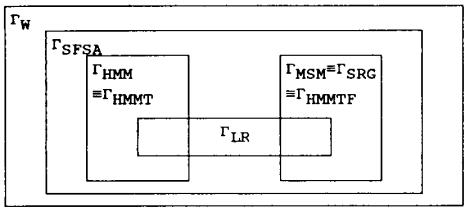


Fig. 1. Some relations among the classes of weighted languages associated to SFSA's, HMM's, LRHMM's, MSM's, and SRG's.

nition, the here-adopted probabilistic formalism is widely used for learning or estimation purposes (Baum–Welch algorithm), and for recognition. However, a general problem is to find an optimal sequence of states which are compatible with an observed sequence of acoustic symbols (Viterbi algorithm). In this case, the results of this work are not directly applicable, and further work is necessary in this direction. Nevertheless, the difference between both points of view is merely the weighted space over which the weighted languages are defined. The relationships between weighted languages generated by stochastic finite automata by using one space or another has been already studied by Santos [25].

We would like to point out that LRHMM's and LRHMMT's are among the most widely used in automatic speech recognition, and there are particular cases that are not equivalent from a probabilistic point of view. From a theoretical perspective, this can affect the application of reestimation and recognition procedures and the mathematical properties of the estimated models (e.g., Baum–Welch theorem). From a practical point of view, on the other hand, it is not clear what is the actual impact of the differences in the applications of the LRHMM and the LRHMMT to automatic speech recognition. Consequently, it is not obvious which formalism can be better used to represent the knowledge in automatic speech recognition.

All the results presented in this correspondence are of an essentially theoretical nature. Further research is clearly necessary, to establish what the actual consequences of these results are on the practical aspects of the application of stochastic finite state networks to automatic speech recognition.

Acknowledgment

The author wishes to thank Dr. E. Vidal and the anonymous reviewers for their criticisms and suggestions.

References

[1] L. Bahl, F. Jelinek, and R. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, no. 2, pp. 179–190, 1983.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[3] L. Rabiner and B. Juang, "Introduction to hidden Markov models," *IEEE ASSP Mag.*, pp. 4–16, Jan. 1986.

[4] J. Baker, "The Dragon system—An overview," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 1, pp. 24–29, 1975.

[5] L. Rabiner, S. Levinson, and M. Sondhi, "On the application of vector quantification and HMM's to speaker independent, isolated word recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1075–1105, 1983.

[6] S. Levinson, L. Rabiner, and M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.

[7] S. Levinson, "Structural methods in automatic speech recognition," *Proc. IEEE*, vol. 73, no. 11, pp. 1625–1650, 1985.

[8] L. Rabiner, "Mathematical foundations of hidden Markov models," in *Recent Advances in Automatic Speech Understanding and Dialog*, H. Niemann, Ed. (NATO ASI Series). New York: Springer, 1988.

[9] A. Salomaa, *Formal Languages*. New York: Academic, 1973.

[10] R. Kashyap, "Syntactic decision rules for recognition of spoken words and phrases using stochastic automaton," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-1, no. 2, pp. 154–163, 1979.

[11] R. Schwarth *et al.*, "Acoustic-phonetic decoding of speech," in *Recent Advances in Speech Understanding and Dialog Systems*, H. Niemann, Ed. New York: Springer, 1988.

[12] H. Cerf-Danon *et al.*, "Speech recognition experiments with 10000 word dictionary," in *Pattern Recognition and Applications*, P. Devijver and J. Kittler, Eds. New York: Springer, 1987.

[13] K. Fu, *Syntactic Pattern Recognition and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[14] R. Gonzalez and M. Thomason, *Syntactic Pattern Recognition: An Introduction*. Reading, MA: Addison-Wesley, 1978.

[15] L. Miclet, *Structural Methods in Pattern Recognition*. North Oxford Academic, 1986.

[16] N. Abramson, *Information Theory and Coding*. New York: McGraw-Hill, 1966.

[17] C. Wetherell, "Probabilistic languages: A review and some open questions," *Comput. Surveys*, vol. 12, no. 4, pp. 361–379, 1980.

[18] H. Rulot and E. Vidal, "An efficient algorithm for the inference of circuit-free automata," in *Proc. NATO Advanced Research Workshop Syntactic Pattern Recognition*, Ferrate *et al.*, Eds. New York: Springer, 1988.

[19] S. Levinson, "A unified theory of composite pattern analysis for automatic speech recognition," in *Computer Speech Processing*, F. Fallside and W. Woods, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[20] ——, "Some experiments with a linguistic processor for continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, no. 6, pp. 1549–1556, 1983.

[21] A. Derouault and B. Merialdo, "Natural language modeling for phoneme to text transcription," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, no. 6, pp. 742–749, 1986.

[22] S. Nakagawa and M. Jilan, "Syllable-based connected spoken word recognition by two pass $O(n)$ DP matching and hidden Markov model," in *Proc. ICASSP86*, 1986, pp. 21.14.1–21.14.4.

[23] B. Merialdo, "Phonetic recognition using HMM's and maximum mutual information training," in *Proc. ICASSP88*, 1988, pp. 111–114.

[24] P. F. Brown, "The acoustic-modeling problem in automatic speech recognition," IBM Res. Center, Yorktown Heights, NY, Res. Rep. RC-12750, May 1987.

[25] E. S. Santos, "Realizations of fuzzy languages by probabilistic max-product and max-min automata," *Inform. Sci.*, vol. 8, pp. 39–53, 1975.

[26] K. F. Lee, "Large-vocabulary speaker-independent continuous speech recognition: The SPHNIX systems," Carnegie-Mellon Univ., Tech. Rep. CMU-CS-88-148, 1988.

A Mandarin Dictation Machine Based Upon a Hierarchical Recognition Approach and Chinese Natural Language Analysis

LIN-SHAN LEE, CHIU-YU TSENG, K. J. CHEN, JAMES HUANG, CHIA-HWA HWANG, PEI-YIH TING, LONG-JI LIN, AND C. C. CHEN

***Abstract*—This correspondence describes the first experimental Mandarin dictation machine developed in the world for the input of**

Manuscript received May 17, 1988; revised November 23, 1989. Recommended for acceptance by C. Y. Suen.

L.-s. Lee, C.-h. Hwang, P.-Y. Ting, L.-j. Lin, and C. C. Chen are with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, Republic of China.

C.-y. Tseng and K. J. Chen are with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of China.

J. Huang is with the Department of Modern Languages and Linguistics, Cornell University, Ithaca, NY 14853.

IEEE Log Number 9034828.

Mandarin speech (spoken Chinese language) into computers. Considering the special characteristics of the Chinese language, syllables are chosen as the basic units for dictation. The machine is designed based on a hierarchical language recognition approach, in which acoustic signals are first recognized as a sequence of syllables, possible word hypotheses are then formed from the syllables, and the complete sentences are finally obtained. This approach is implemented by two subsystems. The first recognizes the syllables using speech signal processing techniques, including the recognition of the finals, initials, and tones of the syllables, respectively. Because every syllable can represent many different characters with completely different meaning, and can possibly form different multisyllabic 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 by first forming all possible word hypotheses from the syllables then finding out one combination of the word hypotheses which is grammatically valid in a sentence. The detailed syllable recognition algorithms, word formation rules, parser, grammar, and the syntactic checking algorithms are described in the correspondence. Using everyday newspaper text in the form of isolated syllables as input, the preliminary test results indicate that such a dictation machine is not only practically attractive, but technically achievable.

