training user interface in Fig. 22(a), the training sentence together with the corresponding phonetic symbols are shown on the top, the segmented input speech waveform in the middle, the on-line recognized output on the bottom, and an averaged learning curve together with current accuracy status on the right. The user can easily realize the progressive improvements made when producing the training sentences one by one during the training process and so on. On the dictation user interface in Fig. 22(b), for incorrectly recognized characters the user can open a window to see all the tonal-syllable candidates or character candidates and select the desired ones easily using the mouse. The user can also correct the errors directly by voice by producing the words or phrases incorrectly recognized and editing the sentences.

Related Spoken Language Applications

Up to this point, we have primarily focused on dictation applications, i.e., the input of Chinese characters into computers using voice, because this area is extremely attractive although difficult. However, if dictation technology is available, it can easily be extended to other related spoken language applications, especially when a very large vocabulary can be taken care of. Here, two examples will be presented to



22 (a) The training user interface and (b) the dictation user interface of the Golden Mandarin (III) Windows 95 version prototype system.

show the potential in this direction. These examples simply indicate there will be plenty of space for future applications to be developed in such areas in addition to the dictation applications discussed previously.

The first example is for voice retrieval for Chinese databases. From experiences with other languages, use of speech-recognition technology in information retrieval (IR) for databases provides users with a convenient computer interface environment [89-92]. For the Chinese language, because the language is not alphabetic and input of Chinese characters into computers is difficult, voice retrieval of Chinese databases is apparently another important application area of Mandarin speech recognition in addition to dictation applications. In fact, by properly utilizing the special monosyllable-based structure of the Chinese language, the many areas of technology discussed in this article can in fact be easily applied to voice retrieval of Chinese databases with very large vocabulary. Here, a typical example for such application systems will be presented [93-95], a nice feature of which is that the content of the target database can be used to train a special database-specific linguistic decoder for spoken queries for the database. Such a database-specific linguistic decoder can thus automatically transform the original Mandarin dictation system discussed above into a Chinese database voice retrieval system, with the output being the desired documents in the database instead of the corresponding characters for the input speech in a dictation system.

The block diagram of this first example system is shown in Fig. 23. First, the document-analysis subsystem segments the documents in the database into words and deletes irrelevant words for retrieval (such as function words). It also uses the rest of the texts in the database to construct two sets of statistical parameters, i.e. the syllable-based document characteristics and the document feature vectors, to be used by the other two subsystems (the speech recognition subsystem and the IR subsystem, respectively). These two sets of statistical parameters obtained from the documents in the database in fact constitute the core of the database-specific linguistic decoder mentioned above. When a natural-language speech query is entered, the speech recognition subsystem transcribes it into a relevant syllable string with syllables irrelevant for retrieval deleted. This subsystem includes two modules: the syllable-recognition module and the syllable-string-search module. The syllable-recognition module produces several tonal-syllable candidates for each syllable in the input query to construct a tonal-syllable lattice. The syllable-string-search module then performs the Viterbi search algorithm over the obtained tonal-syllable lattice using the syllable-based document characteristics of the database provided by the document analysis subsystem as the first-stage linguistic decoder.

On the other hand, the IR subsystem is composed of two modules, the vector-matching module and the document retrieval module. The vector-matching module compares the input query, q, with each piece of document, d, by comparing their feature vectors, $\mathbf{u}_q$ and $\mathbf{u}_d$, respectively, as the second-stage linguistic decoder. Here a set of very useful syllable-based feature parameters precisely describing the syllable-level statistical characteristics of the input query and the pieces of documents are carefully selected and used to construct $\mathbf{u}_q$ and $\mathbf{u}_d$ for q and d. In this way, the similarity measure between q and d, $S(q,d)$, can be easily defined by the angle between the two feature vectors as evaluated by inner products:

$$S(q,d) \overset{def}{=} \cos^{-1}\left[\frac{\mathbf{u}_q \cdot \mathbf{u}_d}{|\mathbf{u}_q||\mathbf{u}_d|}\right], \tag{9}$$

The smallest angle gives the desired documents. The document-retrieval module finally retrieves those documents selected by the vector-matching module.

The above database retrieval system has been successfully implemented as a working experimental system. The database tested includes 2,500 Chinese news items, with the average length of the news items on the order of 500 characters. A total of 200 natural-language queries were tested in the form of continuous speech produced by 10 speakers. The acoustic recognition module of the Golden Mandarin (III) Windows 95 version presented previously is used here to produce the tonal-syllable lattice. The precision rate for top-10 retrieved news items, i.e., the percentage of the news items addressing the desired subjects of the queries among the retrieved top-10

