Golden Mandarin (III) — A User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary

Ren-Yuan Lyu[1], Lee-Feng Chien[3], Shiao-Hong Hwang[1], Hung-Yun Hsieh[1], Rung-Chiuan Yang[2], Bo-Ren Bai[1], Jia-Chi Weng[2], Yen-Ju Yang[2], Shi-Wei Lin[2], Keh-Jiann Chen[3], Chiu-Yu Tseng[4], Lin-Shan Lee[1,3]

[1]Dept. of Electrical Engineering, National Taiwan University,
[2] Dept. of Computer Science and Information Engineering, National Taiwan University,
[3] Institute of Information Science, Academia Sinica,
[4] Institute of History & Philogogy, Academia Sinica,
Taipei, Taiwan, R.O.C.

Abstract

This paper presents a prototype prosodic-segment-based Mandarin dictation machine for the Chinese language with very large vocabulary. It accepts utterances continuous within a prosodic segment which is composed of one or a few word(s). It also possesses various on-line learning capabilities for fast adaptation to a new user in acoustic, lexical and linguistic levels. The overall system is implemented on an IBM/PC with an additional DSP card including a Motorola DSP 96002 chip. The word accuracy can achieve nearly 90% for a new user after he produces about 10 minutes of speech to train the system, and the accuracy can be further improved with the on-line learning functions.

I. Introduction

Today, the input of a large number of different Chinese characters into computers is still a very difficult and unsolved problem. It has long been believed that voice input will be a very attractive solution. This is the basic motivation for the development of a Mandarin dictation machine. Such a machine should thus be able to recognize Mandarin speech with very large vocabulary and almost unlimited texts. A series of prototype Mandarin dictation machines, namely Golden Mandarin (I) and (II) [1][2], have been successfully developed under the constraints of isolated syllabic utterances. The present machine, Golden Mandarin (III), makes an important step forward beyond the previous versions. It now releases the isolated syllabic constraint and allows the speakers to utter a sentence as several prosodic segments, each composed of one to several words and uttered continuously. In this way, not only the dictation process becomes very natural, but also the dictation speed can be improved because the pauses between syllables becomes unnecessary.

There exist at least more than 80,000 commonly used words, more than 10,000 commonly used characters, but only a total of 1345 syllables in Mandarin Chinese . Also, Mandarin Chinese is a tonal language, there exist 5 different tones (4 lexical tones plus a neutral tone), and when the differences in tones are disregarded, the 1345 different syllables are reduced to only 408 different base-syllables (BS's, i.e., tone-independent syllables). Every Chinese word is composed of from one to several characters, but all characters are pronounced as a mono-syllable. For this mono-syllabic structure of Mandarin Chinese, it was a natural choice to develop the previous versions of the Mandarin dictation machine based on isolated syllable recognition[1][2]. However, there exist several limitations in the isolated syllable approach. First, it is not natural to produce a sentence in isolated syllable mode, and such a mode inevitably limits the dictation speed. Secondly, the set of Mandarin syllables composes of many confusing subsets, which make accurate recognition difficult. Thirdly, there exist many homonym characters sharing the same syllable, and this makes the linguistic decoding very difficult in choosing the correct characters. These are probably the reasons why several large vocabulary poly-syllabic Chinese word recognition systems have been proposed recently[3][4]. However, since the Chinese words are not well defined, unlike the English language where a blank is a natural word boundary in sentences, the segmentation of a Chinese sentence into words is usually not unique and the segmented words obtained by a user are very probably not included in the given lexicon. This situation makes the conventional word-based recognition systems impractical to be used for dictation purposes, because these systems can recognize correctly only the words defined in the lexicon. In fact, when dictating a sentence, especially a long one, the users will very naturally make pauses for each breath group to form several prosodic segments. Here a prosodic segment is an utterance easily produced as a breath group and usually composed of one to several words. Compared with a sentence, a prosodic segment is shorter in duration, and simpler in structure. For all the reasons stated above, the prosodic segment was chosen as the input unit to the present system.