***Index Terms*—Chart parser, hierarchical approach, Mandarin Chinese, natural language, speech recognition, syllables, syntactic checking, word hypotheses.**

I. Introduction

Although the computer was introduced to the Chinese community more than 20 years ago, the input of Chinese characters (ideographs) into computers is still a very difficult and unsolved problem [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. There are at least 20 thousand commonly used different characters. Today, for the input of Chinese characters to computers, although there are at least more than 100 different methods developed, it is well known that none of them can provide the users a convenient input system with efficiency comparable to alphabetic languages [1]. These methods are either too slow, or too complicated, or need special training. For example, the radical input systems have rules too difficult to remember and the phonetic symbol input systems are too slow. This is the basic motivation for the development of a Mandarin dictation machine. From the computer input point of view, the desire for a dictation machine is much stronger for Chinese than for English.

The development of such a dictation machine is very difficult; we therefore 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 continuous speech waveforms. Also, due to the monosyllabic nature of Mandarin, distinct syllables appear to be rather acceptable and even more enunciated in Chinese. In other words, even if the voice input is made by isolated syllables, such an input method is not only much more efficient than any of the currently existing Chinese input systems, but considered very convenient and natural in practical applications. The dictation machine is speaker dependent only. The fact that it is trained for only one user at a time is completely acceptable considering practical applications. The first phase goal of this system is 90% accuracy for the sentences in the Chinese textbooks of the primary schools in Taiwan, Republic of China. This means for a sentence of 10 characters, one of the characters will be wrong on the average and will 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. The final goal of such research is of course to have a real-time system, although currently the real-time requirement is not considered crucial. In other words, the first phase goal is to develop a machine which, although not real-time, has the potential to be improved and implemented in real-time in the future. This is because the primary purpose of this research is to demonstrate the technical feasibility of a Mandarin dictation machine instead of actually implementing a prototype system. Due to the same reason, only a small dictionary for demonstration purposes is to be established in the first phase. As will be discussed in detail later, for the Chinese textbooks of the primary schools, it is estimated that a dictionary consisting of about 35 thousand Chinese words is necessary. However, our first phase dictionary consists of only 5 thousand words. Therefore the machine will work well for sentences formed by these words, but new words will have to be added to the dictionary as needed. Such a system is in fact sufficient to demonstrate the technical feasibility in dictating Mandarin speech by machines, as will be clear later. Based upon the above definitions and limitations on the task goals, such a Mandarin dictation machine is not only practically attractive if implemented as a product, but technically obtainable using currently available technologies. As will be clear later in this paper, the above goals are almost achieved in our system. Although similar systems have been designed or implemented for other languages with different approaches [2]–[5], this is believed to be the first experimental dictation machine developed for Mandarin Chinese in the world [6].

Based on the detailed discussion in the next section considering the special structure of the Chinese language, Mandarin syllables instead of Chinese word hypotheses (most of them are multisyllabic) are chosen as the basic units for the dictation. The dictation machine is designed based on a hierarchical language recognition approach, in which acoustic signals are first recognized as a sequence of syllables, possible word hypotheses are then formed from the syllables, and the complete sentences are finally obtained. This approach is implemented by two subsystems. The first one is to recognize the syllables using speech signal processing techniques. However, this is not very helpful at all because in general every syllable can represent many different characters with completely different meaning, and can possibly form different multisyllabic words with syllables on its right or left. Therefore the second subsystem is to identify the correct characters from the syllables by forming correct word hypotheses (most of them have more than one characters) which are grammatically valid in a sentence. This subsystem is designed by carefully considering the characteristics of the Chinese natural language, primarily the syntactic structure. In the following, the basic approach of the machine considering the special structure of the Chinese language will be discussed in Section II, the complete language recognition hierarchy will be described in Section III, and the detailed techniques will be presented in the next few sections. Test results and conclusions 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 [7]. Even if word formation rules are used to reduce the number of different words which are necessary in the vocabulary as will be discussed in detail later, it is estimated that at least 35 thousand words are needed for everyday Chinese language. Such a size is still prohibitively large for today's speech recognition technology. Therefore the words cannot be used as the dictation units if a practical dictation machine is to be designed. On the other hand, there 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 multisyllabic 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, a 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 multisyllabic words with adjacent syllables on its right or left, and it can also be a monosyllabic word. However, as far as the complete grammatical sentence is concerned, there will be only one correct solution. This is where the Chinese natural language analysis becomes very important, and is exactly the way the Chinese people listen to their language. There 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 multisyllabic 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 [7], 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 four 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), and the high-falling tone (the fourth tone). It has been shown [8], [9] that the primary difference for the tones is the pitch contours, there exist standard patterns for the pitch contours for the different 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 four tones of three vowels and two diphthongs [a[1], 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 418 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 418 possible syllables (disregarding the tones) and the tone among the four possible choices [8], [9]. 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 is reduced to 418, which is a very reasonable size.

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 sequence in Mandarin

[1]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.

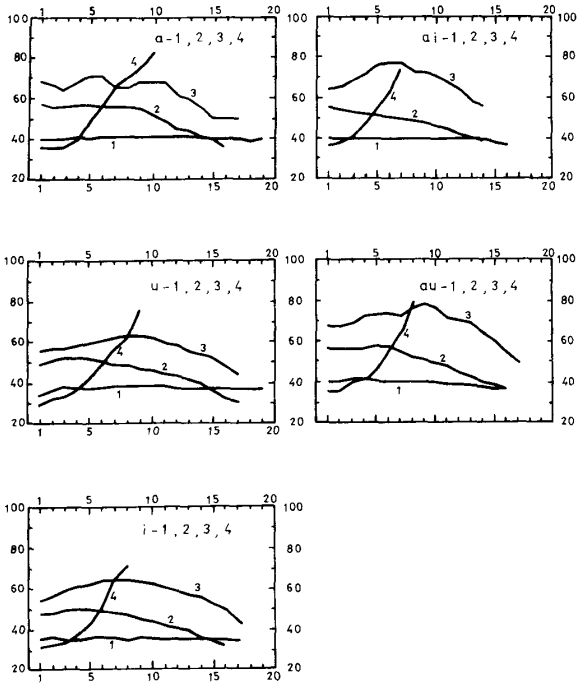


Fig. 1. The pitch contours of [a, u, i, ai, au, − 1, 2, 3, 4] for the same speaker, sampling period versus frame number. The horizontal axis is the time in units of frame number, the vertical axis the pitch period in units of sampling period. The number on each curve indicates the tone.

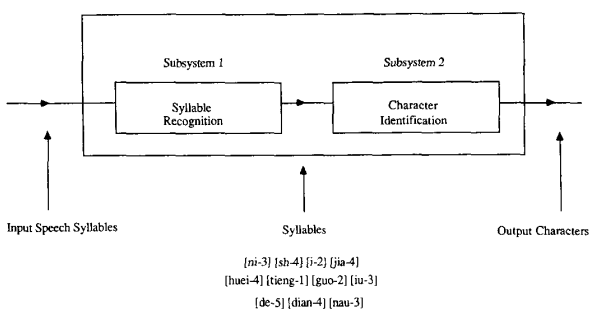


Fig. 2. The basic structure of the Mandarin dictation machine.