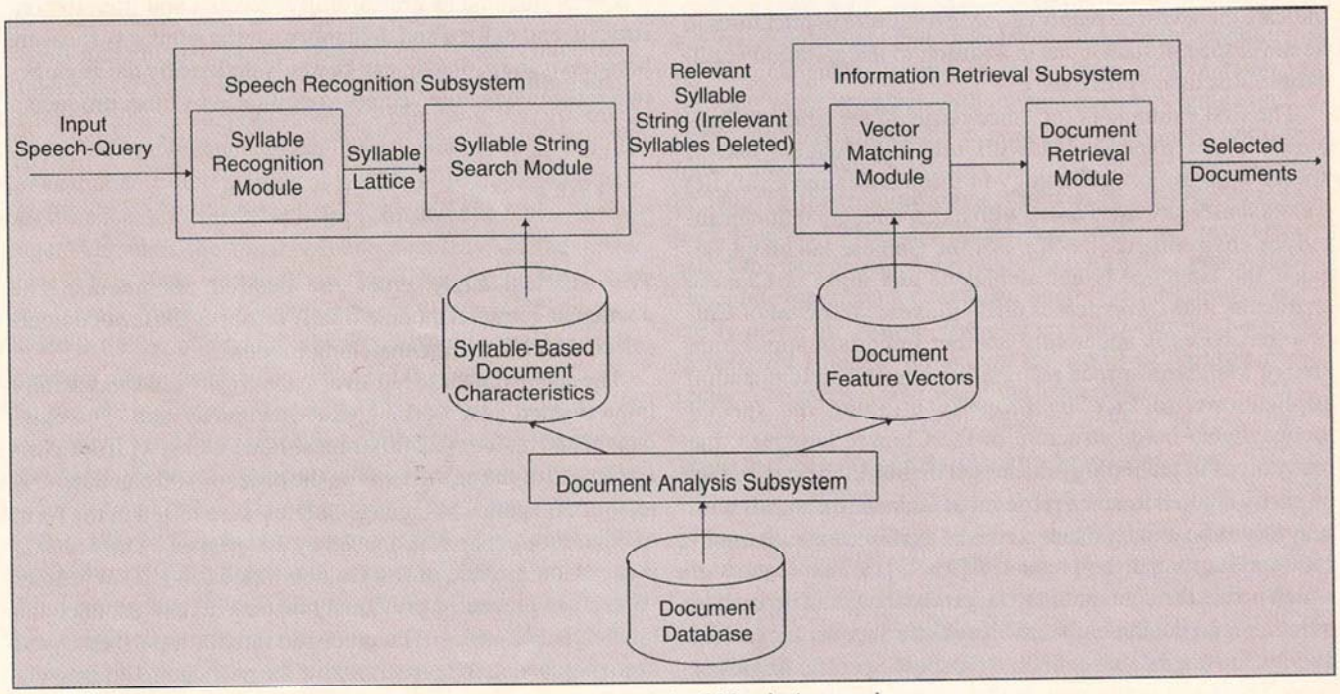
news items, is 86.4%, as compared to the precision rate of 89.6% for top-10 items retrieved by queries in the form of typed character strings.

The second spoken-language application example presented below involves the retrieval of Chinese resources over the Internet using unconstrained Mandarin speech queries. With the rapid growth of the electronic resources published and distributed over the Internet, the increasing demand for efficient, high-performance networked information retrieval with convenient and user-friendly interfaces is obvious [96]. Many efficient Internet search tools have been developed to allow users to formulate a request subject with unconstrained quasi-natural language queries, and it is always highly desired that such systems are capable of accepting queries with unconstrained speech. In particular, IR systems with speech-recognition capabilities are especially needed in the Chinese community because of the difficulties of entering Chinese characters into computers. The example system presented here for Internet information retrieval utilizes a syllable-based client-server architecture with a set of reliable character/syllable-level statistical feature parameters for efficient information retrieval using speech queries. These parameters make it possible to move the linguistic decoding processes to the server side. This can simplify the client-end requirements and make the language models easily adapted to the dynamic network resources.

