S2.7

A REAL-TIME MANDARIN DICTATION MACHINE FOR CHINESE LANGUAGE WITH UNLIMITED TEXTS AND VERY LARGE VOCABULARY

Lin-shan Lee[1], *Chiu-yu Tseng*[3], *Hung-yan Gu*[1],

F. H. Liu[2], *C. H. Chang*[2], *S. H. Hsieh*[2], *C. H. Chen*[2]

[1] Department of Computer Science and Information Engineering, National Taiwan University,
[2] Department of Electrical Engineering, National Taiwan University,
[3] Institute of History and Philology, Academic Sinica,
Taipei, Taiwan, Republic of China

ABSTRACT

This paper describes the first successfully implemented real-time Mandarin dictation machine developed in the world which recognizes Mandarin speech with unlimited texts and very large vocabulary for the input of Chinese characters to computers. Isolated syllables including the tones are first recognized using specially trained hidden Markov models with special feature parameters, the exact characters are then identified from the syllables using a Markov Chinese language model, because every syllable can represent many different homonym characters. The real-time implementation is on an IBM PC/AT, connected to a set of special hardware boards on which ten TMS 320C25 chips operate in parallel. It takes only 0.45 sec to dictate a characters.

I. Introduction

Today, the input of Chinese characters into computers is still a very difficult and unsolved problem. All the currently existing input methods either are too slow or need special training, therefore can't be conveniently used by most people. This is the basic motivation for the development of a Mandarin dictation machine. We defined the scope of this 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). The machine is speaker dependent. The first stage goal of this system is to have 90% correction only for the sentences in the Chinese textbooks of the primary schools in Taiwan, Rep. of China, because the errors can be found by the user on the screen and corrected from the keyboard very easily using convenient software system. Such a performance is still much more efficient than any of the currently existing input systems. However, on the other hand, the machine has to be able to recognize Mandarin speech with very large vocabulary (at least the 15 thousands of commonly used Chinese characters and the 60 thousands of commonly used Chinese words have to be covered) and unlimited texts (at least for sentence structures appearing in primary school text books) because we assume the input to computers can be arbitrary Chinese texts. Also, the machine has to work in real-time for computer input applications. As will be shown later in this paper, the above goals are almost achieved in this research. This is the first successfully implemented real-time Mandarin dictation machine developed in the world for very large vocabulary and unlimited texts.

II. Considerations for the special structure of Chinese Language and the overall system structure

There are at least 60 thousands of commonly used words in Chinese. Therefore the words can not be used as the dictation units. There are at least 15 thousands of commonly Chinese characters, each character is mono-syllabic. A nice feature is that the total number of different syllables in Mandarin speech is only about 1300. If we use the 1300 syllables as the dictation units, all the words or characters will be covered. However, the small number of syllables implies another difficult problem, that is, many different homonym characters will share the same syllable. This problem will be solved later in this paper using a specially designed Markov Chinese language model. Based on the above observations on the special structure of Chinese language, the use of syllable as the dictation unit becomes a very natural choice.

Another very special important feature of Mandarin Chinese language is the lexical tones for the syllables. Mandarin Chinese is a tonal language. Every character is assigned a tone in general. There are basically five different tones. It has been shown that the primary difference for the tones is in the pitch contours, and the tones are essentially independent of the other acoustic properties of the syllables. If the differences among the syllables due to lexical tones are disregarded, only 409 syllables are required to represent all the pronunciations for Mandarin Chinese. This means the recognition of the syllables can be divided into two parallel procedures, the recognition of the tones, and of the 409 syllables disregarding the tones.