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 Chinese natural language analysis approaches. 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 meaningful multisyllabic words such as "國語 (Mandarin)" and "電腦 (Computer)" and a grammatically valid sentence. 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.

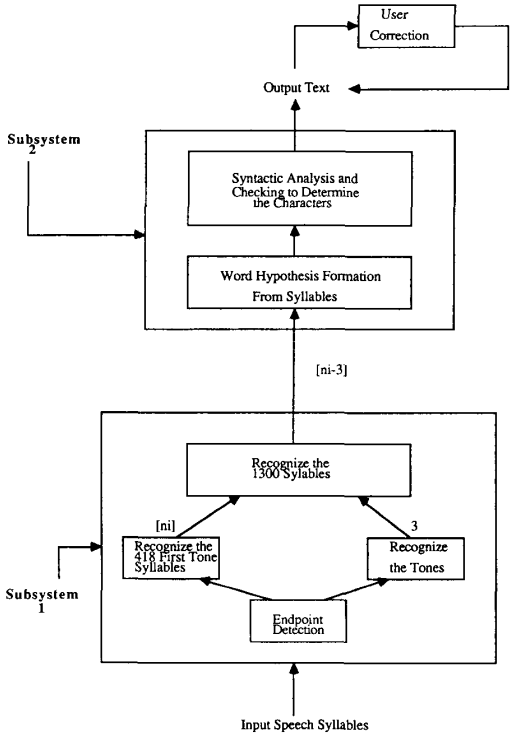


Fig. 3. The complete hierarchy and overall system structure for the Mandarin dictation machine.

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, 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 syntactic level identification.

For the second subsystem of identifying the correct characters for each syllable, we need to first form multisyllabic word hypotheses from each syllables. To use the above example, although there are many characters all correspond to the syllable [guo-2] and many to [iu-3], there is only one multisyllabic word "國語(Mandarin)" has the pronunciation "[guo-2] [iu-3]." Similarly, there are many characters all pronounced as the syllable [dian-4] and many as the syllable [nau-3], but there is only one multisyllabic word "電腦(Computer)" pronounced as "[dian-4] [nau-3]." This can be achieved by matching with the words in a dictionary. But this does not solve the problem well. First, although there is a relatively small number of monosyllabic words, they appear very frequently in everyday Chinese language. In the above example, syllables such as [ni-3], [shr-4], [i-2], [jia-4], [huei-4], [tieng-1] all correspond to monosyllabic words. The corresponding characters or words for them cannot be identified in the above way. They can even form ambiguous multisyllabic word hypotheses; for example, the syllable [i-2] can combine with the syllable to its left [shr-4] to form a wrong word hypothesis "適宜(Suitable, [shr-4] [i-2])," or with the syllable to its right [jia-4] to form a wrong word hypothesis "移駕(Move, [i-2] [jia-4])." The problem becomes even worse when errors occur in the recognized syllable, for example, if the tone of the first syllable [ni-3] is incorrectly recognized as the fourth tone, than the wrong syllable [ni-4] can be combined with the syllable [shr-4] to its right to form a wrong word hypothesis "逆勢(Bad Situation, [ni-4] [shr-4])." The next operation of syntactic analysis then serves as a filter to rule out all ungrammatical combinations of multisyllabic and monosyllabic word hypotheses and only a single syntactically valid sentence will be obtained and appear on the screen as the output text. Any errors in this output sentence can then be further corrected by the user mannually from the keyboard. From the above description, it can be seen that in this approach although the first phase dictionary has only five thousand words which is significantly smaller than the necessary size, the vocabulary can be easily extended by simply enlarging the dictionary size without affecting the syllable recognition procedure or the entire system operation. This is why only a small experimental dictionary is used in the first phase research.

The recognition of the 418 first tone syllables in the first subsystem is in fact very difficult [10]. This is because the 418 first tone syllables consist of about 38 confusing sets, each of which has from about 4 to 19 confusing syllables [10]. 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 [10] that with standard approaches of the Dynamic Time Warping or Hidden Markov Models, LPC, and Itakura distance measures to recognize these syllables, the achievable recognition rates are as low as 60–70%. This tells how difficult it is to correctly recognize these syllables and why curently available techniques for English words cannot be applied directly. An initial/final two-phase recognition approach is thus specially designed [11] to recognize these very confusing syllables, whose block diagram is in Fig. 4. Here "final" means the vowel or diphthong parts of the syllable but including the medials and nasal ending (if any), and "initial" is the inital consonant of the syllable. Because of the open-syllabic structure, the ending point is easy to detect and the final of every syllable is relatively long and steady with a clear single peak in energy, we can therefore first detect the final part and recognize the final form a total of 38 different finals. Once the final is determined, we then try to recognize the initial preceding the final among at most 19 candidate initials. In this way the complicated problem of recognizing 418 very confusing syllables is reduced into two smaller problems, but the accurate recognition of the final is very important [11]. Although in this way segmentation of the syllable into two parts would very likely cause new problems, and different approaches in which the syllables are recognized as a whole are also under development currently, our results show that this initial/final two-phase recognition approach is the most attractive and feasible in terms of achievable recognition rates and computation complexity requirements.

IV. The Recognition of the Finals

The initial/final discrimination can first be performed by checking the periodicity of several successive frames [12]. For the final frames, there must be strong evidence of periodicity. For the initial frames except the initials [m], [n], [l], and [r], no periodicity can be observed. Spectrum characteristics are then needed to provide reliable initial/final discrimination for syllables beginning with the above four special initials.

Table I lists all the 38 possible finals for the 418 Mandarin syllables. The recognition of these finals is still very difficult because there are again many confusing final sets, such as {[a], [ia], [ua]}, {[ai], ⟨iai], [uai]}, etc., in which the medials [i], [u] are very difficult to recognize. In our research it was found that using multi-

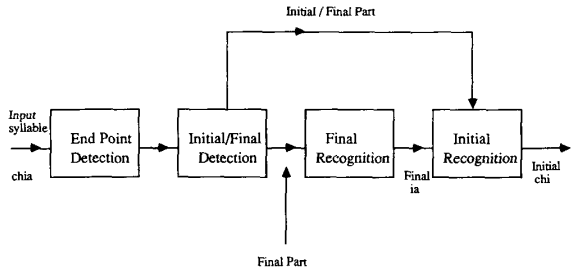


Fig. 4. The initial/final two-phase recognition scheme.

TABLE I
THE LIST OF ALL FINALS IN MANDARIN SYLLABLES

Finals		vowels or medials		
		i	u	iu
vowels	a	ia	ua	
	o	io	uo	
	e			
	eh	ieh		iueh
diphthongs	ai	iai	uai	
	ei		uei	
	au	iau		
	ou	iou		
nasal ending	an	ian	uan	iuan
	en	ien	uen	iuen
	ang	iang	uang	
	eng	ieng	ueng	iueng
special	er			
	$			

section vector quantization (MSVQ) techniques with branch-and-bound algorithm operated on the filter bank coefficients can give very good results while maintaining relatively simple hardware/software implementation [12]. The multisection vector quantization (MSVQ) technique is first proposed by Burton *et al.* [13]. In this approach, isolated words are recognized by means of sequences of VQ codebooks, called multisection codebooks. A separate multisection codebook is designed for each word in the vocabulary by dividing the word into equal-length sections and designing a standard VQ codebook for each section. Unknown words are classified by dividing them into corresponding sections, encoding them with the multisection code-books, and finding the multisection codebook that yields the minimum average distortion. Our final recognition system basically adopts the above concept.