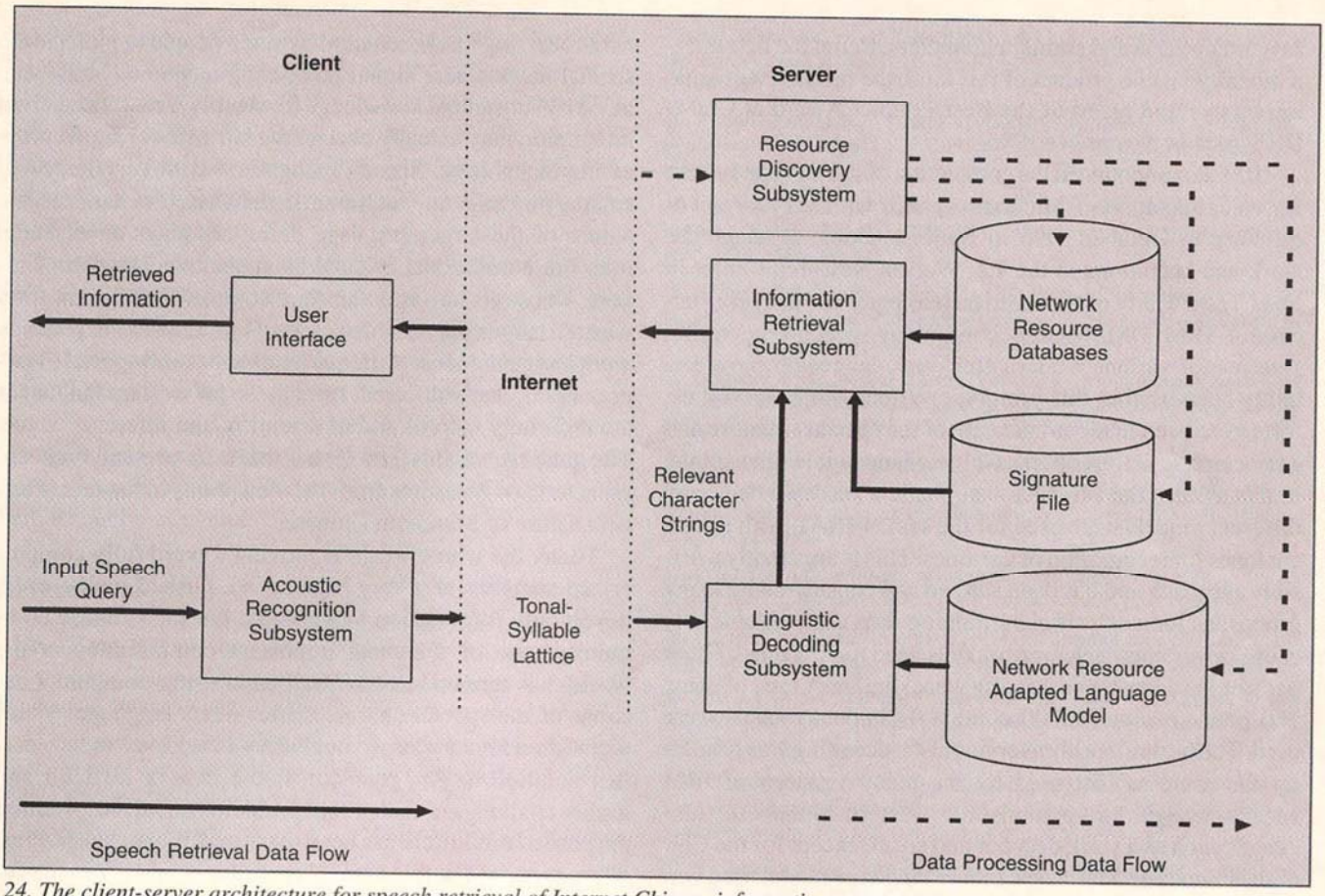
The basic client-server architecture of the example system is shown in Fig. 24. The server part includes a resource-discovery subsystem, an IR subsystem, and a linguistic-decoding subsystem, while the client end includes a user interface and an acoustic-processing subsystem. The resource-discovery subsystem at the server side automatically extracts pieces of relevant information (i.e., records) and uses them to construct the network resource databases. Signatures (i.e., statistical indices) for each of the records in the network resource databases are also generated to be stored in the signature file. These specially designed signatures for all the records include character-based and syllable-level statistical feature parameters of the records, especially considering the monosyllabic structure of the Chinese language. These signatures provide full-text indexing of the network resource databases. The network resource-adapted language model is also constructed based on the discovered resources using a specially designed data structure.

When a speech query is received at the client end, the acoustic-recognition subsystem first produces a tonal-syllable lattice to be transmitted to the server. The linguistic decoding subsystem at the server then generates a relevant character/syllable string based on this tonal-syllable lattice and the network resource-adapted language model. The IR subsystem finally receives this decoded relevant character/syllable string and retrieves the desired records from the network resource databases by evaluating the statistical similarity between the received relevant character/syllable string and the record signatures in the signature file. Also, as shown in Fig. 24, the acoustic recognition and linguistic decoding processes for very-large-vocabulary Mandarin speech recognition are now separated, one at the client end and the other at the server side. Without the overhead space necessary for linguistic decoding at the client end, it is easier for the acoustic-recognition subsystem to be combined with navigation tools such as Netscape to allow many users to enter their speech queries simultaneously. On the other hand, the linguistic decoding processes at the server side can have sufficient space to store a large number of statistical parameters for a powerful language model, and it is easier for this language model to be adapted according to the dynamic network resources.

The application system presented above has been successfully implemented as a working experimental system that provides unconstrained speech retrieval for real-time Chi-



23. A Chinese document database-retrieval system using unconstrained Mandarin speech.



24. The client-server architecture for speech retrieval of Internet Chinese information.

nese news services obtained from Internet news groups. The acoustic recognition module of the Golden Mandarin (III) Windows 95 version is used in the acoustic-recognition subsystem. The client end, including the user interface and the acoustic-recognition subsystem, is implemented on Pentium PCs under MS Windows 95, and the server is implemented on a Sparc 20 workstation. The network resource database contains more than 100,000 real-time news items. In the preliminary experiments, 200 speech queries produced by 10 speakers were tested. The character accuracy for these queries was roughly 90%, apparently due to the very powerful network resource-adapted language model at the server. The precision rate for top-10 retrieved news items was about 82.5% on average, which is almost the same as the top-10 precision rate for typed text queries.

Initial Industrial Efforts and Products

The first internationally visible product for voice dictation of Mandarin Chinese (and probably the only one up to the time of writing this article) commercially available on the market is, to the knowledge of the author, the Apple Chinese Dictation Kit produced by Apple Computer Inc., which was available in November 1995. A few other products may have appeared, but they are practically almost invisible internationally for various reasons. The Apple Chinese Dictation Kit works on a Chinese-language-equipped Power Macintosh computer with at least 4 MB of free memory. It is in principle an isolated-word-based system. In other words, it recognizes isolated words with very large vocabulary (12,000 multi-character words plus 3,500 single characters), and the input mode is primarily in isolated words. Since the user has to segment the input sentences into words and the words segmented by the user are not necessarily the ones in the lexicon stored in the system, it suffers from the problems with the isolated-word input mode as discussed previously.