Based on the considerations described above, the overall system structure for the Mandarin dictation machine is shown in Fig. 1. The

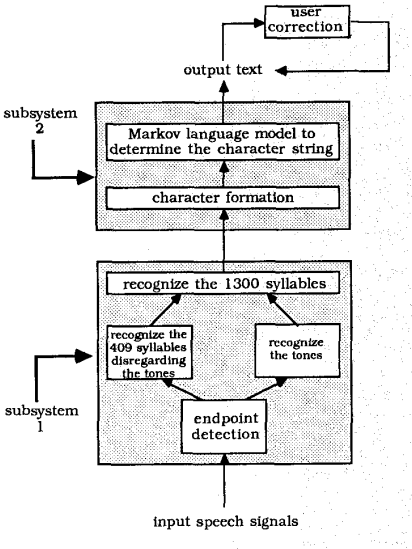


Fig. 1 The overall system structure for the Mandarin dictation machine

system is basically divided into two subsystems. The first is to recognize the syllables, and the second is to transform the series of syllables into the characters. For the first subsystem of syllable recognition, the corresponding syllable (disregarding the tones) and the tone are then recognized independently in parallel. Because errors always happen, we therefore have to provide information for confusing syllables, and confusing tones. For the second subsystem we need to first obtain all possible character hypothesis from a stored dictionary and then use the Markov Chinese language model to select the most probable (maximum likelihood) concatenation of them as the output sentence. All these processes will be described in detail in the following.

CH2847-2/90/0000-0065 $1.00 © 1990 IEEE

III. Syllable Recognition Disregarding the Tones

The recognition of the 409 Mandarin syllables disregarding the tones is very difficult because there exist 38 confusing sets in this vocabulary, each of which contains at most 19 very confusing syllables. Conventionally each Mandarin syllable is decomposed into an "initial/final" format, in which "initial" means the initial consonant of the syllable, while "final" includes the vowel or diphthong part but including possible medial or nasal ending. Each confusing set mentioned above then consists of syllables sharing the same final but with different initials. Direct application of standard approaches of hidden Markov models (HMM's) gives recognition rates on the order of only 70%-80% due to the difficulties caused by these confusing sets, as can be seen in the first two rows in Table 1, where top n rates are the rates for which the correct syllable is among the top n candidates chosen in the recognition phase. In the following, various approaches are developed to improve the recognition rates, and any of them can be used in the dictation machine because only the parameters for the HMM's should be modified.

A two-pass training approach is first developed by specially considering the characteristics of the vocabulary. Because all the syllables in a confusing set share the same final, in this approach 38 final HMM's are first trained in the first pass, each for the final of a confusing set, using the segmented final parts of the training utterances, and 409 initial HMM's are then trained using the segmented initial parts in the second pass. These HMM's are finally smoothly cascaded to form 409 syllable HMM's by requiring that in each syllable HMM the last state of the initial HMM is exactly the first state of the final HMM which was trained primarily from the transition region of the speech signal. In this way not only the initial HMM's and final HMM's can be separately trained and the short initial parts can be assigned more number of states, but the HMM's for the syllables in a given confusing set will have exactly identical parameters in the last few states, thus the effect of the final in the recognition phase can be minimized while the differences in initials can be emphasized to better distinguish these syllables. The results of this approach for continuous HMM's can be seen in the third row of Table 1, where top 1 rate is improved by more than 10%, while the top 3 rate now exceeds 99%.

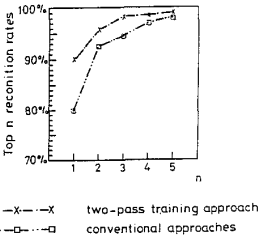
Because the errors in the above two-pass training approach are primarily caused by errors in distinguishing the initials, it is believed that the limited number of training utterances (five for each syllable in the experiments here) make the 409 initial HMM's less robust. Considering the fact that very often quite a few finals approximately start with some common phoneme (for example, a, ai, au, an, ang all start with the phoneme a), syllables with these finals but the same initial (such as sa, sai, sau, san, sang) can in fact share the same initial HMM. In this way, the total number of initial HMM's can be reduced from 409 to 99, and the number of training utterances for each initial HMM's can be significantly increased (4 times in average). This is the revised two-pass training approach. The results in the fourth row of Table 1 show that in this way the top 1 recognition rate is in fact slightly degraded, while the top 2, 3, 4, 5 rates are all improved, probably because the initial HMM's thus obtained is really relatively less accurate, although more robust. A three--pass training approach is therefore further developed, in which the 99 initial HMM's created in the second pass of the revised two-pass training is now taken as the initial values, they then go through a third training process such that eventually 409 instead of 99 initial HMM's are obtained and cascaded with the 38 final HMM's to form 409 syllable HMM's. The results in the fifth row of Table 1 indicate that the top 1 rate is improved to 92% while the high top 2, 3, 4, 5 rates are preserved, i.e., the advantages of the above two approaches, the accuracy and robustness, are now combined. Note that in the dictation machine, the Markov Chinese language model can possibly correct some of the errors made in the syllable recognition, therefore high top 2, 3, 4, 5 rates are also helpful and desired.