The MSVQ recognition approach is demonstrated in Fig. 5. Let C_k, $k = 1, 2, \cdots, M$, represent the multisection codebook for the kth final, where k is the index for the candidate finals and M is the total number (38) of possible finals. Each multisection codebook C_k is composed of N section codebooks, each for one section of the final, i.e., $C_k = \{C_{kj}, j = 1, 2, \cdots, N\}$, where j is the index for sections, and N is the total number of sections for every final, and C_{kj} is the codebook for the jth section of the kth final. Each section codebook C_{kj} further consists of a total of n_{kj} codewords, i.e., $C_{kj} = \{c_{kji}, i = 1, 2, \cdots, n_{kj}\}$, where i is the index for codewords, and n_{kj} is the total number of codewords in the section codebook C_{kj}. Let $x_t, t = 1, 2, \cdots, Nm$ be the sequence of feature vectors for the frames of an unknown test final, Nm is the total number of frames. Apparently when the test final is divided into N equal-length sections, each section has exactly m frames. The distortion for the jth section with respect to the kth final is then

$$d_{kj} = \sum_{t=(j-1)m+1}^{jm} \left[\operatorname{Min}_i d(x_t, c_{kji})\right]$$

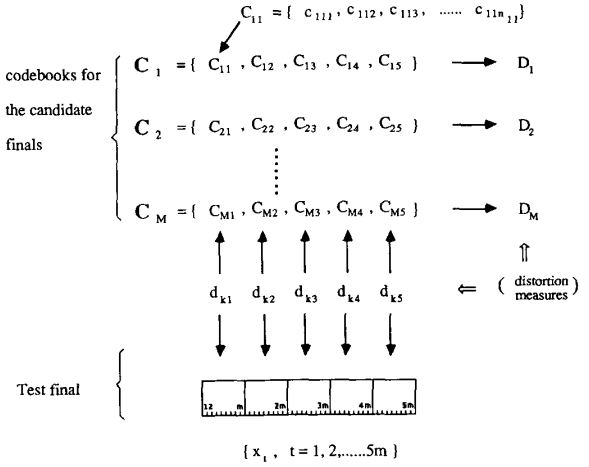


Fig. 5. The MSVQ recognition procedure for $N = 5$.

where $d(x, y)$ is the distance measure between two feature vectors x, y, and the average distortion with respect to the kth final is

$$D_k = \frac{1}{Nm} \sum_{j=1}^{N} d_{kj}.$$

The final with minimum average distortion with respect to the test final is then the recognition result. In this way, the section determination has the consequences of linear-time-warping which achieves gross time alignment. Fine time alignment is then achieved by codeword search. Since the section determination is only gross, this method is less robust than DTW if the words have many syllables. However, the MSVQ approach is in fact superior to the traditional DTW system for monosyllabic vocabulary such as Mandarin finals for the following two aspects [12]. First, time alignment is approximately done by section determination and codeword search, which is less time-consuming than dynamic programming based time alignment (DTW). Second, the MSVQ codebooks are composed of codewords which are more flexible than the word templates in most DTW systems, and thus storage saving can also be achieved. In fact, we have found that due to the monosyllabic structure and spectrum continuity properties of the Mandarin finals, the multisection recognition procedure will cause much less problem in Mandarin finals than English words.

A branch-and-bound classification algorithm is further developed for the search process as follows to improve the computation efficiency. During the classification process, each multisection codebook maintains an accumulated distortion value. The most promising path and the impossible paths are then defined according to the accumulated distortion. When the current most promising path is forwarded one step, i.e., one more frame is calculated, the decision about the next most promising path is made. When a certain path is ending (reaching the ending frame), the impossible paths are deleted and the surviving paths go on as the above strategy. The key points are how to find the most promising path and the impossible paths efficiently. This can be solved by maintaining a queue sorted by the current accumulated distortion value (from minimum to maximum) of each path. So the first element of the queue is for the most promising path. The paths behind the ending path are the impossible paths. The first path must be inserted into the queue after proceeding one step if it is no longer the most promising. This can also be done efficiently because the queue is sorted. Further improvements can be made when the common codebook concept is used in which common codebooks are trained for the first few sections of the finals having the same first phoneme such as [a], [ai], [au], [an], etc.

In our experiments, the speech signals are low-pass filtered at 4

kHz bandwidth, sampled at 10 kHz rate, and digitized in 16 bits. All the 418 possible Mandarin syllables in first tone are recorded. Two male speakers each uttered five randomized repetitions of these 418 isolated syllables in five sessions respectively. There is an interval of at least two days between every two sessions. These syllables are then segmented by locating the end points, and the final parts and initial parts are obtained separately for experiments. Because the total number of finals is 38, on the average more than 10 different syllables having the same final but different initials will contribute to the training of a given final. In every test, the finals obtained from one session (test session) are used as the test finals, and those from the other four sessions (training sessions) are used in training. After all the five sessions of data are used as test sessions, the averaged result is taken as the recognition result. In our experiments, it was found that feature vectors obtained from uniform filter bank with 16 filters, Euclidean distance measures, five equal-length sections for each final ($N = 5$), fixed-size codebook and clustering training algorithm together provide a recognition rate of 93.4 %, for which most parameters have been optimized, and this rate is almost identical to that of dynamic time warping, but with much less computation. Although the results for two speakers only here are not sufficient to determine the robustness as well as the achievable recognition rates, they at least serve as a rough indicator to tell the possible achievable performance of the approach.

In order to provide confusing finals for the second subsystem to correct the recognition errors in the syntactic level, an algorithm is designed such that when the distortion measures for the two most possible candidate finals are close enough, the output will be a first choice final and a second choice final. They will be combined with the possible choices of initials and possible choices of tones to obtain possible choices of syllables in addition to the first choice syllable. Some of the combinations will be automatically deleted, for example, the syllable [bia-4] does not exist in Mandarin.

V. The Recognition of the Initials

All the 21 possible initials of Mandarin syllables are listed in Table II. The recognition of these initials is even more difficult then the finals because the initials are relatively short, unstable, and confusing. In our research, it was found that many well-known approaches cannot yield very good recognition results even with complicated algorithms, for example, the continuous hidden Markov models [14]. We found that finite-state vector quantization (FSVQ) seems to be one of the best approaches to recognize the Mandarin initials [15] when both the achievable recognition rates and computation complexity requirements are considered. A finite-state vector quantizer (VQr) is a trellis structured vector encoder, as was illustrated in Fig. 6. It consists of a set of VQr's which are switched by some well trained transition functions. The encoding process includes a full-search comparison stage and a next-VQr-prediction stage alternatively. It can be specified by a finite state space S, an initial state s_0, and three functions: 1) an encoder α: $A \times S \rightarrow B$ where A denotes the observation space and B is a finite set of all encoded symbols; 2) a decoder β: $S \times B \rightarrow \hat{A}$ where $\hat{A}$ is the reproduction space; 3) a next state function or transition function f: $S \times B \rightarrow S$. They operate as follows, where x_t is the feature vector for the initial frames at the time t, u_t is the encoded symbol, and s_t is the state at time t:

$$\text{encoder} \begin{cases} u_t = \alpha(x_t, s_t) & u_t \in B,\ x_t \in A \\ s_{t+1} = f(s_t, u_t) & s_t, s_{t+1} \in S \end{cases}$$

$$\text{decoder} \begin{cases} \hat{x}_t = \beta(s_t, u_t) & \hat{x}_t \in \hat{A}. \end{cases}$$