Special measures have been developed to remedy this problem slightly. Since some frequently used words (most of them are mono-character), such as those standing for "in," "to," and quite a few function words are very naturally produced in concatenation with the preceding or following words to form simple phrases (for example, the words standing for "in" and "evening" becomes a phrase standing for "in the evening"; those for "to" and "Taipei" become "to Taipei"; "beauty" followed by a function word becomes "beautiful," etc. (see (h) in box), special efforts have been made to take care of such simple phrases (similar to "short prosodic segments"). As a result, the system can accept such simple phrases produced as a single continuous utterance. This is a good engineering solution, but it only solves a very small part of the problem. The system is speaker dependent, i.e., the user needs to spend more than two hours to read as many as 33 pages of texts to train the system before being able to use it. The base syllables are recognized using some kind of INITIAL/FINAL units with some degree of context dependency, while the tones are recognized using HMMs. Some language-model capabilities have been equipped, such

as word bi-gram and similar parameters. Being the first internationally visible product of this kind, the product was quite impressive and received the Best Product Award at COMDEX Asia in November 1995.

IBM also announced the completion of a prototype system for voice dictation of Mandarin speech with very large vocabulary in October 1996 in Beijing, China, although the work had been done at the T.J. Watson Research Center in New York. Many of the basic technologies for the very successful IBM HMM-based continuous-speech-recognition systems for various western alphabetic languages have certainly been used in this prototype system, while special efforts have been made to take care of the special structure and characteristic features of the Chinese language. For example, in this system the FINALs were made tone-dependent, and different models were used for the same FINAL with different tones for recognition of the tones. This is apparently a feasible approach and has been studied and considered by many groups, as long as sufficient training data are available, because in this approach more models need to be trained. Other parts of the acoustic processing were similar. A total of some 160 phonemes and 3,000 context-dependent models were used. The basic algorithm for linguistic decoding was primarily the same as that used by the many versions of IBM speech-recognition systems for western alphabetic languages such as a stack decoder and so on, except for the Chinese language model. The Chinese language model used was a word tri-gram trained from a huge Chinese text corpus, segmented with a carefully selected vocabulary of the 29,000 most frequently used words based on some simplifying rules to handle the problem of Chinese words being not well defined. This is also a good engineering solution, though it only solves a very small part of the problems as well. The system was trained by the speech data produced by many speakers, and it claimed to be speaker independent, and continuous speech could be accepted very well. Test data for six speakers were reported, with character accuracy ranging from 69.6% to 88.0% for texts taken from daily newspapers. Higher accuracies were obtained for native speakers of standard Beijing Mandarin.

Also, Motorola announced in November 1996 the successful development of technologies for recognition of continuous Mandarin speech of 10,000 words on an industry-standard PC. But no further information was available to this author at the time of writing this article.

Concluding Remarks

Research in speech recognition with very large vocabulary requires integration of expertise in many different disciplines, from signal processing to computer science and linguistics. The fundamental framework is certainly based on signal processing technology, but various areas of computer science and linguistics apparently play very crucial roles in solving this highly difficult problem. Without the many key techniques offered by computer science and the various levels of search constraints provided by linguistic knowledge, any solution is simply impossible. A title of "Intelligent Signal Processing" has been used by some people to indicate the special areas where signal processing technology has been aided by substantial knowledge from other areas; and such an integration may actually change the intrinsics of signal processing technology. Speech recognition with very large vocabulary is really one such area. In this case, the characteristic nature of the target language definitely plays an essential role. For a traditional oriental language like Mandarin Chinese, whose feature structures are completely different from western languages on which most of the mainstream research work has been focused, unique approaches and special measures significantly different from those for western languages are definitely helpful and of scientific and reference value. The purpose of this article has been to present such approaches and measures from the viewpoints of the characteristic nature of Mandarin Chinese.

Today the whole world is moving toward fully computerized societies at a very high speed, pushed by the ever-developing information technology, but the Chinese community, one of the most important communities in the world, has tremendous difficulties in using computers because of the special characteristics of its language. Voice dictation with very large vocabulary is believed to be a perfect solution to this problem, but it is very difficult and highly challenging with many problems unsolved. Another purpose of this article has been to try to stimulate the interest of more people in this area. Hopefully, some day in the future, all Chinese people, north and south, old and young, including our moms, dads, and kids, will be able to use computers easily and freely with voice input in their daily lives. Of course, there is still a very long way to go before this dream can come true. There was an old Chinese saying, "The integration of great efforts made by many people can build a castle" (see (i) in box). With advanced technology and great efforts made by many people, there is hope that such a great and beautiful day can come early.