The system presented in this paper not only recognizes the prosodic segments using improved techniques, but also includes a Chinese language model to differentiate the large number of homonym mono-syllabic and bi-syllabic words and handle other linguistic problems. Furthermore, various adaptive/learning functions were developed to make the system more intelligent and friendlier. The overall system is shown in <fig. 1>. The speech waveform is first decoded into a base-syllable string (BSS), which corresponds to one word, several homonym words, or the concatenation of several words. The tone pattern (TP) of the utterance is also recognized by the TP recognizer. This TP is then combined with the BSS to access

0-7803-2431-5/95 $4.00

the correct word by the lexical accessor. The above three blocks constitute the word recognizer to be described in more detail in section II. Usually, there still exist more than one words (or concatenation of words) corresponding to the same BSS plus TP, thus a word lattice is constructed and a word-class-based Chinese language model is used in the linguistic decoder to be described in section III. Furthermore, the system can on-line adapt itself from the feedback of the user. For example, it will emphasize or de-emphasize the corresponding acoustic, lexical or linguistic parameters by knowing the correct words for the input speech. These user-adaptive functions will be described in section IV, while the system implementation issues are given in the last section.

II. Mandarin Poly-syllabic Word Recognition

Unlike the other large vocabulary word recognition systems which can only recognize the words defined in the lexicon, our word recognizer here can recognize the prosodic segments composed of several words (mono-syllabic or poly-syllabic), and output a word lattice to be post-processed by a linguistic decoder. This is achieved by three blocks, namely BSS recognizer, TP recognizer, and lexical accessor.

The lexicon used in the present system has a total of 84,495 words, for which there are 52,813 different BSS's . To deal with the vocabulary of this quantity, instead of using the conventional sub-word unit for Mandarin speech, the INITIAL-FINAL's as used in the previous works [3][4], some kind of smaller units were used in this study. It was found that only 33 phone-like-units (PLU's) as listed in <table. 1> are enough to transcribe all Mandarin speech, and if only the intra-syllabic coarticulation is considered, which is the most important coarticulation effect in Mandarin speech, a total of 149 right-context-dependent (RCD) PLU's will be very useful in recognition of the 52,813 BSS's for Chinese words. These RCD-PLU's were used as the basic acoustic units for our BSS recognizer and each of them was modeled as a continuous Hidden Markov Model (CHMM) with 2 states and 3 Gaussian mixtures per state. Compared with the INITIAL-FINAL's, this set of units is easier to be modeled. Also, the smaller state number of these units requires less speaker-specific training data in the speaker adaptation process, such that a new user can use the dictation machine as soon as possible. In the training phase, the segmental K-means algorithm [5] was adopted, while in the recognition phase, those RCD-PLU-CHMMs were concatenated into the 408 phonologically allowed base-syllables (BS's) first. These BS models were further connected into a syllabic net as in <fig. 2> where the final state of each BS model was linked with the first state of each BS. A frame-synchronous dynamic programming search algorithm for matching an utterance with N best BSS's [6] was then used as the kernel algorithm.

For the tone patterns (TP), it was observed that only mono-syllabic and bi-syllabic words need TP recognition. The reason is that for poly-syllabic words containing more than two syllables, each BSS corresponds to exactly one word and thus only BSS is enough to access the word correctly. The TP recognizer used here was composed of a pitch analyzer and a Gaussian classifier. This classifier includes 5 discriminative functions for the 5 possible TP's of mono-syllabic words, and 25 discriminative functions for the 25 possible TP's of bi-syllabic words.

Combining the candidates of BSS's and TP's for each unknown utterance , the corresponding word or concatenation of words can be accessed from the lexicon by a set of specially designed prosodic segment rules which describe how several words can be combined into a prosodic segment. For example, it is very often that mono-syllabic words are naturally pronounced together with a preceding or subsequent word, primarily depending on the categories and length of the preceding or subsequent words, e.g., the word "要(want)" is often pronounced together with its previous proper noun like "我要(I want)", the word "好(good)" with its subsequent objects like "好東西 (good things)", etc. Because the number of frequently used mono-syllabic words is not large, such rules can be acquired from analysis of corpus.