Apparently the point is that the next $\mathrm{VQr}(s_{t+1})$ to be used is a function of the current $\mathrm{VQr}(s_t)$ being used and the corresponding encoded symbol (u_t). The encoder α keeps a copy of the decoder β and selects the encoded symbol u_t by minimum distortion rule:

$$u_t = \alpha(x_t, s_t) = \arg\min_{u \in B} d(x_t, \beta(s_t, u)).$$

TABLE II
Classification of Mandarin Consonants According to Their Places and Manners of Articulation

manner / place	liquid	nasal	plosive		affricate		fricative	
			unas	as	unas	as	unv	voi
labial		m	b	p				
labiodental							f	
front part of tongue tip					tz	ts	s	
tongue tip	l	n	d	t				
back part of tongue tip (retroflex)					jr	chr	shr	r
alveolar palatal					ji	chi	shi	
velar			g	k			h	

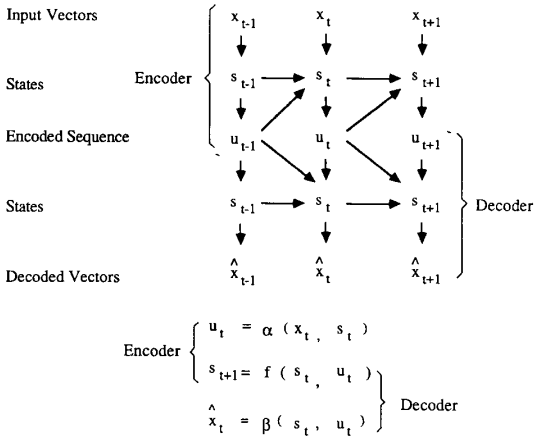
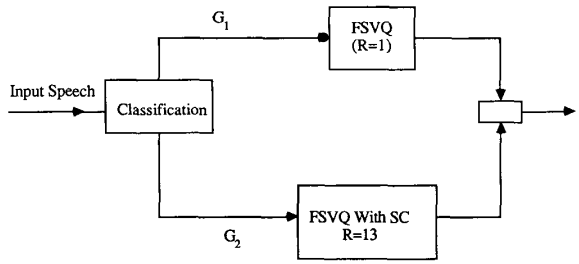


Fig. 6. The concept of finite state vector quantization (FSVQ).

This is demonstrated in Fig. 6. Just as done for finals previously, FSVQ can be applied easily in initial recognition. FSVQr's can be trained for all candidate initials. The test initial is then encoded by these FSVQr's and the average distortion are evaluated. The candidate giving minimum distortion will be the recognized initial. Experiments discussed later show that in this way the switching among many possible VQr's by well-trained transition functions will minimize the distortion between the test initial and the correct candidates, thus improves the recognition performance [15]. The computation required for this algorithm, however, is on the same order as MSVQ previously discussed for the recognition of finals and much less than Dynamic Time Warping.

Our initial recognition approach using FSVQ has a block diagram shown in Fig. 7. The experiments show that such a structure gives the highest recognition rate. The test initial is first classified into two categories, G_1 and G_2, where G_1 = {[b], [d], [g], [m], [n], [l], [r], [?]} and G_2 = {[p], [f], [t], [k], [h], [jr], [chr], [shr], [tz], [s], [ji], [chi], [shi]}. [?] in G_1 represents the initial part for syllables without a consonant in the beginning. The classification is performed by evaluation of a classification function based on some simple time domain features such as average energy, average vibration count, zero-crossing rate, and the length of the initial. It was found that the two categories of initials can be recognized using slightly different algorithms such that the overall recognition results can be optimized. For the category G_1, the standard FSVQ described above can be used for recognition. For the category G_2, on the other hand, a modified FSVQ algorithm should be used, in



G_1 = { [b], [d], [g], [m], [n], [l], [r], [?]}

G_2 = { [p], [f], [t], [k], [h], [jr], [chr], [shr], [tz], [ts], [s], [ji], [chi], [shi]}

Fig. 7. The block diagram of the initial recognition algorithm.

which the present state is a function of the state and encoded symbol R frames earlier instead of simply one frame earlier, i.e.,

$$s_{t+R} = f(s_t, u_t), \qquad R \neq 1$$

and simulations indicate that $R = 13$ gives the highest recognition rate.

In the experiments, the initials are taken from exactly the same database for the 418 syllables as described previously. Because the total number of initials is 22 (21 initials plus [?]), on the average about 20 syllables with the same initial but different finals will contribute to the training of an initial. The overall recognition rate is on the order of 93.6% for two speakers only, which is the highest result we have ever obtained for the Mandarin initials.

VI. The Recognition of the Tones and the Combined Result for the Syllables

Although there are only four different tones, the correct recognition of the tone is in fact difficult. One example problem is the confusion caused by the third and the fourth tones. In most of the cases, only the first half of the third tone will be pronounced, which will be very close to a fourth tone. A good example is shown in Fig. 8, where the two curves are the pitch frequencies (the inverse of the pitch period as used in Fig. 1) plotted as functions of time for two syllables [iu-3] and [iu-4]. It can be seen that the basic shapes are very close, the only difference is in the dynamic range and the slope. We use here the tone recognition method developed by J.-C. Lee [16], in which the sum and difference of the log pitch frequencies for adjacent frames are taken as feature parameters,

$$v_1 = p_{i+1} + p_i$$

$$v_2 = p_{i+1} - p_i$$

where p_i is the log pitch frequency for the i-th frame. In fact these two parameters represent the average pitch level and the pitch slope, respectively. These two feature parameters are then used to form a two-dimensional vector, $v = [v_1, v_2]$, and vector quantization (VQ) codebooks with codebook size 16 for the four tones are trained using the weighted mean square error as the distortion measure, in which the inverse of covariance matrix is taken as the weighting function. Three-state, left-to-right hidden Markov models (HMM) are then developed based on these vectors and used to recognize the tones [16]. The above system parameters have been optimized with respect to a database in which eight males and eight females each uttered about a hundred different syllables, each with four different tones, and it was found that the average recognition rate for the tones is 98.3% for speaker dependent case, and the algorithm can be easily implemented on a personal computer.

Combining the results for the final, initial and tone recognition, the total recognition rate for the syllables is only on the order of

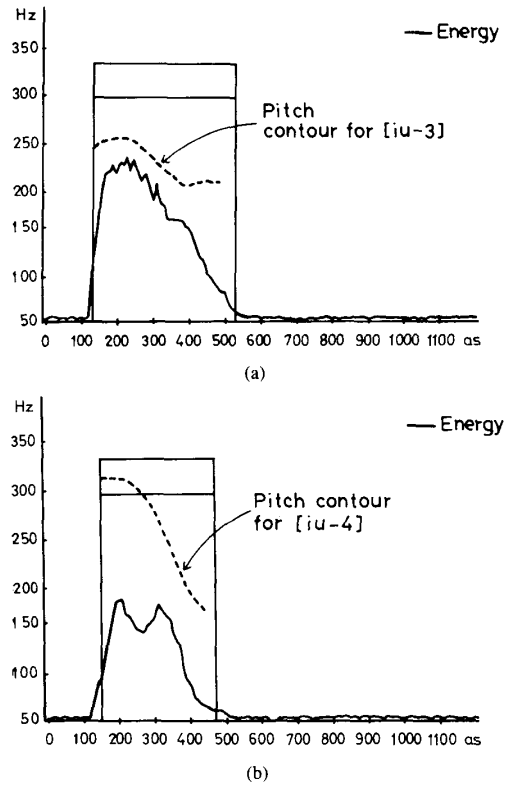


Fig. 8. The pitch contours for (a) a third tone with only the first half and (b) a fourth tone.

