1994 International Symposium on Speech, Image Processing and Neural Networks, 13-16 April 1994, Hong Kong

Golden Mandarin(II) - An Intelligent Mandarin Dictation Machine for Chinese Character Input with Adaptation/Learning Functions

Lin-shan Lee[1,2,3], Keh-Jiann Chen[3], Chiu-yu Tseng[4], Renyuan Lyu[1], Lee-Feng Chien[3],
Hsin-min Wang[1], Jia-lin Shen[1], Sung-Chien Lin[2], Yen-Ju Yang[2], Bo-Ren Bai[1], Chi-ping Nee[3],
Chun-Yi Liao[3], Shueh-Sheng Lin[1], Chung-Shu Yang[2], I-Jung Hung[1], Ming-Yu Lee[1], Rei-Chang Wang[1],
Bo-Shen Lin[1], Yuan-Cheng Chang[2], Rung-chiung Yang[2], Yung-Chi Huang[1], Chen-Yuan Lou[1], Tung-Sheng Lin[1]
National Taiwan University and Academia Sinica, Taipei, Taiwan, Republic of China*

ABSTRACT - Golden Mandarin (II) is an intelligent single-chip based real-time Mandarin dictation machine for Chinese language with very large vocabulary for the input of unlimited Chinese texts into computers using voice. This dictation machine can be installed on any personal computer, in which only a single chip Motorola DSP 96002D is used, with a preliminary character correct rate around 95% at a speed of 0.6 sec per character. Various adaptation/ learning functions have been developed for this machine, including fast adaptation to new speakers, on-line learning the voice characteristics, task domains, word patterns and noise environments of the users, so the machine can be easily personalized for each user. These adaptation/ learning functions will be the major subjects of this paper.

I. Introduction

Today, the input of 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. We defined the scope of this research by following limitations. The input speech is in the form of isolated syllables. The machine is speaker dependent, but should be easily adapted to new speakers. Reasonable errors are acceptable because they can be found on the screen and corrected with a mouse by the user very easily. But the machine has to be able to recognize Mandarin speech with very large vocabulary and unlimited texts, because the input to computers can be arbitrary Chinese texts. Also, the machine has to work in real-time for computer input applications. A previous version of such a machine, Golden Mandarin (I), has been developed in 1990 [1][2], but the highly computation-intensive algorithms for Golden Mandarin (I) require 10 TMS 320C25 chips operating in parallel on 9 special hardware boards to meet the real-time requirements, and it is only a speaker dependent system without adaptation / learning functions. This is why the present machine is developed using completely different algorithms. The present machine, Golden Mandarin (II), on the other hand, achieved almost real-time operation at a speed of 0.6 sec per character and around 95% character correct rate using a single chip Motorola DSP 96002D [3], and various adaptation/ learning functions were developed to make the machine friendlier to the user. The latter will be the major subject of this paper although the former will be briefly summarized also.

There are at least 10^5 commonly used Chinese words, each composed of one to several characters. There are at least 10^4 commonly used Chinese characters, all produced as monosyllables. However, the total number of different syllables in Mandarin speech is only 1345. Based on such observation, the use of syllable as the dictation unit becomes a very natural choice. Another very special feature of Mandarin Chinese is that it is a tonal language. Every syllable is assigned a tone in general. There are basically four lexical tones and one neutral tone in Mandarin. 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 carrier syllables. If the differences in tones are disregarded, only 408 base-syllables (each bearing different tones) are required 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 408 base-syllables disregarding the tones. Based on the above considerations, the overall system structure for the Golden Mandarin (II) dictation machine is shown in Fig.1. The system is basically divided into two subsystems. The first is to recognize the Mandarin syllables, and the second is to transform the series of syllables into Chinese characters, because every syllable can be shared by many homonym characters. For the first subsystem of Mandarin syllable recognition, the base-syllable (disregarding the tones) and the tone are recognized independently in parallel. For the second subsystem of Chinese language model, we need to first obtain all possible word hypothesis to construct a Chinese word lattice, and then use Chinese word class bigram to select the most probable concatenation of word hypotheses as the output sentence.