Acknowledgments

The author wishes to thank his colleagues Dr. K.J. Chen and Dr. C.Y. Tseng, for their long-term support and collaboration on the research work on this subject. Special thanks also go to many of his former and current graduate students and research assistants (some have now become his colleagues also) who have been working on related subjects with him for years and making great contributions on various technologies and different versions of prototype systems developed, in particular, Dr. L.F. Chien, Dr. H.M. Wang, Dr. R.Y. Lyu, Dr. J.L. Shen, Dr. S.C. Lin, Ms Y.J. Yang, Mr. B.R. Bai, Mr. T.H. Ho, Dr. H.Y. Gu and Mr. F.C. Chou. This list continues with Mr. S.S. Lin, Mr. C.S. Yang, Mr. I.J. Hung, Mr. R.C. Yang, Mr. S.H. Hwang, Mr. H.Y. Hsieh, Mr. W.P. Chen, Mr. C.C. Hung, Mr. J.C. Weng, Mr. T.L. Yu, Mr. B.H. Cheng, Mr. C.W. Chen, Dr. M.H. Yu, Ms. G.H. Chen, Mr. M.C. Chen, Mr. C.D. Huang, Dr. J.K. Chen, Mr. R.S. Lin, Mr. C.L. Chen, Mr. S.L. Tu, Mr. Y.H. Lin, Mr. Y. Lee, Dr. F.H. Liu, Dr. C.H. Chang, Mr. S.H. Hsieh, Mr. C.H. Chen, Mr. C.H. Hwang, Dr. P.Y. Ting, Dr. L.J. Lin, Dr. C.C. Wu, Mr. C.C. Chen, Mr. I.H.

Kuo, Mr. H.J. Lin, Dr. J.H. Tseng, Dr. J.A. Liang, Mr. M.Y. Lee, Mr. R.C. Wang, Mr. B.S. Lin, Mr. Y.C. Chang, Mr. Y.C. Huang, Mr. C.Y. Lou, Mr. T.S. Lin, Mr. C.P. Nee, Mr. C.Y. Liao, Mr. Y.C. Wu, Mr. S.W. Lin, Mr. C.H. Su, Mr. Y.C. Huang, Mr. T.M. Hsu, Mr. J.L. Tsai, Mr. L.L. Hsu, Mr. C.C. Chao, Ms. K.W. Lu, and many others. The author really can't list all the names here.

The project has been supported by the National Science Council of the Republic of China on Taiwan since 1983 and by Academia Sinica at Taipei since 1986. The long-term financial support for the research in this area by these two organizations is also highly appreciated.

Lin-Shan Lee is a research fellow with Academia Sinica and a professor with National Taiwan University in Taipei, Taiwan, Republic of China.

References

1. "Guoyurbao Tzdian (Mandarin Chinese Daily Dictionary)," R. He, Ed., Taipei, Taiwan, R.O.C. : Guoyurbao Company, 1976.

2. L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc. 1993.

3. "Special Issue on Speech Recognition for Different Languages," *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific Publishing Co. Pte. Ltd., Vol. 8, No. 1, Feb. 1994.

4. "Special Issue on Speech Processing." *IEEE Communications Magazine*, Vol. 31, No. 11, Nov. 1993.

5. L.R. Rabiner, "Applications of Voice Processing to Telecommunications," *Proceedings of the IEEE*, Vol. 82, No. 2. Feb. 1994, pp. 199-228.

6. S. Furui, "Speaker-independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum, " *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 34, No. 1, pp. 52-59, Feb. 1986.

7. T. Ukita, E. Saito, T. Nitta, and S. Watanabe, "A Speaker Independent Connected Digit Recognition System Concatenating Statistically Discriminated Words," *IEEE Trans. on Signal Processing*, Vol. 40, No. 10, pp. 2412-2424, October 1992.

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

9. X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, and R. Rosenfeld, "The SPHINX-II Speech Recognition System : An Overview," *Computer Speech and Language*, Vol. 2, pp. 137-148, Feb. 1993.

10. L.R. Bahl, "A Fast Approximate Acoustic Match for Large Vocabulary Speech Recognition, " *IEEE Trans. Speech and Audio Processing*, Vol. 1, pp. 59-67, Jan. 1993.

11. T. Hanazawa, K. Kita, S. Nakamura, T. Kawabata, and K. Shikano, " ATR HMM-LR Continuous Speech Recognition System," ICASSP'90, pp. 53-56, New Mexico, U.S.A., April 1990.

12. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, Vol. 77, pp. 257-286, Feb. 1989.

13. F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," *Proceedings of the IEEE*, Vol. 73, pp. 1616-1624, 1985.

14. H. Murveit, et. al, "Large-Vocabulary Dictation Using SRI's Decipher™ Speech Recognition System: Progressive Search Techniques," ICASSP'93, Vol. 2, pp. 319-322, Minneapolis, Minnesota, U.S.A. April 1993.

15. L.R. Bahl, et al., "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task," ICASSP'95, Vol. 1, pp. 41-44, Detroit, Michigan, U.S.A., May 1995.

16. P. Jeanrenaud, et. al., "Reducing Word Error Rate on Conversational Speech from the Switchboard Corpus," ICASSP'95, Vol. 1, pp. 53-56, Detroit, Michigan, U.S.A., May 1995.

17. P.C. Woodland, et. al., "The 1994 HTK Large Vocabulary Speech Recognition System," ICASSP'95, Vol. 1, pp. 53-56, Detroit, Michigan, U.S.A., May 1995.

