An Improved Mandarin Speech Dictation Machine for Intelligent Terminals of Advanced Communication Networks

Lin-shan Lee*, Chiu-yu Tseng**, Hung-yan Gu* and Fu-hua Liu***

*Dept. of Computer Science and Information Engineering, National Taiwan University,
** Institute of History and Philology, Academia Sinica,
***Dept. of Electrial Engineering, National Taiwan University,
Taipei, Taiwan, Republic of China

Abstract

The transmission of text information in future advanced networks is always important, hence the difficulties in the input of Chinese characters to computers and therefore networks become an additional technical constraint for the development of intelligent terminals communication networks in the Chinese community. The primary reason is that Chinese language is not alphabetic and every Chinese character is a complicated square graph. Speech input is considered as one approach to this problem. This paper describes an improved experimental Mandarin dictation machine for the input of Chinese characters to computers and networks.Considering the special characteristics of Chinese language, syllables are chosen as the basic units for dictation. The machine is divided into two subsystems. The first is to recognize the syllables using Hidden Markov Models (HMM's) trained by special algorithms. Because every syllable can represent many different characters with completely different meaning, and can possibly form different multi-syllabic words with syllables on its right or left, the second subsystem then identifies the exact characters from the syllables and corrects the errors in syllable recognition using a specially designed Chinese language model.

I. Introduction

The new information age today can be characterized as a time of exploding demands and challenges for telecommunication developments, among which the development of advanced networks which can provide efficient and versatile services is one of the major trend. For the telecommunications engineers in the Chinese community, there is an additional technical constraint for their efforts made in the development of advanced networks, i.e., the difficulties in the input of Chinese characters (or ideographs) for terminals, computers, and therefore networks[1]. The primary reason is that Chinese language is not alphabetic. Every Chinese character is a complicated square graph, many of which are composed of different radicals organized in a very irregular manner, and there are at least 20 thousand commonly used different characters,. Although there are at least more than 50 different methods developed to input Chinese characters to computers today, none of them can provide the users a convenient input system with efficiency comparable to alphabetic languages[1], because they are either too slow,too complicated, or need special training. The radical input systems and the phonetic symbol input systems are good examples for them[1]. Because the Transmission of text information is always very important in advanced networks, efficient techniques for input of Chinese characters are thus crucial for telecommunication developments in Chinese community. Speech input is considered as one possible approaches to this problem. An initial effort to develop a Mandarin dictation machine has been made earlier[2], and recently an improved experimental Mandarin dictation machine for the speech input of Chinese text was being implemented. This paper will present the design approach and system structure of this Mandarin dictation machine, To our knowledge this is the first successful real-time Mandarin dictation machine developed in the world.

To develop a dictation machine is very difficult, we first define the scope of the research by the following limitations. The input speech is in the form of isolated syllables instead of continuous speech (The choice of syllables as the dictation unit will be discussed in detail later). This avoids the problem of processing or segmenting continuous speech waveforms. Even if the input is made by voice of isolated syllables, such an input method is still much more efficient than any of the currently existing Chinese input systems.The dictation machine is speaker dependent only. The fact that it is trained for only one user at one time is completely acceptable considering practical applications. The first stage goal of this system is to have only 90% correction for sentence structures and words in the Chinese textbooks of the primary schools in Taiwan, Rep. of China. This means for a sentence of 10 characters, one of the characters will be wrong in average and should be found by the user on the screen and corrected from the keyboard. Such a performance is still much more efficient than any of the currently existing input systems, and the primary school Chinese textbooks already cover most of the everyday Chinese language. On the other hand, the operation of such a dictation machine has to be real-time,because the input to a terminal or computer must be performed in real-time. Based upon the above definitions and limitations on the task goal, such a Mandarin dictation machine is not only practically attractive, but technically obtainable. As will be clear later in this paper, the above goal is exactly achieved in our system.