All the above approaches require segmented initial and final parts of training utterances to be used in the training of HMM's, but actually automatic algorithm for segmentation between initial and final parts is very difficult, and manual help is currently needed to avoid errors. Therefore a revised three-pass training approach which requires a minimum number of segmented training utterances is further developed.

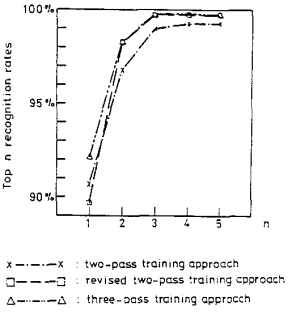
Table 1 The rates for recognition of the 409 Mandarin syllables disregarding the tones based on HMM's.

		top 1	top 2	top 3	top 4	top 5
Standard Approaches	discrete HMM's	70.82%	87.64%	92.17%	93.88%	94.83%
	continuous HMM's	79.85%	92.54%	94.53%	97.01%	98.01%
New Approaches	two-pass training	90.69%	96.81%	99.02%	99.26%	99.26%
	revised two-pass training	89.71%	98.28%	99.75%	99.75%	99.75%
	three-pass training	92.16%	98.28%	99.75%	99.75%	99.75%
	revised three-pass training	91.42%	96.81%	99.02%	99.51%	100.00%

In this approach, only one set of training utterances needs to be segmented, and they are used to train 38 final HMM's and 99 initial HMM's to be cascaded to form the 409 syllable HMM's. These 409 HMM's are then taken as the initial values to go through a third pass training, in which the unsegmented training utterances are used in the forward--backward algorithm, while the parameters of the initial and final parts of the syllable HMM's are reestimated separately. The results in the last row of Table 1 show that in this way most of the training utterances need not be segmented, but at the price that the top 1, 2, 3, 4, rates are all slightly degraded. Some typical recognition rates in Table 1 are also plotted in Fig. 2 for easy comparison.



(a) Two-pass training vs standard continuous HMM's



(b) Comparison among the three new approaches.

Fig. 2 Comparison among the recognition rates for the 409 syllables disregarding the tones

IV. Tone Recognition

There are totally five different lexical tones in Mandarin usually referred as the the first, second, third, fourth and the fifth tones, among them the fifth (or neutral) tone is the most difficult to distinguish. This is because there exist typical pitch contours for the pitch frequencies of the first four tones in general. Some examples are shown in Fig. 3. However, the pitch frequencies of the fifth tone do not necessarily follow any specific pattern. Some initial efforts on Mandarin lexical tone recognition had been reported with very encouraging results[9,10], but they all concentrated on the recognition of the first four tones while the fifth tone was always ignored. However, the occurrence frequency of the fifth tone in everyday Mandarin is in fact not negligible., i.e. 7%, and most of them are function words such as "的", "了" and "着" with special syntactic or

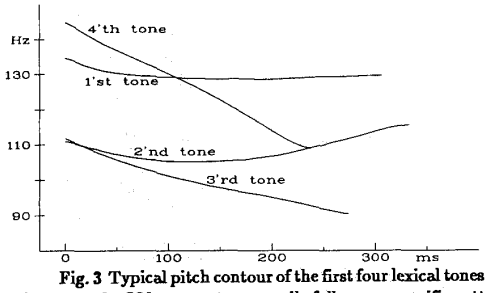


Fig. 3 Typical pitch contour of the first four lexical tones (those for the fifth tone not necessarily follow any specific pattern).