To test the initial performance of this word recognizer, a test set composed of 644 poly-syllabic words selected from the lexicon was first read by two male speakers. The top-1 and top-5 word accuracy's for this test set achieve 90.06% and 98.76% respectively in speaker dependent mode with about 50 minutes of training speech available for each speaker. However, in addition to those poly-syllabic words, almost each character in the Chinese language can also be used as a mono-syllabic word, and some of them are used very frequently in daily used sentences. But it is difficult to access correctly these mono-syllabic words from the lexicon using BS and TP information only, because for each combination of BS and TP, there exist more than 10 homonym mono-syllabic words in average. When the N-best syllable candidates are considered, a large number of word candidates form a large word lattice with possible paths linking these words, a partial listing of an example is shown in <fig. 3>. This is where the language model to be described below becomes necessary.

III. Chinese Word-class-based Language Model

Conventional word-based Markov language models, due to their demand for huge number of properly estimated model parameters, are in general inadequate in the present tasks with very large vocabulary. Word-class-based language models, in which words with similar linguistic properties are clustered into smaller number of word classes, are therefore important and highly desired because the number of model parameters was significantly reduced. In the present system, an efficient three-stage hierarchical word classification algorithm integrating both grammatical and statistical information was proposed, and a word-class-based language model based on the obtained word classes was developed.

The most difficult issue in developing an efficient word-class-based language model is the strategy to classify the words into word classes. In the present system, a new three-stage hierarchical word classification algorithm was developed, which integrated the advantages of both grammatical and statistical approaches. In order to solve the problem that some rarely-used

words did not have sufficient statistical information, the classification process was divided into three stages, each with different strategy as shown in <fig. 4>. This algorithm is in general applicable to any language, but the following description will be based on studies on our Chinese lexicon. In the first stage, all the words were roughly grouped according to their linguistic features of part-of-speeches assigned by linguists. For example, if a word has three different part-of-speeches, e.g., Active-Verb-B, Proper-Noun-A and Proper-Noun-C, it will be grouped with the words exactly having these three part-of-speeches. In this way, about 200 part-of-speeches in Chinese were used and about 950 initial classes were obtained. In the second stage, the words in the same class, which were believed to have similar syntactic behavior, were further grouped into smaller classes based on their statistical behavior[7]. That is, the words having similar word-co-occurrence feature vectors were further grouped into even smaller classes, if the similarity measure between them exceeding a threshold. It is believed that much of the implicit semantic information has been integrated in the second stage. In the third stage, however, in order to avoid too restrict classification, some classes obtained in the second stage with different part-of-speech features can be further merged together according to the statistical similarity between them, even if they have been separated apart in the first stage.

In general, with the above procedure, words clustered in the same class have similar syntactic and semantic behavior, because both of their part-of-speech features and co-occurrence relations with adjacent words in the corpus are very similar. Also, with the above procedure, the number of finally obtained classes is adjustable for different application tasks, considering factors such as accuracy and storage space. In our experiments, it was found that roughly 2,000 classes are very good choice for our Chinese lexicon. Furthermore, for those rarely-used words with insufficient statistical information, the classifications are still satisfactory because the part-of-speech features have been carefully used. This is why the language models based on this word classification algorithm turns out to be more reliable and robust with significantly reduced model parameters of desired number. With the word-class-bigram obtained as above, a Viterbi search can be performed on the word lattice in the linguistic decoder as shown in <fig. 1>, and a best word string was then chosen as the system output.

IV. Several User-adaptive Functions

If a dictation machine is to be practically used by many users efficiently, a key feature is that it should be easily adapted to a new user. Various user-adaptive functions are thus developed in the present system. These functions are divided into two broad categories, namely the acoustic level functions and the linguistic level functions.