Because the design approach of a dictation machine depends heavily on the target language, the machine described here is in fact quite different from several other dictation machines developed for several dictation machines developed for several other languages[3,10] in the world. Based on the detailed discussion in the next section considering the special structure of the Chinese language, the Mandarin syllables instead of Chinese words (most of them are multi-syllabic) are chosen as the basic units for the dictation. The dictation machine is then divided into two subsystems. The first one is to recognize the syllables using signal processing techniques, i.e., Hidden Markov Models (HMM's) trained by special algorithms, but this is not very helpful at all because in general every syllable can represent many defferent characters with completely different meaning, and can possibly form different multi-syllabic words with syllables on its right or left. Therefore the second subsystem, a Chinese language model, is to identify the correct characters from the syllables by forming correct words (most of them have more than one characters) and then sentences with maximum likelihood statistically. This subsystem is in fact trained using a large database of Chinese texts. In the following, the basic design approach of the machine considering the special structure of the Chinese language will be

CH2682-3/89/0000-0965 $1.00 © 1989 IEEE

discussed in section II, the overall system structure will be described in section III, and the detailed techniques will be presented in the next few sections. The test results and the conclusion will finally be given.

II. Considerations for the Special Structure of Chinese Language

There are at least some 80 thousand commonly used words in Chinese[11]. Such a size is prohibitively large for today's speech recognition technology. Therefore the words can not be used as the dictation units if a practical dictation machine is to be designed. On the other hand, there are at least 20 thousand commonly used Chinese characters, each character being monosyllabic. Each of the 80 thousand commonly used Chinese words is composed of from one to several characters (a very small fraction of them have only one character), therefore most of the words are multi-syllabic and a small fraction of them are monosyllabic. Although the total number of monosyllabic words is small, they appear in everyday Chinese language very frequently. Nevertheless, we note that the total number of phonologically allowed syllables in Mandarin speech is only about 1300. In other words, if we use the 1300 syllables as the dictation units, all the words or characters will be covered. Therefore, use of these syllables as the dictation units will allow the replacement of the 20 thousand commonly used characters by the 1300 syllables for computer input. However, the small number of syllables implies another difficult problem, that is, relatively high number of homonyms for which many different characters will share the same syllable. In other words, after a syllable is recognized from speech signal, it may form different multi-syllabic words with adjacent syllables on its right or left, it can also be a monosyllabic word. However, as far as the complete meaningful sentence is concerned, there will be only one correct solution. This is where the Chinese language model becomes very important, and is exactly the way the Chinese people listen to their language. These are also some additional reasons to use syllables as the dictation units. First, all of these syllables are of open syllabic structure, i.e., they always end with a vowel with the exceptions of vowels plus nasals -n and -ng. This makes the detection of the end points relatively easier. Furthermore, although most of the Chinese words are multi-syllabic with several characters, most of the morphemes, i.e., the minimum meaningful units, in Chinese are monosyllabic and composed of only a single character. Based on the above observations on the special structure of Chinese language[11], the use of syllables as the basic units to recognize Mandarin Chinese sentences in the dictation machine is a very natural choice.

Another very special important feature of Mandarin Chinese language is the existence of the lexical tones for the syllables. Chinese is a tonal language, in general every character is assigned a tone and the tones have lexical meaning in Mandarin. There are basically five different tones, i.e., the high-level tone (usually referred to as the first tone), the mid-rising tone (the second tone), the mid-falling-rising tone (the third tone), the high-falling tone (the fourth tone), and the neutral tone (the fifth tone). It has been shown[12,13] that the primary difference for the tones is the pitch contours, there exist standard patterns for the pitch contours of the first four lexical tones, and the tones are essentially independent of other acoustic properties of the syllables. One example is shown in Fig. 1, where the pitch contours for the first four tones of three vowels and two diphthongs [a*, u, i, ai, au-1, 2, 3, 4] for the same speaker are plotted as functions of time. In each drawing the horizontal scale is the time in units of frame number, the vertical scale is the pitch period in units of sampling period. The number on each curve indicates the tone. It can be seen that although the vowels or diphthongs are completely different, the basic patterns for the pitch contours for the four tones are essentially the same, and they are in fact the same for all different syllables. If the differences among the syllables due to lexical tones are disregarded, only 408 syllables are required to represent all the pronunciations for Mandarin Chinese. This means every syllable can be considered as the combination of two completely independent parts, a first-tone syllable among the 408 possible syllables (disregarding the tones) and the tone among the five possible choices[12,13]. This means the recognition of the syllables can also be divided into two parallel procedures, and by removing the effect of the tones the number of different candidates for the syllable recognition of reduced to 408, which is a very reasonable size.