II. Mandarin Syllable Recognition and Chinese Language Model

The recognition of the Mandarin syllables includes two

*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 and Philology, Academia Sinica

parts: recognition of the 408 base syllables (disregarding the tones) and recognition of the tones. The tone recognition is not too difficult. Discrete Hidden Markov Models are used, with special measures applied to distinguish the neutral tone from the four lexical tones. The recognition of the 408 base-syllables (disregarding the tones), however, is very difficult, because there exist 38 confusing sets in this vocabulary. A good example is the A-set, { a, ja, cha, sha, dsa, tsa, sa, ga, ka, ha, da, ta, na, la, ba, pa, ma, fa }. Specially trained continuous density Hidden Markov Models (HMM's) [4][5] for cepstral coefficients were used in the previous version machine [1][2], which are highly computation-intensive. Considering the fact that Mandarin mono-syllables have relatively simple phonetic structure and the primary problem in base-syllable recognition is to distinguish the very confusing initial consonants instead of matching the entire template, it is therefore believed that the time warping functions of the state transition probabilities of HMM's are not very important. Because state transition path searching process in HMM's is highly computation-intensive, a segmental probability model (SPM) specially for Mandarin base-syllables was therefore developed [3], which is very similar to continuous density HMM's with each state modeled by Gaussian mixtures, but the state transition probabilities are deleted and the N states equally segment the syllable utterance. The detailed developments and experiments for the application of such SPM technique to the accurate, fast recognition of the Mandarin base-syllables have been reported previously [3], therefore not repeated here in this paper.

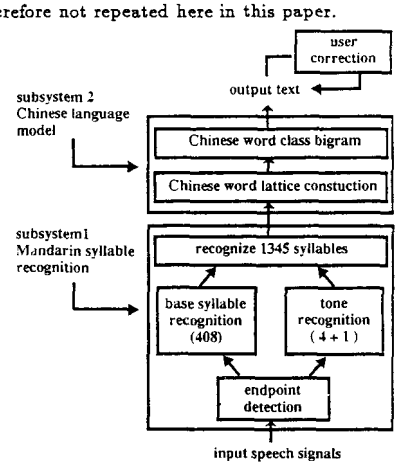

Figure 1: The overall system structure for the Golden Mandarin(II) dictation machine

After the base syllables and tones are recognized by the subsystem 1, the high degree of ambiguity caused by the large number of homonym characters still remain to be solved. The subsystem 2 thus acts as a linguistic decoder to identify the characters using context information. In the previous version machine [1][2], a relatively simple Chinese character bigram trained by primary school Chinese textbooks [6][7] was used, whose function was in fact limited. In Chinese language every word is composed of from one to several characters and there is no blanks between two adjacent words, thus a sentence can be considered as a sequence of words, or a sequence of characters. The 10^4 characters or 10^5 words require a character bigram of $10^4 \times 10^4$ probabilities or a word bigram of $10^5 \times 10^5$ probabilities respectively. Preliminary tests indicated that the word bigram is much more powerful than the character bigram [6], probably because the Chinese sentences are really built by words rather than by characters. But the word bigram is difficult to train and implement because of the much larger size. A new approach considering the special structure of Chinese language using a Chinese word class bigram was thus developed to solve this problem [3]. In this approach, the sequence of syllables obtained from the subsystem 1 is first matched with the words in a lexicon of 10^5 words to find all possible word hypotheses to construct a word lattice, which is a graph of all possible paths connecting all word hypotheses. The paths on the word lattice are then searched through by a Chinese word class bigram. The path with the highest probability is then chosen as the output, with the conditional probabilities estimated based on the beginning and ending characters of the word classes. This is because in this approach words with identical beginning or ending characters are categorized as a word class. The detailed developments and experiments on the application of such Chinese word class bigram on Mandarin dictation have been reported previously [3], therefore not reported here in this paper.