85.9%. This number is relatively low. However, considering the fact that the Mandarin syllables are highly confusing and the recognition rate of 60–70% for the standard approaches of dynamic time warping or hidden Markov models, LPC and Itakura distance measures [10], very significant improvements have been made and the results are rather satisfactory.

VII. Formation of Word Hypotheses from Syllables

This is the next level operation in our recognition hierarchy, or the first part of the second subsystem. It transforms the series of input syllables into corresponding characters in forms of possible monosyllabic and multisyllabic word hypotheses. Its operation will be described here using the previously mentioned example. The output characters and word hypotheses for the example, [ni-3] [shr-4] [i-2] [jia-4] [huei-4] [tieng-1] [guo-2] [iu-3] [de-5] [dian-4] [nan-3], is shown in Fig. 9. First, all multisyllabic word hypotheses which can be found by matching with the words in the dictionary will be obtained, such as the word hypotheses "適宜 (Suitable, [shr-4] [i-2])," "移駕 (Move, [i-2] [jia-4])," "國語 (Mandarin, [guo-2] [iu-3])," "電腦 (Computer, [dian-4] [nau-3])." Note that the first two are incorrect word hypotheses, but the last two are desired words. Secondly, all syllables which cannot form multisyllabic word hypotheses with adjacent syllables on its left or right must correspond to monosyllabic word hypotheses. But usually more than one monosyllabic words can share the same syllable, therefore only the most frequently used and second frequently used (if any) monosyllable word hypotheses will be found from the dictionary and given, such as "你 (You)" and "擬 (Plan)" for [ni-3] "會 (Can)" for [huei-4], and "聽 (Listen to)" and "廳 (Hall)" for [tieng-1]. Thirdly, if some of the syllables in the multisyllabic word hypotheses obtained in the first step corre-

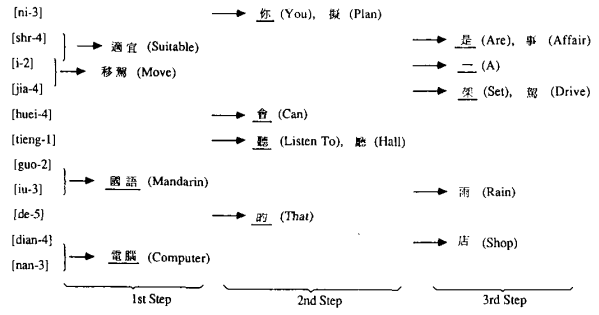


Fig. 9. The steps in word hypothesis formation from syllables. Only the underlined words are correct.

spond to very frequently used monosyllabic words, such as "是 (Are)" and "事 (Affair)" for [Shr-4], "一 (A)" for [i-2], "架 (Set)" and "駕 (Drive)" for [jia-4] in the multisyllabic word hypotheses "適宜 (Suitable, [Shr-4] [i-2])" and "移駕 (Move, [i-2] [jia-4])" in the above example, these very frequently used monosyllabic word hypotheses are also given. In other words, the operation tries to provide all possible multisyllabic and monosyllabic word hypotheses which can be found in the dictionary.

In the dictionary, the words are organized in a special structure called word trees such that all words (monosyllabic and multisyllabic) having the same first syllable will be organized in the same tree, and all words in the same tree having the same second syllable will be organized under the same node, etc. Such a structure will improve the searching speed in matching processes. In order to reduce the total number of words which have to be stored in the dictionary, some additional word formation rules are designed to generate new words from given words in the dictionary [17]. For example, a verb followed by the character "過" ([guo-4]) simply represent the past tense, such as "吃 (Eat, [chr-1])" and "過" form "吃過 (Ate, [chr-1] [guo-4]), " "看 (See, [kan-4]), " and "過" form "看過 (Saw, [kan-4] [guo-4]), " etc. Also, a very large number of nouns can be formed by combining two other nouns. For example, the nouns "豬 (Pig, [ju-1])" and "肉 (Meat, [rou-4])" form a new noun "豬肉 (Pork, [ju-1] [rou-4])" based on very simple general rules, etc. According to these word formation rules, it is estimated that the total number of commonly used Chinese words which have to be stored in the dictionary can be reduced from about 80 thousand to about 35 thousand.

Some additional word deletion rules are also designed such that redundant or impossible word hypotheses can be automatically deleted without going to the syntactic level [17]. One example rule is as follows. Any shorter word hypotheses within a longer word hypothesis should be deleted if they do not extend across the boundary of the longer word hypothesis. For example, the shorter word hypotheses "台灣 (Taiwan, [tai-2] [wan-1])" and "大學 University, [da-4] [shiue-2])" will both be deleted if the longer word hypothesis "台灣大學 (Taiwan University, [tai-2] [wan-1] [da-4] [shiue-2])" exists and only this longer word hypothesis will be left. Another example rule tells that all word hypotheses which cannot be connected by other word hypotheses on both ends (except in the sentence beginning and ending positions) should be deleted. For example, for the four syllables [jueng-1] [wen-2] [shu-1] [ru-4], the word hypotheses "中文 (Chinese Character, [jueng-1] [wen-2]), " "輸入 (Input, [shu-1] [ru-4]), " "中 (Center, [jueng-1]), " "文書 (Documents, [wen-2] [shu-1])" will all be formed, but only the first two will be left and the last two will be deleted because the syllable [ru-4] cannot represent a monosyllabic word in today's Chinese sentence and therefore the word hypotheses "中 (Center)" and "文書 (Documents)" cannot find any word hypothesis to be connected to the right, thus should be deleted.

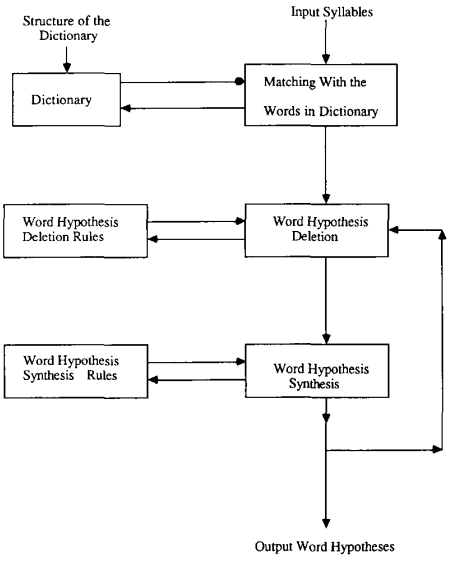


Fig. 10. The complete word hypothesis formation processes.