* The transliteration symbols used in this paper are the Mandarin Phonetic Symbols II (MPS II). The numerical numbers following each syllable denotes the lexical tone of the syllable.

III. The Complete Language Recognition Hierarchy

Based on the considerations described above, the basic system structure for the Mandarin dictation machine is shown in Fig. 2. The system is divided into two subsystems. The first subsystem is used to recognize the syllables using speech signal processing techniques, for example, to transform the input sentence in Mandarin speech form, such as "你是一架會聽國語的電腦 (You are a computer who can listen to Mandarin)" into its corresponding syllables, i.e., [ni-3] [shr-4] [i-2] [jia-4] [huei-4] [tieng-1] [guo-2] [iu-3] [de-5] [dian-4] [nau-3]. Here the input speech signal is assumed to be a sequence of isolated syllables, therefore the endpoint detection is easy and the primary task is to recognize the syllables. The second subsystem is then to identify the correct character for each syllable using a Chinese language model. Every syllable like [ni-3] or [shr-4] in the above example can represent many completely different characters with the same pronunciation, but there exists only one set of characters, such as in the above example, "你是一架會聽國語的電腦", which can form frequently used multi-syllabic words such as "國語 (Mandarin)" and "電腦 (Computer)" and a sentence with maximum likelihood statistically. Therefore the task of the second subsystem is to transform the input sequence of syllables into the output text formed by characters. As long as 90% of the characters are correct, the performance will be satisfactory.

The complete recognition hierarchy is shown in Fig. 3. For the first subsystem of syllable recognition, the endpoints for each syllable are first detected, the corresponding first-tone syllable (disregarding the tones, such as [ni] for the first syllable in the above example) and the tone (such as the third tone for the same example) are then recognized independently in parallel, because as discussed previously every syllable can be considered as the combination of these two independent parts. The results are then combined to determine the syllable [ni-3]. It will be shown later that the recognition of the first-tone syllable and the tones are both difficult, and errors always occur. We therefore have to provide information for confusing first-tone syllables. For example, [ni] being not a very confident result and the second choice being [mi], and confusing tones, for example, the second choice being the fourth tone and so on, to the second subsystem such that the errors in the first subsystem or acoustic level recognition can hopefully be corrected by the second subsystem or linguistic level identification.

For the second subsystem of identifying the correct characters

for each syllable, we need to first form monosyllabic character hypotheses from the syllables, and then find character strings with maximum likelihood statistically using the Chinese language model. To use the above example, although there are many characters all correspond to the syllable [ni-3], many to the syllable [shr-4] and so on, but only when the syllable [ni-3] represents the character "你 (You)" and the syllable [shr-4] represents the character "是 (Are)" can the combination of them form the most frequently used (or, with maximum likelihood) character string "你是 (You Are)". Similarly, although there are many characters all correspond to the syllable [guo-2] and many to [iu-3], there is only one very frequently used multi-syllabic word "國語 (Mandarn)" has the pronunciation" [guo-2] [iu-3]". In this way, all the recognized syllables are first used to form possible character hypotheses, and the Chinese language model will then try to select a single character string with the maximum likelihood. The likelihood of a character string is computed using the frequency of occurrence of the character string from a large database of Chinese texts. Finally a single character string will be obtained and appear on the screen as the output text. Any errors in this output can then be further corrected by the user manually from the keyboard.