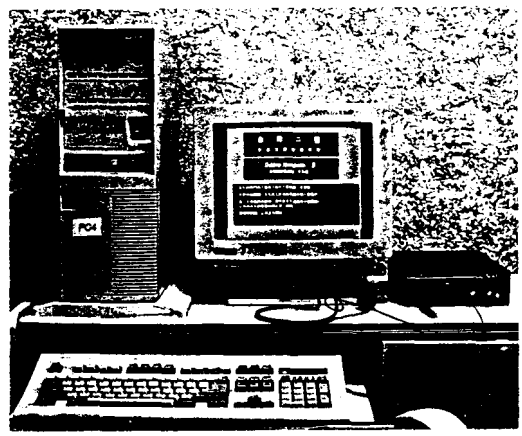

Figure 2: The picture of the Golden Mandarin(II) dictation machine

In the real-time implementation of the Golden Mandarin (II) [3] all necessary computation for the Mandarin syllable recognition is performed in a single chip Motorola DSP 96002D installed on an IBM PC/AT, and the Chinese language model is implemented on the IBM PC/AT itself, while a ProPort Model/ 656 acts as the front end for acoustic signals. The waveform of the input unknown syllable is filtered and sampled in ProPort and transformed into 16-bit integer format, DSP96002D and the IBM PC/AT then

sponsors all the following processes including endpoint detection, pre-emphasis, Mandarin syllable recognition and the Chinese language model. A picture of the completed Golden Mandarin (II) is shown in Fig.2, which just looks like a normal IBM PC/AT. The only difference is the D-SP96002D and the Golden Mandarin(II), both of which are invisible in the picture.

III. Fast Speaker Adaptation for Golden Mandarin (II)

In the part experiences of developing Mandarin dictation machines, the huge number of training utterances needed for a new speaker to train the machine has always been a difficult problem, especially due to the fact that it is a very boring process for a new user to produce large number of isolated monosyllables one by one, because these monosyllables generally do not bear any meaning. It is therefore obvious that some efficient/intelligent adaptation functions for the machine to adapt to a new speaker is highly desired. This includes at least the following features: the number of needed training utterances should be reduced to minimum, the recognition rate should be improved as soon as possible, the necessary syllables should be grouped into meaningful sentences to make the training processes interesting for the new user, and some on-line learning functions should be provided such that the new user can start to use the machine as soon as possible and the performance can be continuously improved after the new user start using the machine. The first few features will be discussed in this section, while the last will be discussed in the next session.

The first approach adopted by Golden Mandarin (II) is that each training utterance should be shared by as many syllable as possible as long as they have some common INITIAL's or FINAL's. Here INITIAL/ FINAL are conventional acoustic units used for decomposing Mandarin syllables, very close to the consonant/ vowel format in other languages. The INITIAL's are the initial consonants in the Mandarin syllables, while FINAL's include the remaining parts of the syllables, primarily the vowel or diphthong parts but with optional medials (e.g. i, u) and nasal ending (-n or -ng). Note that Mandarin syllables can't have other consonant endings except for the nasal ending. For example, in a syllable /chiang/, /ch/ is the INITIAL part and /iang/ is the FINAL part. The basic idea for the training utterance sharing is that each training utterance should be used by as many syllable models as possible. For example, when a training utterance for the syllable /ba/ is uttered by a new user, models for the syllables /ban/, /bang/, /bau/, /bai/ ... etc. can all use its INITIAL part for adaptation because they share the common INITIAL's, while the models for the syllables /ma/, /pa/, /cha/, /ja/ ... etc. Can all use its FINAL part for adaptation because they share the common FINAL's. In this way, the necessary number of training utterances for achieving some desired recognition rates can be reduced to minimum.