The above word hypothesis formation process is summarized in Fig. 10, which has been successfully implemented on a VAX minicomputer using Franz Lisp. The primary problem is that to build a large enough dictionary is a very time consuming task. Therefore our first phase dictionary consists of 5 thousand words only, which practically limits the possibility of a complete and rigorous test, but the feasibility to transform syllables into word hypotheses is completely demonstrated as discussed earlier.

VIII. Syntactic Analysis to Identify Correct Words

This is the highest level operation in our recognition hierarchy, or the second part of the second subsystem. The input is the set of all possible monosyllabic and multisyllabic word hypotheses corresponding to the recognized syllables. The goal is to find the only correct combination of these word hypotheses which form a grammatically valid sentence. The approach is simply to take all possible combinations of the word hypotheses and parse them one by one based on a Chinese grammar [18], [19]. The ungrammatical combinations will be ruled out and only the correct solution will be left and shown on the screen. There are a few cases in which two combinations are both grammatically valid, or none of the combinations are valid. In both cases the output will be either a random choice or a most probable choice (for example, with the most frequently used monosyllabic word hypothesis).

A syntactic analysis system for Chinese sentences (SASC) is employed here to parse the possible sentences and rule out ungrammatical combinations [20], [21]. In the SASC system, Chinese sentences are syntactically analyzed from the viewpoints of generative grammar. It uses a bottom-up parser instead of a top-down parser, because the former tends to be more efficient for Chinese sentence analysis. The parser uses charts as global working structures, because many natural language processing systems have proved the chart to be an efficient data structure to record what has been done so far in the course of parsing. A parser based on charts can avoid the inefficiency in duplicating many computations that a top-down parser often suffers when backtracking occurs. The parser parses sentences in such a way that phrases are built up on the chart by starting with their heads and adjoining constituents on the left or right of the heads. For example, according to the phrase structure rule (PSR), "NP → QP N," N (noun) is the head of NP (Noun Phrase). When encountering a noun, the parser will try to build an

NP by starting with the noun and adjoining the preceding quantity phrase (QP). According to the PSR, "VP → V-n NP," V-n (transitive verb) is the head of VP (Verb Phrase). When encountering a transitive verb, the parser's action is similar except that it tries to adjoin the following NP as its object. But if its following NP is not yet parsed by the parser, the expectation to build a VP is suspended until an NP is built up in the object position. The parser using the above algorithm constructs syntax trees of input sentences exactly from bottom to top. The algorithm used seems to be a good combination of data-driven parsing and hypothesis-driven parsing. The grammar used is a modification of ATN grammar, which is made up of states classified into HEADS (Head State) and STATE. A state indicates a stage of parsing and describes the information and actions needed to deal with the current situation. A list of some fundamental PSR's for Chinese sentences implemented in our system is shown in Fig. 11.

In order to solve the complicated syntactic phenomena such as passivization, relativization, topicalization, and Ba-transformation in Chinese sentences, a raise-bind mechanism based upon the linguistic theory of empty categories is developed [21]. In this way, the SASC parser will treat the above different transformations very easily in the same way, i.e., if an empty element is inserted into the right position, the syntactic structure will become easy. Fig. 12 lists some typical examples of Chinese sentences with empty categories to be used in analysis. In sentences (1)–(6) the solid lines indicate the antecedant of each empty category. The notation [s · · ·] denotes the presence of a clause. The missing element in sentence (7), however, does not refer to any element within the sentence. For these sentences, the parser first generates an empty NP inserted into the vacant position where an NP is expected to appear. Then the empty NP will be raised up in some way along the parsing tree, when the tree is growing up (recall that the parser works bottom-up), until its antecedant is parsed. At this point, the parser binds the empty NP by setting it to refer to its antecedant. Once being bound, the empty NP will not be raised any further. This is because an empty NP has exactly one antecedant and cannot be bound more than one time. Using this approach, many syntactically valid sentences can be easily parsed.

The SASC has been successfully implemented on a VAX minicomputer using Franz Lisp. The parsing operation can be performed with very high speed. In average it takes only 15 ~ 20 ms per word to parse the sentences. Due to the limited vocabulary of the dictionary implemented in the system as discussed previously, only preliminary tests can be performed. However, the results show that at least 92–95% of sentences in Chinese textbooks of primary schools in this country can be successfully analyzed by the system, if the words used are already in or keyed into the dictionary.

(1) S = bar → S-bar PRTAG | S PRTAG

(2) S-bar → Topic S

(3) S → (NP) VP

(4) NP → (XPDE) (QP) (ADJ) N |
(QP) (XPDE) (ADJ) N | NP Localizer

(5) XPDE → S 的 | NP 的 | PP 的

(6) VP → (AUX | ADV | PP | NP)* V-bar

(7) V-bar → V QP | V 得 S | V 得 ADV |
V (NP) (NP | PP | VP | S | S-bar)

(8) PP → PREP NP

Fig. 11. The list of some fundamental PSR's for Chinese language used in our system.

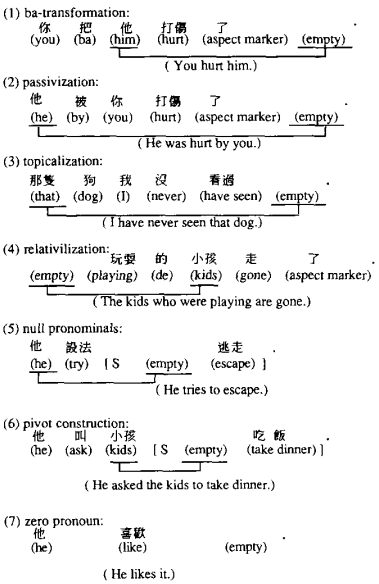


Fig. 12. Some typical examples of Chinese sentences with empty categories to be used in analysis.

IX. Preliminary Test Results and Concluding Remarks

All the different operations described above have been successfully implemented individually. The recognition of finals, initials, and tones of syllables are separately implemented on three IBM personal computers with additional TMS 320 signal processing boards. They all work in real-time (the recognition is completed in 0.2 s), but the recognition rates are slightly worse than the software simulation results described previously. The word hypothesis formation and syntactic analysis operations, on the other hand, are implemented in a VAX 11/785 minicomputer. They work well except for problems with a relatively small dictionary which limits the scope of practical tests. Although the integration of different parts into a single system is currently in progress, these individual operations are connected to perform some preliminary dictation tests. The text for the test sentences is taken from everyday newspapers and read in isolated syllables by the same speaker as in the training of the syllable recognition systems. New words have to be keyed into the dictionary if they are not in the dictionary already. The syllables together with at most two other confusing syllables (if any) can be recognized in real-time and sent to the VAX computer for higher level recognition. The problem with higher level recognition is that very often large numbers of word hypotheses give hundreds of possible combinations for syntactic checking even in a relatively simple sentence. It is therefore hard to tell now the average speed in dictating a sentence. Apparently, a more efficient algorithm is needed to direct the system toward most probable word hypothesis combinations instead of simply analyzing all of them exhaustively. A total of 600 sentences are dictated in the first phase test. The rate for syllable recognition (first choice) is on the average 81.6% which is 4.3% lower than the simulation result, but the final rate for dictation (i.e., the correction rate for the characters) is on the order of 88–90% depending on the text of the sentences. In other words, some errors made in the acoustic level can be finally corrected in the syntactic level; this is the result of our hierarchical language recognition approach, and the number is close to the first phase goal of the research as described in the beginning. Further improvements should be made on each part of the system and there is still a very long way to go before a real-time, integrated

prototype system can actually be implemented. However, it has been demonstrated clearly that a Mandarin dictation machine will be not only practically attractive, but technically achievable; and this is definitely one possible solution to the problem of input of Chinese characters into computers.