semantic roles. Therefore the recognition of the fifth tone should be considered in the Mandarin dictation machine. In this research, a study on the complete lexical tone recognition (i.e., including the fifth tone) for isolated syllables is performed. Various versions of hidden Markov models are considered including discrete HMM's, continuous HMM's and the recently proposed modified version of HMM's with bounded state durations (HMM/BSD), in which the duration for a model to stay at a certain state is upper and lower bounded by two bounding parameters to prevent a state from occupying too many or too few signal frames than appropriate. In other words, the probability density function $P_i(d)$ for a model to stay at a state i for a duration d is in general geometric just as the conventional HMM's, but with the upper and lower portions removed from the upper and lower bounds. These bounds, on the other hand, can be estimated from the maximum likelihood state transition sequences obtained during the training process.

The test results obtained in experiments are listed in Table 2. When only the first four tones were tested, the conventional approaches

Table 2 The recognition rates for the tones

Test conditions		Continuous HMM's	Discr. HMM's (codebook size 32)	Discr. HMM's (codebook size 16)
Only four tones were tested	Conventional feature vectors	96.5%	97.0%	97.5%
All five tones were tested	Conventional feature vectors	90.3%	89.6%	90.0%
	New feature vectors	93.6%	92.5%	91.2%
	New feature vectors Bounded state duration	94.7%	93.8%	92.0%

previously proposed [9,10] using feature vectors simply obtained from pitch frequencies gave very good results (97.5%). However, this number was seriously degraded when all the five tones were tested (on the order of 90.0%). Apparently this is due to the fact that the pitch frequency contour of the fifth tone do not necessarily follow any specific pattern as mentioned above, therefore the feature vectors simply obtained from pitch frequencies can not very well distinguish the fifth tone from the other tones. In this research a new form of feature vector is thus developed, which includes not only the pitch contour features, but the short-time energy and syllable duration of voiced part. This is because it was found that although the fifth tone does not necessarily follow any typical pitch contour as the other four tones, but lower energy and shorter duration are very often observed. The last two rows of Table 2 indicate that using new form of feature vectors, the recognition rate can be improved to 93.6%, and further improvement up to 94.7% can be achieved using the approach of bounded state duration mentioned above.

V. The Markov Chinese Language Model

Even if the syllables and tones can be correctly recognized using speech processing approaches, the high degree of ambiguities due to homonym characters still causes serious problems for the linguistic decoding process in the Mandarin dictation machine. This is because each Chinese character is pronounced as a monosyllable and in average each phonologically allowed syllable can represent at least 15 homonym characters. Thus the choice of the correct characters each syllable presents is a very difficult task.

In this research, various versions of Markov models for Chinese language are developed to solved this difficult linguistic decoding problem. The likelihood value for a given Chinese sentence is the product of the successive state transition probabilities, where the state can be either one (first order) or r (r-th order) characters or words (a word is composed of one to several characters), because unlike English language there is no boundary marker between two adjacent words or two adjacent characters in Chinese sentence, and every input syllable can have several possible candidates each with some given probabilities. The state transition probabilities are trained using a large quantity of training texts. For each sequence of input candidate syllables and probabilities, a word lattice can be constructed and the most probable output Chinese sentence can be obtained from the maximum likelihood path formed in this lattice. A good example for such a Chinese word lattice is shown in Fig. 4. Although the number of possible paths in the lattice is of

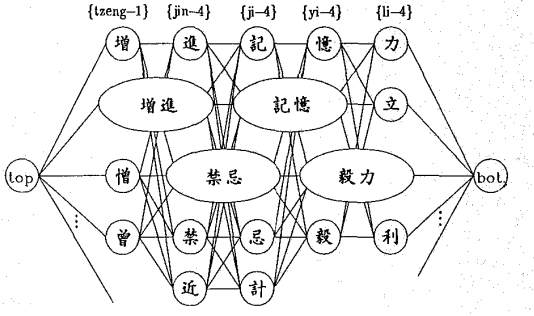


Fig. 4 A typical example for a Chinese word lattice.

exponential size in general, an efficient search algorithm based on dynamic programming was also developed to find the best solution in polynomial time.