IV. The First Subsystem - Syllable Recognition or Acoustic Level

The recognition of the 408 first tone syllables in the first subsystem is in fact very difficult[14]. This is because the 408 first tone syllables consist of about 38 confusing sets, each of which has from about 4 to 19 confusing syllables[14]. Good examples are the A-set: {[a], [ba], [pa], [ma], [fa], [da], [ta], [na], [la], [ga], [ka], [ha], [ja], [cha], [sha], [tza], [tsa], [sa]} and AN-set: {[an], [ban], [pan], [man], [fan], [dan], [tan], [nan], [lan], [gan], [kan], [han], [jan], [chan], [shan], [ran], [tzan], [tsan], [san]}. It has been shown[14] that with standard approaches of Dynamic Time Warping, LPC and Itakura distance measures to recognize a subset of these syllables, the achievable recognition rates are as low as 25%-30%. This tells how difficult it is to correctly recognize these syllables and why currently available techniques for English words can not be applied directly. An initial/final two-pass training approach for Hidden Markov Models (HMM's) is thus specially designed[15] to recognize these very confusing syllables. Here "final" means the vowel or diphthong parts of the syllable but including the medials and nasal ending (if any), and "initial" is the initial consonant of the syllable. Table 1 is a list of all the 408 syllables. The vertical scale of the table lists all the 38 finals (including a null final), and the horizontal scale of the table lists all the 22 possible initials (including a null initial). Therefore every row of the table represents a confusing set, consisting of syllables sharing the same final but with different initials.

Hidden Markov Models (HMM's) have been successfully applied to speech recognition in many different cases[16] in recent years. In this approach a Markov model is trained for each recognition candidate using training samples, and in the recognition phase the candidate model with the highest probability to generate the unknown test samples is taken as the recognition result. Each model is specified by a series of states described a set of parameters evaluated from training samples. In order to correctly recognize the above highly confusing 408 syllables, a special initial/final two-pass training approach to HMM's are designed to take care of the 38 confusing sets in the 408 syllables. As discussed previously, in each confusing set all the syllables possess the same final and the only differences is in the initials. Therefore it is desired that the rear parts of the models for syllables in the same confusing set are identical so that the effect of the final will be reduced to minimum and the different initials be emphasized. In other words, if two syllables belong to the same confusing set, the parameters of their models should be equal in the last several states such that their differences in the initials can better differentiate the two syllables. This is the basic idea of the two-pass training concept. The block diagram of the initial/final two-pass training approach is shown in Fig. 4. The training syllables are first end-point detected and segmented into initials and finals. In the first pass, the 408 syllables are classified into 38 confusing sets as discussed above, and for each confusing set a single HMM for the common final shared by all the syllables in the set is trained using the finals of all these syllables obtained in the segmentation process. After the 38 final HMM's have been created, the second-pass training starts to obtain totally 408 initial syllables. By considering the information carried in the transition region from initial to final and the fact that only relatively few training frames are available for initials due to the very short duration in many of Mandarin initials, we use the segmented initial combined with a portion of the final following the initial to train the front parts of the HMM's for the syllables. This means some portion of final frames are used in the training of both the final models and the initial models, and we require that the final model and the initial model in a syllable model share a common state such that they can be smoothly combined into a single syllable model.

The recognition of the lexical tones is also difficult, especially when the fifth tone causes serious confusion. Although some initial efforts in the lexical tone recognition have been made and very encouraging results have been obtained [17], slightly different method is used in this dictation machine to better distinguish the fifth tone from the other four tones. HMM's are trained for the five lexical tones with feature vectors defined as $x_t=[\log(f_t)+\log(f_{t+1}), \log(f_t)-\log(f_{t+1}), \log(e_t)+\log(e),p]$ where f_t and e_t are respectively the pitch frequency and short time energy at frame t, e is the maximum of the energy e_t's for a specific syllable, and p is the time length of the voiced part for the syllable. The energy and duration parameters are used here to emphasize the difference between the fifth tone and other four tones.

V. The Second Subsystem - Chinese Language Model or Linguistic Level

Assume a string of n characters is denoted by

$$C = c_1, c_2, c_3 ..... c_n$$

where c_i is the i'th character. Let S denote the unknown speech signal on the basis of which the dictation machine will make its decision about which character string was spoken. A very natural decision rule for the dictation machine to decide in favor of a character string C^* is that

$$P(C^* \mid S) = \max_{C} P(C \mid S)$$

in other words, the character string C^* is chosen if it maximizes the probability P(C | S) for all possible character string C. This is actually a well-known decision rule which has been applied in many different problems including communication problems in which the receiver has to decide which symbol string C is transmitted given the received signal S. This decision rule is also intuitively acceptable, because the best the machine can do is to choose a character string which has the highest probability given the observed speech signal.