18. X. Aubert and H. Ney, "Large Vocabulary Continuous Speech Recognition Using Word Graphs," ICASSP'95, Vol. 1, pp. 49-52, Detroit, Michigan, U.S.A., May 1995.

19. J.-L. Gauvain, et. al., "Developments in Continuous Speech Dictation using the ARPA WSJ Task," ICASSP' 95, Vol. 1, pp. 65-68, Detroit, Michigan, U.S.A., May 1995.

20. D. Pye, et. al., "Large Vocabulary Multilingual Speech Recognition Using HTK," 4th European Conference on Speech Communication and Technology, Vol. 1, pp. 181-184, Madrid, Spain, Sept. 1995.

21. L. Lamel, et. al., "Issues in Large Vocabulary, Multilingual Speech Recognition," 4th European Conference on Speech Communication and Technology, Vol. 1, pp. 185-188, Madrid, Spain, Sept. 1995.

22. C. Dugast, et. al., "The Philips Large-Vocabulary Recognition System for American-English, French and German," 4th European Conference on Speech Communication and Technology, Vo. 1, pp. 197-200, Madrid, Spain, Sept. 1995.

23. D.B. Roe, J.G. Wilpon, "Whither Speech Recognition : The Next 25 years," *IEEE Communications Magazine*, Nov. 1993, pp. 54-62.

24. T. Lee, et. al., "Tone Recognition of Isolated Cantonese Syllables," *IEEE Trans. on Speech & Audio Processing*, Vol. 3, May 1995, pp. 294-209.

25. Z. Wang, et. al. "Methods Towards the Very Large Vocabulary Chinese Speech Recognition," 4th European Conference on Speech Communication and Technology, Vol. 1, pp. 215-218, Madrid, Spain, Sept. 1995.

26. T. Huang, et. al., " Language Processing for Chinese Speech Recognition," 1994 International Symposium on Speech, Image Processing and Neural Networks, pp. 151-154, Hong Kong, Apr. 1994.

27. Y. Hao, D. Fang, "Speech Recognition Using Speaker Adaptation by System Parameter Transformation," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 1, pp. 63-68, Apr. 1994.

28. Y. Gau, et. al., "Tangerine: A Large Vocabulary Mandarin Dictation System," ICASSP' 95, Vol. 1, pp. 77-80, Detroit, Michigan, U.S.A., May 1995.

29. H.-W. Hon, K.-F. Lee, et. al, "Toward Large Vocabulary Mandarin Chinese Speech Recognition," ICASSP'94, Adelaide, Australia, Vol. 1, pp. 545-548, 1994.

30. J.-K. Chen, et. al., "Large Vocabulary Word Recognition Based on Tree-Trellis search," ICASSP'94, Vol. 2, pp. 137-140, Adelaide, Australia, 1994.

31. J.-K. Chen, et. al., "Large Vocabulary Word-based Mandarin Dictation System," 4th European Conference on Speech Communication and Technology, Madrid, Spain, Sept. 1995.

32. Y.R. Chao, "A Grammar of Spoken Chinese," Univ. of California at Berkeley Press, 1986.

33. C.-C. Cheng, "A Synchronic Phonology of Mandarin Chinese," The Hague, Mouton, 1973.

34. S.-M. Lei and L.-S. Lee, "Digital Synthesis of Mandarin Speech Using Its Special Characteristics," J. Chinese Inst. Eng., Vol. 6, No 2, pp. 107-115, Mar. 1983.

35. V.A. Fromkin, "Tone-A Linguistic Survey," New York: Academic, 1987.

36. J. Barnett, et. al., "Comparative Performance in Large-Vocabulary Isolated Word Recognition in Five European Languages," 4th European Conference on Speech Communication and Technology, Madrid, Spain, Sept. 1995.

37. H.Wang, et. al, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data," to appear on *IEEE Transactions on Speech and Audio Processing*, 1997.

38. T.-H Ho, et. al., "Fast and Accurate Continuous Speech Recognition for Chinese Language with Very Large Vocabulary," 4th European Conference on Speech Communication and Technology, PP. 211-214, Madrid, Spain, Sept. 1995.

39. H. Hsieh, et. al., "Use of Prosodic Information to Integrate Acoustic and Linguistic Knowledge in Continuous Mandarin Speech Recognition with Very Large Vocabulary," International Conference on Spoken Language Processing, Philadelphia, Pensylvania, U.S.A., Oct. 1996.

40. Lin-shan Lee, 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 Processing*, Vol. 1, No. 2, pp. 158-179, April, 1993.

41. L.-S. Lee, et. al, "Golden Mandarin (II) - An Improved Single-chip Real-time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary," ICASSP'93 Vol. 2, pp. 503-506, Minneapolis, Minnesota, U.S.A. May, 1993.

42. L.-S., et. al. "Golden Mandarin (II) - An Intelligent Mandarin Dictation Machine for Chinese Character Input with Adaptation/Learning Functions," 1994 International Symposium on Speech, Image Processing and Neural Networks, Hong Kong, Apr. 1994.

43. S. Fujio, et. al., "Prediction of Prosodic Phrase Boundaries Using Stochastic Context-free Grammar," International Conference on Spoken Language Processing, Sept. 1994, Yokohama, Japan, pp. 839-842.