The Markov models for Chinese language have been trained using primary school Chinese textbooks (Train-0 includes 8 volumes and Train-1 includes 12 volumes) and a dictionary of 40,000 commonly used Chinese words (Train-2). They are tested using three sets of database. The texts for Test-1 and Test-2 are of the same language style as the primary school Chinese textbooks, while Test-3 are obtained from daily newspaper whose language style and wording is in fact different. When the corresponding syllables and tones for these test texts are completely correct, the decoding rates are listed in Table 3. Note that in the

Table 3 Test results for the Markov Chinese language model

		Test-1	Test-2	Test-3
Order zero	Train-2	85.6%	88.0%	79.3%
Order one	Train-0	87.3%	87.4%	——
	Train-1	90.8%	90.8%	64.9%
	Train-1 plus Train-2	89.6%	91.2%	74.9%

Mandarin dictation machine errors always exist in syllables and tones and the Markov Chinese language model is provided with the first few choices of the syllables and tones each with corresponding probabilities. Therefore the numbers in Table 3 only serve as an upper bound for the actual performance of the language model. It can be seen from Table 3 that the decoding rates for Test-3 is apparently significantly lower due to the different language style and wording. It is therefore recommended that currently the Markov Chinese language model should be trained for each respective user to include the user's language style and word domain. This is a very reasonable limitation because the Mandarin dictation machine is speaker dependent. If this is considered, Table 3 shows that using order one model the decoding rate is on the order of 90%.

VI. Real-time Implementation of the Machine

The block diagram of the real-time Mandarin dictation machine is shown in Fig. 5, in which an IBM PC/AT is the control center, with three

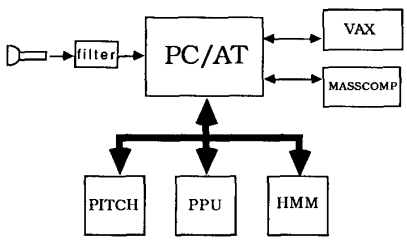


Fig. 5 The block diagram of the real-time Mandarin dictation machine

sets of special hardware boards connected to the PC/AT, the pitch detector, the preprocessing unit and the HMM processor. The waveform of the input unknown syllable is first filtered and sampled. The pitch frequencies are detected by the pitch detector. The preprocessing unit is responsible for the end point detection, pre-emphasis and real-time evaluation of all autocorrelation and LPC feature parameters and vectors, while the 409 HMM's are implemented on the HMM processor, on which all computations needed for the HMM syllable recognition are performed. The IBM PC/AT is not only the monitor and control center, but will take care of the recognition of the tones and the Markov Chinese language model to transform the recognized syllables into Chinese characters. The MASSCOMP-5400 workstation and the VAX 730 minicomputer, on the other hand, are responsible for the training processes for the syllable recognition and the Chinese language model respectively, and the parameters obtained in the training processes will be transferred to the PC/AT.

The pre-processing unit includes three Texas Instrument TMS 320C25 digital signal processor chips C25-1, C25-2 and C25-3, where C25-1 and C25-2 are responsible for the processing of the speech data, and C25-3 is to take care of the connection and data transmission among various units of the system, including the PC/AT, the pitch detector, and the HMM processor. The HMM processor, on the other hand, has one main board on which seven sub-boards are plugged in. The center of each sub-board is also a TMS 320C25 chip providing all necessary computations. The program for recognition of the unknown syllable using the 409 syllable HMM's is written on the seven sub-board, with the 409 syllable HMM's divided into seven groups, each having 58-59 syllable HMM's implemented on one sub-board. Therefore totally ten TMS 320C25 chips are used, operating in parallel to complete all necessary computations in time. the picture of the completed real-time Mandarin dictation machine is is Fig. 6. The tests indicate that the average time for this machine to dictate a syllable is 0.45 sec, which is about the duration for a speaker to pronounce an isolated syllable.