In the acoustic level, fast on-line MAP-based speaker adaptation was developed. The system parameters, e.g. those for RCD-PLU-CHMM's, were first initialized by a speaker-independent (SI) database. The top-1 word accuracy for the same test set mentioned in section II for a new outside speaker is only 62.68% for this SI training, simply because the SI database we adopted covers too few speakers. However, four stages of incremental speaker adaptation based on four sets of pre-selected phonetically balanced words covering all necessary phonetic events with minimal number of word utterances were developed here. The first set of utterances (SA1) covers all the necessary RCD-PLU's, the second (SA2) covers all BS's, the third (SA3) covers all transitions among PLU's, and the fourth (SA4) covers all 1345 phonologically allowed tonal syllables. Testing results showed that with 4.35 minutes of speech of the target speaker (up to the third stage) the top-1 word accuracy will be raised to 85.40%, and with 9.4 minutes of speech (up to the fourth stage), to 89.60%, very close to the speaker dependent (SD) training result (90.06%) which needs about 50 minutes of speech. The complete speaker adaptation (SA) results are listed in <table. 2>. Further on-line adaptation after these four stages is still possible. The on-line adaptation functions also include adaptation to environmental noise and acoustic variations.

In the linguistic level, not only the new words can be learned on-line but an on-line adaptation technique based on linear interpolation was applied on the parameters of the word-class-based language model. Furthermore, a short-term learning mechanism was also developed, in which the dynamic parameters were stored in a cache memory accessed prior to the static parameters trained for general use. In this way, both the special words for a special dictation task and the personal wording and sentence style for the user can be learned by the system. As a result, the performance of the system can be improved after a user has used the system for some time.

V. System Implementation and Concluding Remarks

Golden Mandarin (III) is currently implemented on a Motorola DSP 96002 chip buried in an Ariel DSP-96D developing card plugged in an IBM/PC compatible, with an additional Ariel ProPort 656 as the pre-amplifier and A/D converter at 16-bit resolution and 16 KHz sampling frequency. The system runs under Microsoft Windows and provides users a fully graphical user interface (GUI). It includes an acoustic manager to record the new user's speech and then create a speaker specific acoustic model set. It also includes a linguistic manager to allow users to add new words, or train a user specific language model. Another on-line learning module performs the various functions mentioned previously. With all the on-line learning capabilities, this word accuracy of the system can easily exceed 90% for a new user.

Golden Mandarin (III) is the third version of prototype systems developed in a long term project, with the goal of solving the difficult problem of input of Chinese characters into computers using voice. The major breakthrough achieved in this version is that it accepts utterance continuous in prosodic segments, which is very close to continuous speech except the length of utterance is usually shorter. This is achieved by several new techniques, including PLU based acoustic modeling, an N-best frame synchronous dynamic network searching algorithm, a new three-stage hierarchical word classification algorithm and a new word-class-based Chinese language model.

Reference:

[1] Lin-shan Lee, et al. "Golden Mandarin (I)- A Real-time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary", IEEE Tran. on Speech and Audio Processing, April, 1993. pp158-179

[2] Lin-shan Lee, et al. "Golden Mandarin (II)- An Improved Single-Chip Real-time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary", ICASSP-93, Minneapolis USA, pp.503-506

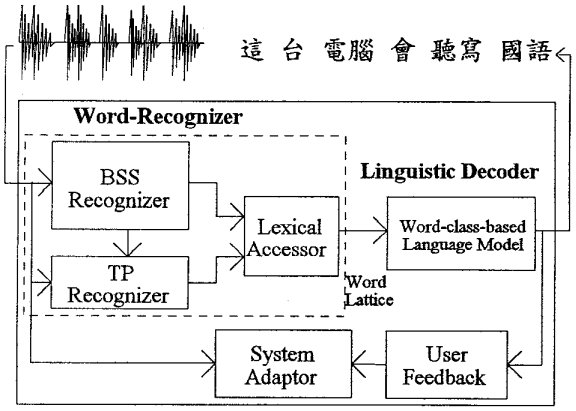
[3] Hsiao-Wuen Hon, Kai-Fu Lee, et.al, "Toward Large Vocabulary Mandarin Chinese Speech Recognition", ICASSP-94, Adelaide, Australia, Vol.1 p.545-548