Using the Bayes formula,

$$P(C \mid S) = \frac{P(S \mid C)\, P(C)}{P(S)}$$

For a given observed speech signal S, P(S) is fixed for all possible character strings C. Therefore all the machine has to do is to find a character string C which maximizes P(S|C) P(C). The first term P(S|C) can be evaluated using the syllable components of the character string C and the corresponding probabilities obtained from the syllable HMM's as described in the previous section, while the second term P(C) should be evaluated by the Chinese language model. A direct relation to compute P(C) is

$$P(C) = \prod_{i=1}^{n} P(c_i | c_{i-1}, c_{i-2}, \ldots c_1)$$

but practically such a relation will cause difficulties in implementation because even for a moderate vocabulary size the probabilities $P(c_i\, c_{i-1}\, c_{i-2}, \ldots c_1)$ would be too large to be estimated, stored or retrieved. In our dictation machine we simplify the above relation by assuming a first-order markov language model, i.e.,

$$P(c_i | c_{i-1}, c_{i-2}, c_1) = P(c_i | c_{i-1})$$

and the probabilities $P(c_i\, c_{i-1})$ are estimated using a large database of Chinese texts. In our dictation machine in order to properly adjust the relative weight with which the decision depends on the first term P(S C) from the acoustic level evaluation and the second term P(C) from the linguistic level, the actual objective function to be

$$F = P(S|C)\, P(C)^{W}$$

where W is a parameter to be adjusted empirically to optimize the recognition rate. This is because it is hard to tell whether the two probabilities P(S|C) and P(C) can be estimated with equal accuracy and which one of the two causes more errors in recognition. In actual implementation a large database for Chinese texts are used to train the probabilities $P(c_i | c_{i-1})$. These probabilities are estimated by simply counting the frequency of occurrence for different characters to appear in adjacent positions. The training database includes the Chinese textbooks for primary schools in Taiwan, Republic of China, the texts taken from everyday newspapers and magazines, etc.

VI. Conclusion

A Mandarin dictation machine has been successfully designed and under implementation currently. Such a machine is probably a very important step to develop intelligent terminals for advanced communication networks for the Chinese community. Further improvements for the machine is still under progress.

References

[1] Proceedings of 1986, 1987 and 1988 International Conference on Chinese Computing, Aug. 1986, Singapore, Jun 1987, Chicago USA, and Aug 1988, Toronto, CANADA, Chinese Language Computer Society.

[2] L.-S. Lee, C.-Y. Tseng, K.J. Chen, J. Hunang, "The Preliminary Results for a Mandarin Dictation Machine Based Upon Chinese National Language Analysis", 1987 International Joint Conference on Artificial Intelligence, Milano, Italy, Aug. 1987.

[3] F. jelinek, "The development of an Experimental Descrete Dictation Recognizer", Proc IEEE Vol. 73, NO. 11, Nov 1985, pp. 1616-1623.

[4] A. Averbuch, et al. "An IBM PC Based Large-vocabulary Isolated-utterance Speech Recognizer", 1986 International Conference on Acoustics, Speech and Signal Processing, Tokyo, Apr. 1986, Vol. 1, pp. 53-56.

[5] M. Picheny, et al. "A Real-time IBM PC Based Large-vocabulary Isolated-word Speech Recognizer", Voice Processing, Online Publication, Pinner, UK, 1986

[6] A. M. Derouault, "Context-dependent Markov Models for Large-vocabulary Speech Recognition", 1987 International Conference on Acoustics, Speech and Signal Processing, Dallas, Apr. 1987.

[7] B. Merialdo, "Speech Recognition with Very large Size Dictionary", 1987 International Conference on Acoustics, Speech and Signal Processing, Dallas, Apr. 1987, pp. 364-367.

[8] A. Averbuch, et al, "Experiments with the Tangora 20,000 Word Speech Recognizer", 1987 International Conference on Acoustics, Speech and Signal Processing, Dallas, Apr 1987, pp. 701-704.