VII. Conclusion

The first successfully implemented real-time Mandarin dictation machine developed in the world which recognized Mandarin speech with unlimited texts and very large vocabulary has been completed. Isolated syllables are recognized using hidden Markov model techniques and transformed into Chinese characters through a Markov Chinese language model under speaker dependant mode. The overall system performance is still under test. The technology used in this machine is quite different from those machines for other languages due to the very special characteristics of Chinese language.

References

[1] Bahl, L. R., et al., "Large vocabualry natural language continuous speech recognition", ICASSP (Glasgow, Scotland), pp. 465-467, 1989.

[2] Merialdo, B., "Multilevel Decoding for Very-large-size dictionary Speech Recognition", IBM J. Res. Develop., Vol. 32(2), pp. 227-237, Mar 1988.

[3] D'Orta, P., et al., "Large-vocabulary speech recognition: a system for the Italian language", IBM J. Res. Develop., Vol. 32(2), pp. 217-226, Mar. 1988.

[4] Averbuch, A., et al., "Experiments with the Tangora 20,000 word speech recognizer", ICASSP (Dallas, TX), pp. 701-704, 1987.

[5] Jelinek, F., "The development of an experimental discrete dictation Recognizer", Proc. IEEE, vol. 73(11), pp. 1616-1624, Nov. 1985.

[6] Rabiner, L. R., B. H. Juang, S. E. Levinson and M. M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities", AT&T Tech. J., pp. 1211-1234, July-Aug. 1985.

[7] Juang, B. H. and L. R. Rabiner,"Mixture Autoregressive Hidden Markov Models for Speech Signals", IEEE trans. ASSP, pp. 1404-1413, 1985.

[8] Russell, M. J. and R. K. Moore, "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition", ICASSP, pp. 5-8, 1985.

[9] Yang, W. J., J. C. Lee, Y. C. Chang and H. C. Wang, "Hidden Markov Model for Mandarin Lexical Tone Recognition", IEEE trans. ASSP, Vol. 36, pp. 988-992, July 1988.

[10] Chen, X. X., C. N. Cai, P. Guo and Y. Sun, "A Hidden Markov Model Applied to Chinese Four-tone Recognition", ICASSP (Dallas, TX), pp. 797-800, 1987.

[11] Katz, S. M., "Estimation of probabilities from sparse data for the language model component of a speech recognizer", IEEE trans. ASSP, vol. 35(3), pp. 400-401, Mar. 1987.

Acknowledgement

The work described in this paper is supported by the National Science Concil of Taiwan, Republic of China from 1984 to 1989.

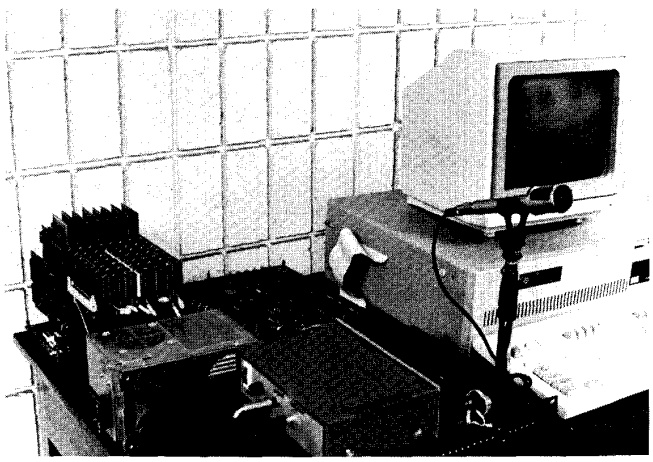


Fig. 6 The picture of the completed real-time Mandarin dictation machine