The second approach adopted by Golden Mandarin (II) is to use computer algorithm to select automatically some sets of necessary phonetically balanced sentences from a large text corpus. Each of such sets includes almost minimum number of sentences but covers all necessary phonetic units, even with the statistical distribution of these phonetic units very close to that in the large text corpus. Therefore when the new speaker have read these sentences, all necessary phonetic units for the new speaker will be available and the training utterances will be sufficient to train the machine. A nice feature is that when the statistical distribution of the phonetic units are close to that in the large texts corpus, more frequently used phonetic units will appear more frequently in these specially selected sentences. Thus these phonetic units will be trained better and recognized more accurately. This can improve the overall recognition rate. In this way the time needed for a new speaker to achieve some desired recognition rate can be very short, but the processes of reading these phonetically balanced sentences won't be boring any more for the new user, because they are meaningful sentences.

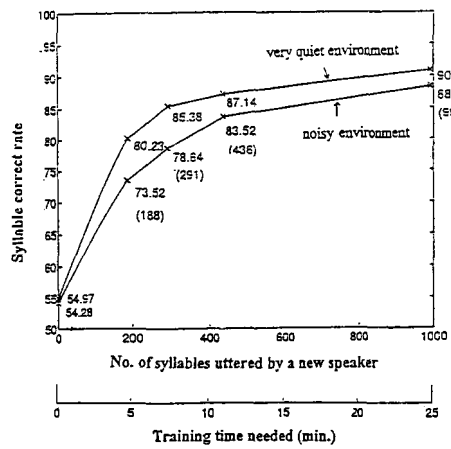

Figure 3: The Learning curves for Golden Mandarin(II) to adapt to a new speaker, with the lower curve for noisy environment

With the above adaptation approaches, four sets of phonetically balanced sentences have been selected to construct a four-stage incremental adaptation procedure. When the new user have not produced any training utterance, the multi-speaker model previously obtained provides an average syllable recognition rate of only about 55% for a new outside speaker. After a new user reads the first set of 24 phonetically balanced sentences (with a total of 188 characters or syllables, and it takes less than 5 minutes to read these sentences including overhead time), the syllable recognition rate is immediately improved to around 80%. Note that the 188 syllables constitute only less than one seventh of the total 1345 syllables, but they provide very good speaker adaptation function due to the training utterance sharing approach mentioned above. When the new speaker reads the second set of 13 additional sentences (with 103 additional syllables and 3 additional minutes only), the syllable recognition rate is improved to

about 85%. When the new speaker further reads the third set of 18 sentences and the fourth set of 69 sentences (145 syllables with 4 minutes and 556 syllables with 13 minutes respectively), the syllable recognition rate is improved to about 87% and 91% respectively. Note that the total syllables read in the four stages is only 992 and the total time needed is less than 25 minutes, but the syllable recognition rates of 91% is already quite acceptable for a new user to start using the machine. Fig.3 shows the learning curves for Golden Mandarin (II), in which the upper curve exactly sketches the four-stage incremental adaptation procedure mentioned above.

IV. Other Adaptation/ Learning Functions for Golden Mandarin (II)

The nice features of the SPM techniques for Mandarin syllable recognition briefly summarized above include not only the robustness in speaker adaptation and training utterances sharing as mentioned above, but also the relatively less computation involved in the adaptation processes due to the simple representation of Gaussian mixtures for each state. The result is that a real-time on-line learning routine has been implemented such that a new user can just go through the above four-stage incremental adaptation procedures sentence by sentence, and feel the improvements in recognition rate syllable by syllable in real time. The new user can also stop the training process and start using the machine any time, either within the four-stage adaptation procedure or after the four-stage adaptation procedure is finished. As long as he corrects the errors made by the machine on the screen via a mouse and operates the machine in "learning" mode, every utterance he produces in using the machine (correctly recognized or with errors corrected)will be used in further training the machine. Therefore the recognition rate will be continuously improved as long as the user continuously uses the machine and lets the machine learn.