[9] P. D'Orta, "Large-vocabulary Speech Recognition: A System for the Italian Language", IBM Journal of Research and Development, Vol. 32, No. 2, Mar 1988, pp. 217-226.

[10] B. Merialdo, "Multi-level Decoding for Very-large-size dictionary Speech Recognition", IBm Journal of Research and Development, Vol. 32, no.2, Mar 1988. pp. 227-237.

[11] Chao, Y.R. (1968) A Grammar of Spoken Chinese, University of California Press, Berkeley.

[12] S. -M. Lei, L. -S. Lee, "Digital Synthesis of Mandarin Speech Using Its Special Characteristics", Journal of the Chinese Institute of Engineers, Vol. 6, No.2, Mar. 1983, pp. 107-115.

[13] V.A. Fromkin, "Tone-A Linguistic Survery", Academic Press, New York, 1978.

[14] Michael Wagner, Wei Wang, Helen Ho, "Isolated-word Recognition of the Complete Vocabulary of Spoken Chinese",1986 International Conference on Acoustic, Speech and Signal Processing, Apr. 1986, Tokyo, JAPAN.

[15] Fu-hua Liu, "Mandarin Syllable Recognition Based Upon Hidden Markov Models with Two-pass Training", Master Thesis, Dept. of Electrical Engineering National Taiwan Uninersity, Jun 1988.

[16] L. R. Rabiner, B. H. Juang, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, Vol. 3. No. 1, Jan 1986.

[17] W. J. Yang, et al. "Hidden Markov Models for Mandarin Lexical Tone Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 36, No. 7, July 1988.

				2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
					ch	ish	r	tz	ts	s	g	k	h	j	ch	sh	d	t	n	l	b	p	m	f
1				2	3	4	5	6	7															
2	a	5	9	10	11		12	13	14	15	16	17					18	19	20	21	22	23	24	25
3	o	25																			27	28	29	30
4	e	31	32	33	34	35	36	37	38	39	40	41					42	43	44	45				
5	ai	45	47	48	49		50	51	52	53	54	55					56	57	58	59	60	61	62	
6	eh	63																						
7	ei	64	65			66	67			68	69		70				71		72	73	74	75	76	77
8	au	78	79	80	81	82	83	84		85	86	87	88				89	90	91	92	93	94	95	
9	ou	96	97	98	99	100	101	102	103	104	105	106					107	108	109	110		111	112	113
10	an	114	115	116	117	118	119	120	121	122	123	124					125	126	127	128	129	130	131	132
11	en	133	134	135	136	137	138	139	140	141	142	143							144		145	146	147	148
12	ang	149	150	151	152	153	154	155	156	157	158	159					160	161	162	163	164	165	166	167
13	eng	168	169	170	171	172	173	174	175	176	177	178					179	180	181	182	183	184	185	186
14	i	187												188	189	190	191	192	193	194	195	196	197	
15	u	198	199	200	201	202	203	204	205	206	207	208					209	210	211	212	213	214	215	216
16	iu	217												218	219	220			221	222				
17	ia	223												224	225	226				227				
18	ie	228												229	230	231	232	233	234	235	236	237	238	
19	iai	239																						
20	iau	240												241	242	243	244	245	246	247	248	249	250	
21	iou	251												252	253	254	255		256	257			258	
22	ian	259												260	261	262	263	264	265	266	267	268	269	
23	in	270												271	272	273			274	275	276	277	278	
24	iang	279												280	281	282			283	284				
25	ing	285												286	287	288	289	290	291	292	293	294	295	
26	ua	296	297	298	299					300	301	302												
27	uo	303	304	305	306	307	308	309	310	311	312	313					314	315	316	317				
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	365		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	iin	400												401	402	403								
37	iung	404												405	406	407								
38	er	408																						

Table 1. All the 408 Mandarin syllables disregarding the tones. The number indicates the sequence number used in our database.