[4] Jung-Kuei Chen, F.K.Soong, "Large Vocabulary Word Recognition Based on Tree-Trellis search", ICASSP-94, Adelaide, Australia, Vol.2,p.137-140

[5] L.R. Rabiner, et al. "A Segmental K-Means Training Procedures for Connected Word Recognition", AT&T Technical Journal. May/June Vol. 65 Issue 3, p21-31.

[6] C.H.Lee, et al, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition", IEEE T-ASSP, Vol.37, No.11, Nov. 1989 p1649-p1658

[7] Hinrich Schutze, "Part-of-speech Induction from Scratch", ACL'93, pp. 251-258, 1993.

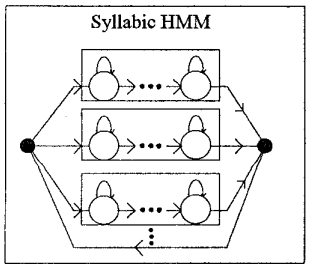


<fig. 1> The overall system block diagram

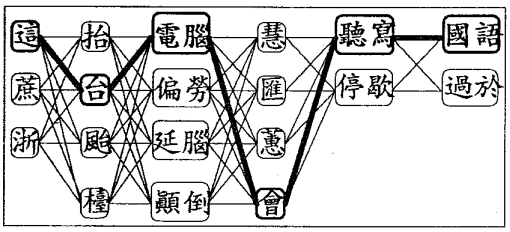
p	t	k	p'	t'	k'	ts	tʂ	tɕ	ts'	tʂ'
tɕ'	m	n	ŋ	l	f	s	ʂ	ɕ	x	z
a	o	ɤ	e	i	u	y	ɿ	ʅ	ɚ	#

<table. 1> The PLU's for Mandarin speech used in this paper

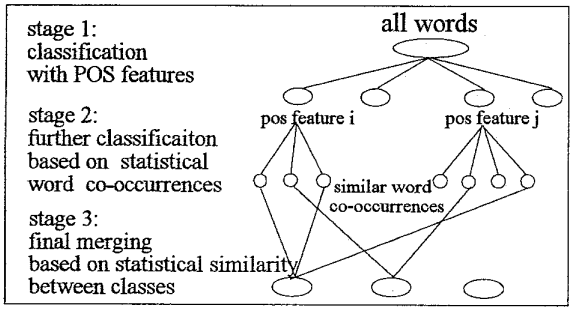
Note: The International Phonetic Alphabet (IPA) is used here. The symbol # is used for the Null INITIAL in Mandarin speech.



<fig .2> The syllable net composed of syllabic HMM



<fig. 3> A partial listing of an example word lattice.
Note: The correct word string here is "這 台 電腦 會 聽寫 國語 (This computer can be dictated using Mandarin.)"



<fig. 4> The three-stage hierarchical word classification algorithm

	Data provided by the new speaker			*BSS Acc%*	*Word Acc %*
	number of words	number of syllables	length of speech		
SI	0	0	0 min	***64.74***	***62.58***
SA1	53	162	0.88min	***75.76***	***73.76***
SA2	186	511	2.81min	***85.27***	***81.15***
SA3	286	794	4.35min	***89.30***	***85.40***
SA4	665	1685	9.4 min	***92.25***	***89.60***

<table.2> Database and system performance for four-stage incremental speaker adaptation

Note: SI means speaker independent training, and SA1, SA2, SA3, SA4 mean speaker adaptation for the four stages