44. R. Lyu, et. al., "Golden Mandarin (III) A User Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary," ICASSP'95, Vol. 1, pp. 57-60, Detroit, Michigan, U.S.A., May 1995.

45. L.-S. Lee, et. al, "A Mandarin Dictation Machine Based upon Chinese Natural Language Analysis," the 10th International Joint Conference on Artificial Intelligence, AAAI, Aug. 1987, Milano, Italy, pp. 619-621.

46. W.J. Yang, J.C. Lee, Y.C. Chang and H.C. Wang, "Hidden Markov Model for Mandarin Lexical Tone Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 36, pp. 988-992, July 1988.

47. X.X.Chen, et.al.,"A Hidden Markov Model Applied to Chinese Four-tone Recognition," ICASSP'87, pp. 797-800, Dallas, TX, 1987.

48. S.-H. Chen, Y. -R. Wang, "Tone Recognition of Continuous Mandarin Speech Based on Neural Networks," *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 2, pp. 146-150, March 1995.

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

50. X. Huang, et. al., "Unified Stochastic Engin (USE) for Speech Recognition," ICASSP'93, pp. 636-639, Minneapolis, Minnesota, U.S.A., Apr. 1993.

51. M. Hwang, et. al., "Improving Speech Recognition Performance via Phone-Dependent VQ Codebooks and Adaptive Language Models in SPHINX-II," ICASSP'94, Vol. 1 pp. 549-552, Adelaide, Australia, Apr. 1994.

52. L.S. Lee, et. al., "Special Speech Recognition Approaches for the Highly Confusing Mandarin Syllables Based on Hidden Markov Models," *Computer Speech and Language*, Vol. 5, No. 2, pp. 181-201, Apr. 1991.

53. F. Liu, Y. Lee, and L.S. Lee, "A Direct-Concatenation Approach to Train Hidden Markov Models to Recognize the Highly Confusing Mandarin Syllables with Very Limited Training Data," *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 1, pp. 113-119, Jan. 1993.

54. Y.M. Lee, L.S. Lee. "Continuous Hidden Markov Models Integrating Transitional and Instantaneous Features for Mandarin Syllable Recognition," *Computer Speech and Language*, Vol. 7, pp. 247-263, 1993.

55. Ren-yuan Lyu, et. al., "A new Approach for Mandarin Base-syllable Recognition Based Upon Segmental Probability Model (SPM)," Proc. Int. Conf. Computer Processing of Oriental Languages, pp. 201-206, Taejon, Korea, May 1994.

56. W. Chou, et. al., "Segmental GPD Training of HMM Based Speech Recognizer," Proc. ICASSP'92, Vol. 1, pp. 473-476, San Francisco, CA, U. S. A., March 1992.

57. P.C. Chang, B.H. Juang, "Discriminative Training of Dynamic Programming Based Speech Recognizer," ICASSP'92, Vol. 1, pp. 493-496, San Francisco, CA, U.S.A. March 1992.

58. B.H. Juang, and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. Signal Processing*, Vol. 40, pp. 3043-3054, Dec. 1992.

59. L.R. Rabiner, J.G. Wilpon, and B.H. Juang, "A Segmental K-means Training Procedure for Connected Word Recognition," *AT&T Technical Journal*, Vol. 65, No. 3, pp. 21-31, May-June 1986.

60. B.H. Juang, L.R. Rabiner,"The Segmental K-Means Algorithm for Estimating Parameters of Hidden Markov Models," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 38, No. 9, pp. 1639-1641, September 1990.

61. C.-H. Lee, et. al, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition," *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. 37, No. 11, pp. 1649-1658, Nov. 1989.

62. J.K. Baker, "The Dragon System - An Overview," *IEEE Trans. Acoustic, Speech and Signal Processing*, Vol. ASSP-23, pp. 24-29, Feb. 1975.

63. L.R. Bahl, et. al., "Recognition of Isolated-Word Sentences From A 5000-Word Vocabulary Office Correspondence Task," ICASSP'83, Boston, MA, U. S. A., pp. 1065-1067, 1983.

64. Eng-Fong Huang, Frank K. Soong and Hsiao-Chuan Wang, "The Use of Tree-trellis Search for Large-vocabulary Mandarin Polysyllabic Word Speech Recognition," *Computer Speech and Language*, pp. 39-50, August 1994.

65. S. Furui, "Unsupervised Speaker Adaptation Based on Hierarchical Spectral Clustering," *IEEE Trans. on Acoustic, Speech, and Signal Processing*, Vol. 37, No. 12, pp. 1923-1930, Dec. 1989.

66. C.-H. Lee, et. al., "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on Signal Processing*, Vol. 39, No. 4, pp. 806-814, Apr. 1991.

67. X. Huang, and K.F. Lee, "On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition," *IEEE Trans. On Speech and Audio Processing*, Vol. 1, No. 2, pp. 150-157, Apr. 1993.

68. J.-L. Shen, et. al., "Incremental Speaker Adaptation Using Phonetically Balanced Training Sentences for Mandarin Syllable Recognition Based on Segmental Probability Models," International Conference on Spoken Language Processing, Yokohama, Japan, pp. 443-446, Sept. 1994.

69. S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoustic, Speech and Signal Processing*, Vol. 35, pp. 400-411, 1987.