References

[1] *Proc. 1986 and 1987 Int. Conf. Chinese Computing*, Chinese Lang. Comput. Soc., Singapore, Aug. 1986, Chicago, IL, June 1987.

[2] A. Averbuch *et al.*, "An IBM PC based large-vocabulary isolated-word recognizer," in *Proc. 1986 Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Tokyo, Japan, Apr. 1986, pp. 53-56.

[3] M. Picheny *et al.*, "A real-time IBM PC based large-vocabulary isolated-word speech recognizer," Pinner, England, Voice Processing, Online Publ., 1986.

[4] A. M. Derouault, "Context-dependent Markov models for large-vocabulary speech recognition," in *Proc. 1987 Int. Conf. Acoustics, Speech, Signal Processing*, Dallas, TX, Apr. 1987.

[5] B. Merialdo, "Speech recognition with very large size dictionary," in *Proc. 1987 Int. Conf. Acoustics, Speech, Signal Processing*, Dallas, TX, Apr. 1987, pp. 364-367.

[6] L.-S. Lee, C.-Y. Tseng, K. J. Chen, and J. Huang, "The preliminary results for a Mandarin dictation machine based upon Chinese natural language analysis," in *Proc. 1987 Int. Joint Conf. Artificial Intelligence*, Milano, Italy, Aug. 1987.

[7] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press, 1968.

[8] S.-M. Lei and L.-S. Lee, "Digital synthesis of Mandarin speech using its special characteristic," *J. Chinese Inst. Eng.*, vol. 6, no. 2, pp. 107-115, Mar. 1983.

[9] V. A. Fromkin, *Tone—A Linguistic Survey*. New York: Academic, 1978.

[10] H.-y. Gu, C.-y. Tseng, and L.-s. Lee, "A comparative study on the performance of several speech recognition techniques applied on the highly confusing Mandarin syllables," *J. Chinese Inst. Eng.*, 1989.

[11] M.-S. Yu, G.-S. Chen, C.-C. Hsiao, C.-Y. Tseng, and L.-S. Lee, "A preliminary approach to complete vocabulary Mandarin syllable recognition," in *Proc. 1986 Int. Conf. Chinese Computing*, Singapore, Aug. 1986, pp. 168-171.

[12] C.-W. Hwang, C.-Y. Tseng, and L.-S. Lee, "An efficient Mandarin vowel recognition system based upon multi-section vector quentization and branch-and-bound classification techniques," in *Proc. 1986 Int. Computer Symp.*, Tainan, Taiwan, Rep. of China, Dec. 1986.

[13] D. K. Burton, J. E. Shore, and J. T. Buck, "Isolated-word speech recognition using multi-section vector quantization codebooks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 4, Aug. 1985.

[14] C.-H. Wu, C.-Y. Tseng, and L.-S. Lee, "New speech recognition approaches for the Mandarin consonants based upon hidden Markov models," in *Proc. 1987 Nat. Computer Symp.*, Taipei, Taiwan, Rep. of China, Dec. 1987, pp. 971-979.

[15] P.-Y. Ting, C.-Y. Tseng, and L.-S. Lee, "New speech recognition approaches based upon finite state vector quantization with structural constraints," in *Proc. 1988 Int. Conf. Acoustics, Speech, Signal Processing*, New York, Apr. 1988.

[16] J.-C. Lee, "Mandarin lexical tone recognition based on vector quantization and hidden Macov models," Master's thesis, Tsing-Hua Univ., Taiwan, Rep. of China, May 1986.

[17] C.-G. Chen, K.-J. Chen, and L.-S. Lee, "A model for lexical analysis and parsing of Chinese sentences," in *Proc. 1986 Int. Conf. Chinese Computing*, Singapore, Aug. 1986, pp. 33-40.

[18] C. Li and S. Thompson, *Mandarin Chinese: A Functional Reference Grammar*. Berkeley, CA: University of California Press, 1981.

[19] J. Huang, "Logical relations in Chinese and the theory of grammar," Ph.D. dissertation, Massachusetts Inst. Technol., 1982.

[20] L.-J. Lin, K.-J. Chen, J. Huang, and L.-S. Lee, "SASC: A syntactic analysis system for Chinese sentences," in *Proc. 1986 Int. Conf. Chinese Computing*, Singapore, Aug. 1986, pp. 29-32.

[21] L.-J. Lin, J. Huang, K.-J. Chen, and L.-S. Lee, "A Chinese natural language processing system based upon the theory of empty categories," in *Proc. Fifth Nat. Conf. Artificial Intelligence*, AAAI, Philadelphia, PA, Aug. 1986, pp. 1059-1062.

On-Line Recognition of Handwritten Arabic Characters

SAMIR AL-EMAMI AND MIKE USHER

***Abstract*—Arabic characters are always in cursive script. Handwritten words were entered into an IBM PC via a graphics tablet and a segmentation process applied to the points; the length and the slope of each segment was then found, and the slope categorized to one of four directions. In the learning process, specifications of the strokes of each character are fed to the computer. In the recognition process, the parameters of each stroke are found and special rules applied to select the collection of strokes which best matches the features of one of the stored characters. The results are promising, and suggestions for improvements leading to 100% recognition are proposed.**

***Index Terms*—Arabic characters, cursive script, handwritten characters, segmentation, structure analysis, trees.**

I. Introduction

Much work has been done on the recognition of English characters, both separated and in cursive script. There has been a little research on Chinese, Indian, Korean, and some other characters, but that on Arabic characters is very limited [1], [2]. This field is important, not only for Arabic speaking countries, but also for Curds, Persians, and Aurdo-speaking Indians, who have similar character sets.

Arabic characters are always in cursive script. The language comprises 28 main characters and is written from right to left. Most characters have four versions, depending on their position within a word, which is a group of joined or separated characters. A set of Arabic characters is shown in Fig. 1(a) and some Arabic words in Fig. 1(b).

The four versions of characters are the beginning, median, termination, and separated types. The beginning version is always connected to another character at its tail (زقاق), the median to characters at both sides (محله) and the termination to another at its start (حي); otherwise they are regarded as separated versions (د ا). The median versions are usually similar to the beginning, except for the connecting stroke but there are some which differ in the two positions [e.g., beginning (عـ), median (ـعـ)]. There are a small number which have the same shape in any position [except for the connecting stroke (ـل)] and there are others which can have different shapes in one position depending on the phonetics of the word (e.g., ك ، ى ، د ، ؤ ، إ ، أ).

The widths and lengths of characters differ from character to character and from one version to another, making recognition techniques difficult. Many characters are composed of two parts, the body of the character and a number of complementary dots, either above or below the body. There may be one dot or a group of two or three.

A feature of our work is that the various versions of one character are considered as different characters, as is the case in the Arabic typewriter, which has separate keys for the various versions. The characters chosen for the experiment are the building units of four words (حي، محله ، دار، زقاق), which are used as a

Manuscript received November 2, 1987; revised January 20, 1989. Recommended for acceptance by C. Y. Suen.

The authors are with the Department of Cybernetics, University of Reading, Reading RG6 2AL, England.

IEEE Log Number 9036116.