The above speaker adaptation/ learning functions have another nice feature. During the on-line adaptation/ learning procedure the machine can also simultaneously adapt to some extent to the environmental noise of the new user, simply because the Characteristics of the background noise in the user's environment will also be automatically averaged and included in the SPM model parameters. The lower curve in Fig.3 actually shows such an effect. The upper curve as discussed above is for quiet environment, while the lower curve is for some noisy environment with air condition and more than ten computers running. As can be seen in Fig.3, in the first two stages of speaker adaptation, the noise environment gives a degradation in syllable recognition rate on the order of 7%, but this degradation is quickly reduced to about 3.5% and even 2.5% in the third and fourth stages. Apparently the noise characteristics have been included in the SPM model parameters during the on-line adaptation/ learning procedure.

All the adaptation/learning functions discussed up to this point have to do with the robustness of the SPM technique for Mandarin syllable. In fact, the Chinese word class bigram developed for the Chinese language model and briefly summarized above is also robust in terms of adaptation/ learning functions with respect to the task domains and word patterns of the users or of the texts being dictated. In the past experiences in developing Mandarin dictation machines, the wide variety in task domains of the texts to be dictated and the complicated word patterns in texts with completely different subjects, have caused another difficult problem. Very often when the texts being dictated switch from one subject domain to another, the performance of the machine becomes very difficult to predict. A nice feature of the Chinese word class bigram used in Golden Mandarin (II) is that it is also kind of robust in terms of adaptation/learning with respect to the task domains and word patterns. In Golden Mandarin (II), not only the bigram parameters and word frequencies can be on-line adapted with adjusted weights, but the basic structures of the lexicon and the language model are also specially designed to make such adaptation/ learning functions easy and efficient. Very often when a user is entering some texts on a certain subject, some special word patterns appear repeatedly. As a result, with the adaptation/learning functions mentioned here, if some errors occurring on some special word patterns are corrected, very probably the following similar word patterns will be accurately recognized. In this way the performance of the machine can be continuously improved when some texts with a certain subject is being dictated, or a user continues to enter texts with similar task domain and word patterns.

Combining all the adaptation/ learning functions discussed above together, a new user can in general start using the machine in 5 minutes and find the machine rather convenient to use in about 25 minutes. By letting the machine continue to learn the voice and noise characteristics of the user as well as the task domains and word patterns of the texts being entered, the machine can be easily "personalized", i.e., adapted to the user himself, in which case the character correct rate can be around 95%.

V. User Interfaces for Training/ Recognition Phases

In order to incorporate the adaptation/learning functions efficiently in Golden Mandarin (II) and make the machine more friendlier to the users. special user interface for training and recognition phases are developed. Fig.4 shows the user interface for training phase. In Fig.4, the first training sentence among the 24 phonetically balanced sentences for the first stage speaker adaptation mentioned in section III, 「金聲二號智慧型國語聽寫機」 (Golden Mandarin (II) intelligent dictation machine), is shown on the top of the screen, and the user is requested to read the first character. The input speech waveform as well as the top 15 syllable candidates are also shown on the screen, so the user can easily assess the performance of the machine with respect to the syllable currently uttered. On the bottom of the screen the percentage of training utterances having been collected is shown, and it is also indicated that the first stage requires 18% of the total training utterances needed for the four stages. On the right hand side of the screen. Various functions useful for the training phase are

available, such as "continue the training procedure", "repeat this utterance", "delete this utterance", "repeat this sentence", "take the environmental noise", etc., each of which can be easily used by a mouse.