70. A. Nadas, "Estimation of Probabilities in the Language Model of the IBM Speech Recognition System," *IEEE Trans. Acoustic, Speech and Signal Processing*, Vol. 32, pp. 859-861, 1984.

71. L.R. Bahl, F. Jelinek, and R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligent*, Vol. PAMI-5, pp. 179-190, 1983.

72. A.M. Derouault and B. Merialdo, "Natural Language Modeling for Phoneme-to-text Transcription," *IEEE Trans. Pattern Analysis and Machine Intelligent*, Vol. PAMI-7, pp. 742-749, 1985.

73. L.R. Bahl et.al., "A Tree-based Statistical Language Model for Natural Language Speech Recognition," *IEEE Trans. Acoustic., Speech and Signal Processing*, Vol. 37, pp. 1001-1008, 1989.

74. R. Kuhn, R.D. Mori, "A Cache-based Natural Language Model for Speech Recognition," IEEE Trans. Pattern Analysis and Machine Intelligent, PAMI-12, No. 6, pp. 570-583, Jun. 1990.

75. S.M. Ross, "Introduction to Probability Models." London, UK: Academic, 1985.

76. H.Y. Gu, et. al., "Markov Modeling of Mandarin Chinese for Decoding the Phonetic Sequence into Chinese Characters," Computer Speech and Language, Vol. 5, No. 4, pp. 363-377, Oct. 1991.

77. P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer, "Class-based N-gram Models of Natural Language," *Computational Linguistics*, Vol. 18, No. 4, pp. 467-479, Dec. 1992.

78. Lin-Shan Lee, et. al., "An Efficient Natural Language Processing System Specially Designed for the Chinese Language," *Computational Linguistics*, vol. 17, No. 4, pp.347-374, Dec. 1991.

79. L.-F. Chien, et. al., "A Best-first Language Processing Model Integrating the Unification Grammar and Markov Language Model for Speech Recognition Applications" *IEEE Trans. Speech and Audio Processing*, Vol. 1, No. 2, pp. 221-240, Apr. 1993.

80. Y.-J. Yang, et. al., "An Intelligent and Efficient Word-Class-Based Chinese Language Model for Mandarin Speech Recognition with Very Large Vocabulary," The 1994 International Conference on Spoken Language Processing, pp. 1371-1374, Yokohama, Japan, Sept. 1994.

81. Y.-C. Chang, et. al., "Methodology, Implementation and Applications of Word-class-based Language Models in Mandarin Speech Recognition" R.O.C. Computational Linguistics Conference VII, Hsinchu, Taiwan, R.O.C., Aug 1994, pp. 17-30.

82. K.-J. Chen, C.-R. Huang, "Information-based Case Grammar," COLING'90, Vol. 2, pp. 54-59.

83. H. Schutze, "Part-of-speech Induction from Scratch," ACL'93, pp. 251-258, 1993.

84. J.F. Sowa, "Conceptual Structure," Reading, MA U.S.A.: Addison-Wesley, 1984.

85. S.C. Lin, et. al., "Training, Detection and Adaptation of Multi-domain Language Models for Mandarin Speech Recognition," ROC Computational Linguistics Conference IX, Tainan, Taiwan, ROC, Aug. 1996

86. Jia-lin Shen, et. al., "Robust Speech Recognition Features Based on Temporal Trajectory Filtering of Frequency Band Spectrum," International Conference on Spoken Language Processing, Philadelphia, Pennsylvania, U.S.A., Oct. 1996.

87. M.J.F. Gales and S.J. Young, "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination," *Computer Speech and Language*, pp. 289-307, Sept. 1995.

88. H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, No. 4, Oct. 1994.

89. V. Zue, et. al., "PEGASUS: A Spoken Dialogue Interface for On-line Air Travel Planning," *Speech Communication*, Vol. 15, pp. 331-340, 1994.

90. C.-H. Lee, "Stochastic Modeling in Spoken System Design," *Speech Communication*, Vol. 15, pp. 311-312, 1991.

91. M. Yamada, et. al., "A Spoken Dialogue System with Active/Non-active Word Control for CD-ROM Information Retrieval," *Speech Communication*, Vol. 15, pp. 355-365, 1994.

92. G. Salton, "Introduction to Modern Information Retrieval," NY, McGraw-Hill, 1983.

93. S.-C. Lin, et. al., "A Syllable-Based Very-Large-Vocabulary Voice Retrieval System for Chinese Databases with Textual Attributes," 4th European Conference on Speech Communication and Technology, PP. 203-206, Madrid, Spain, Sept. 1995.

94. S.-C. Lin, et. al. "Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary and Unlimited Domains," 4th European Conference on Speech Communication and Technology, pp. 1203,-1206, Madrid, Spain, Sept. 1995.

95. S-C. Lin, L-F. Chien, K-J Chien, and L-S. Lee, "An Efficient Voice Retrieval System for Very-Large-Vocabulary Chinese Textual Databases with a Clustered Language Model," ICASSP'96, pp. 287-290, Atlanta, USA, May, 1996.

96. L-F. Chien, et. al., "Natural Language Information Retrieval with Speech Recognition Techniques for Chinese Network Resource Discovery," International Workshop on Information Retrieval with Oriental Languages, Korea, June, 1996.