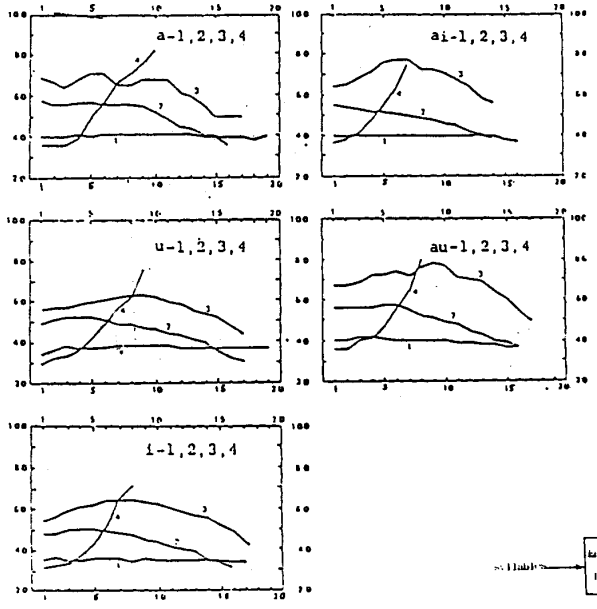

Fig.1. The pitch contours of [a, u, i, ai, au -1, 2, 3, 4] for the same speaker, sampling period vs frame number. The horizontal axis is the time in units of frame number, the vertical axis the pitch period in units of sampling period. The number on each curve indicates the tone.

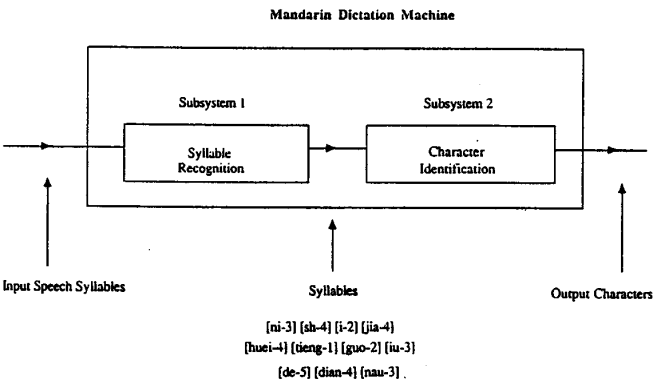

Fig.2 The Basic Structure of the Mandarin Dictation Machine

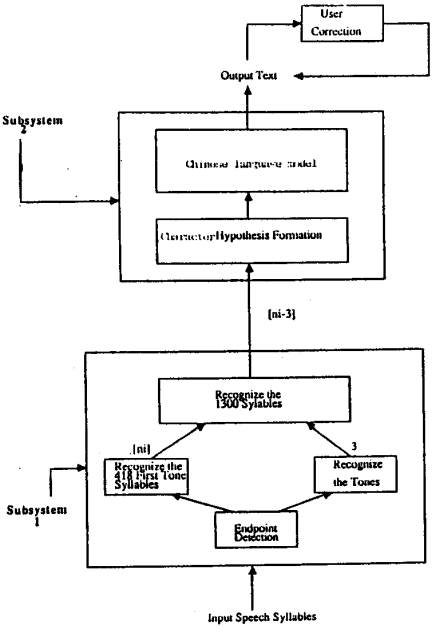

Fig.3 The Complete Hierarchy and Overall System Structure for the Mandarin Dictation Machine

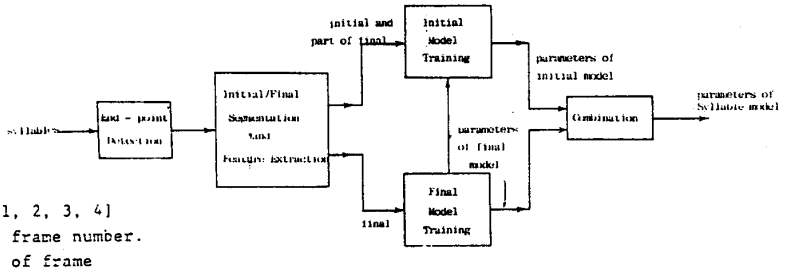

Fig. 4 The block diagram of the two-pass training procedure