Fig.5 shows the user interface for recognition phase. The upper part of the screen is for editing purposes, while the lower part displays the recognized syllables and characters one by one in real time. The first sentence, 「這台電腦聽得懂國語」 (This computer can listen to Mandarin), has been correctly entered and moved to the upper part for further editing, while the next sentence, 「從此中文輸入不用鍵盤」 (The keyboard will not be needed in entering Chinese characters from now on), is currently being entered. In this sentence the second character (and the second syllable) is incorrectly recognized. So the "correction" function can be used by the mouse to open a window to see the top 15 syllable candidates to select. Unfortunately the correct syllable is not among the top 15 candidates, so a second window is opened on which all the Mandarin phonetic symbols are listed and the user can easily enter the correct syllable using the mouse without touching the keyboard. Other functions such as "learning", "deleting", "continuing" are also available at the bottom of the screen. The "learning" function is for the adaptation /learning function of the Chinese language model with respect to task domains and word patterns as mentioned above.

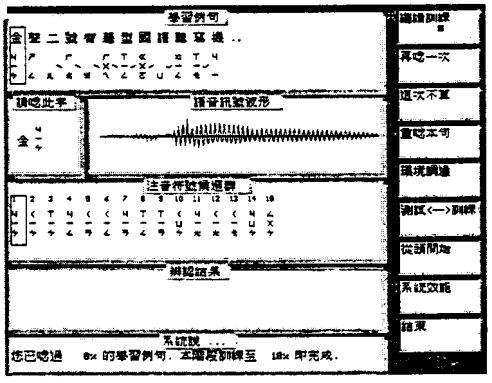

Figure 4: The user interface for training phase

VI. Conclusion

An intelligent Mandarin dictation machine, Golden Mandarin (II), has been successfully developed, Not only a single DSP chip is enough to provide close to real-time speed with very high accuracy, but the various adaptation /learning functions make the machine very friendly for users. Such adaptation /learning functions are believed to be very important if practically feasible products to be used by large number of users are considered.

ACKNOWLEDGMENT

The work presented in this paper has been supported by the National Science Council of Republic of China from 1984 to 1994, and performed in the Speech Laboratory of National Taiwan University in Taipei. The contributions made by many graduate/ undergraduate students working in this laboratory in the last 10 years are highly appreciated, although their work may not be directly included in the Golden Mandarin (II) system thus their names may not listed as a co-author.

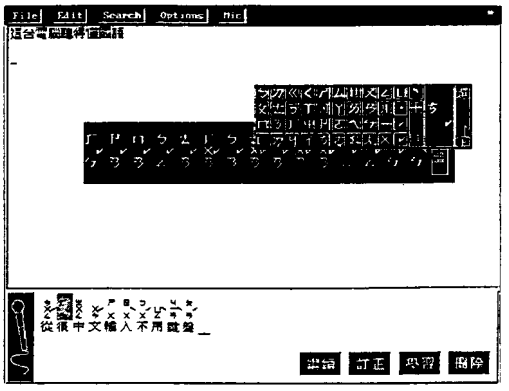

Figure 5: The user interface for recognition phase

References

[1] L.S. Lee, C.Y. Tseng, *et. al.* "A Real-time Mandarin Dictation Machine for Chinese Language with Unlimited Texts and Very Large Vocabulary," *ICASSP*, Apr 1990, Albuquerque, NM, USA, pp.65-68.

[2] L.S. Lee, C.Y. Tseng, *et. al.* "Golden mandarin (I) – A Real-time Mandarin Speech Dictastion Machine for Chinese Language with Very Large Vocabulary," *IEEE Transactions on Speech and Audio Processing*, Vol.1, No.2, Apr 1993, pp.158-179.

[3] L.S. Lee, C.Y. Tseng, *et. al.* "Golden Mandarin(II) - An Improved Single-Chip Real-Time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary," *ICASSP*, Apr 1993, pp.503-506.

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

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

[6] H.Y. Gu, C.Y. Tseng and L.S. Lee, "Markov Modeling of Mandarin Chinese for Decoding the Phonetic Sequence into Chinese Characters," *Computer Speech and Language*, Vol.5, No.4, Oct 1991, pp.363-377.

[7] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model component of a Speech Recognizer," *IEEE Transactions on ASSP*, Vol.35, pp.400-411, 1987.