CCLMDS'96: Towards a Speaker-Independent Large-Vocabulary Mandarin Dictation System

TungHui Chiang, Chung-Mou Pengwu, Shih-Chieh Chien, Chao-Huang Chang

Advanced Technology Center (ATC)
Computer & Communication Research Laboratories (CCL)
Industrial Technology Research Institute (ITRI)
Chutung, HsinChu, Taiwan 31015, R.O.C.

ABSTRACT

This paper presents the first known results for the speaker-independent large-vocabulary **M**andarin **D**ictation **S**ystem, namely **CCLMDS'96**, developed by Computer & Communication Research Laboratories (**CCL**) at Industrial Technology Research Institute (**ITRI**). First, a fast searching algorithm is proposed to improve the searching efficiency such that the CCLMDS'96 can operate in real time running on a personal computer. In addition, a discriminative scoring function is proposed to integrate the speech recognizer and the word-class-based bigram language model. With this discriminative scoring function, the system attains word accuracy rate of 91.3%, which significantly outperforms the conventional integration approach.

1. INTRODUCTION

CCL/ITRI has been devoted efforts towards building speaker-independent medium-vocabulary (100~1,000 words) Mandarin speech recognition systems for voice-command/control applications, such as voice dialing, Windows navigation, etc. A Microsoft speech API compliant Mandarin speech recognition engine has been accomplished allowing users to develop their own applications with Chinese voice input capability. Now, the CCLMDS'96 system is our first attempt on building a Mandarin dictation system based on our previous experiences. At present, like some Mandarin dictation systems [1, 2], the prototyping CCLMDS'96 is designed with isolated-word inputs. However, unlike the Mandarin dictation systems [1, 2, 3, 4] that were designed only either for speaker-dependent or for speaker-adaptive operations, the CCLMDS'96 is specially developed aiming at recognizing *speaker-independent* inputs.

Furthermore, another emphasis for the CCLMDS'96 system is its ability to operate in real time on a personal computer. To do this, a fast two-stage searching algorithm is proposed for speeding up searching process. At the first stage, a screening procedure is carried out on the entire vocabulary to locate the possible candidates with a coarse HMM-based classifier. The computation time spent in this stage is almost independent of the vocabulary size. During the second stage, only the words in the substantially reduced subset are considered by a finer HMM recognizer. The computation time spent in this stage is constant once the number of candidates selected in the first stage is fixed. The experiment results show that this two-stage approach speeds up the recognition process efficiently, only with a negligible sacrifice of the recognition rate.

Even though the speech recognizer would recognize the input uttered syllables with a relatively high accuracy rate, the system is still unable to decide what are the words being uttered because there are a lot of homophones in Mandarin. The homophone problem is quite serious for Mandarin because there are over 10,000 ideographical characters sharing the 1,229 syllables in the language. To decode the syllables into corresponding words, the language module with a word-class-based bigram language model [6, 7] is applied in the present system. In particular, the word classes used in this system are automatically trained by using the simulated annealing algorithm on the 1991 United Daily (UD) newspaper corpus [6, 7]. This corpus is composed of 579,123 sentences (or 4,761,120 words) extracted from 19 days of newspapers. Since the corpus covers very wide ranges of topics, the system is thus robust to applications of general domains.

Regarding the integration of the speech and language modules, the conventional approaches usually assume that the speech and the language modules contribute equally to the overall discrimination power of the system. However, at most time this assumption is inappropriate because uncertainty and ambiguity exist in different extents at speech and language processing modules. In addition, the likelihood values provided by the speech and language modules usually have different dynamic ranges, directly combining those likelihood values would thus over-emphasize the term that has largest dynamic range and de-emphasize the others. Motivated by the above concerns, a discrimination function is proposed to achieve a better integration framework. The joint learning procedure proposed by [8,9] is used to adjust the parameters. With such a learning procedure, not only the discrimination power of the system is ensured, the dynamic range variations of the various likelihood values are also well compensated. In the test on a lexicon of 1,000 words, the accuracy of the system with the discriminative approach achieves 91.3% word accuracy rate, which is significantly better than that attained by the conventional approach.

2. SYSTEM DESCRIPTION

For a dictation system, the preferred output word sequence $\hat{w}_1^N \left(= \hat{w}_1\hat{w}_2\ldots\hat{w}_N\right)$ with respect to the input acoustic feature vectors $\mathbf{a}_1\mathbf{a}_2\ldots\mathbf{a}_N$ is usually expressed as the following formulae:

$$\hat{w}_1^N = \arg\max_{w_1^N} P\left(w_1^N \mid \mathbf{a}_1^N\right)$$
$$= \arg\max_{w_1^N} \left\{\log P\left(w_1^N\right) + \log P\left(\mathbf{a}_1^N \mid w_1^N\right)\right\}, \qquad (1)$$

where $\log P\left(\mathbf{a}_1^N \mid w_1^N\right)$ and $\log P\left(w_1^N\right)$ represent the speech score and the language score, respectively. We are going into the details about these scores in the following section.

2.1 Speech Processing

Sub-Syllable Modeling:

In our speech recognizer, the acoustic feature vectors are extracted from the 8KHz sampled data every 10ms. Each feature vector is composed of the 12-order mel-scaled cepstral coefficients and the corresponding delta cepstral coefficients. To account for intra-syllable coarticulation, a total of 143 right-context-dependent (RCD) phone-like units (PLU), as defined in [4], are used as the acoustic units in the speech recognizer. Each of the RCD-PLU's is modeled as a 3-state continuous Hidden Markov Model (CHMM) with each state being characterized by four Gaussian mixtures.

The training of the RCD-PLU's is carried out by using two speech databases:

[1] utterances from 90 speakers (50 male and 40 female), each speaking 408 base syllables.

[2] utterance from 16 speakers (5 male and 11 female), each speaking 479 poly-syllable words.

Since the training databases contain utterances from more than 100 speakers, the HMM's trained with these databases provide certain degree of robustness for speaker-independent uses.

In the present system the tones of syllables are not considered yet. Therefore, the words of the same base-syllables, regardless of tones, are considered as the homonyms in the current implementation. This works because there are much fewer poly-syllabic homonyms than mono-syllabic ones in Mandarin.

In addition, we have pre-complied a tree lexicon according to the pronunciation lexicon. The recognition of the input utterances can be thus viewed as the traversal in the tree lexicon. However, in large-vocabulary cases the conventional tree searching algorithm that traverses the tree lexicon exhaustively cannot meet our developing requirement of "real-time response on low-end computers." To improve searching efficiency, a fast searching algorithm is employed in our system. This searching algorithm is described as follows.

The Fast Searching Algorithm:

Let the vocabulary is described as $\mathbf{V} = \{ W_1, W_2, \dots W_{|\mathbf{V}|} \}$, $|\mathbf{V}|$ denotes the number of words in the vocabulary, and the length of word W_i (i.e. the number of characters constituting the word W_i) as l_i. Each word W_i is described by the model M_i, which is formed by concatenating the corresponding character models C_i's, i.e., $M_i = C_{i1}C_{i2}\dots C_{il_i}$. Since each character contains only one syllable in Mandarin, each character model C_i can be further described as a concatenation of an initial (consonant) model m_j^I and a final (vowel or diphthong, possible with nasal ending) model m_k^F, i.e. $C_i = [m_j^I] m_k^F$. The bracketed initial model indicates that it may be absent for some syllables.

For the conventional tree search procedure, the computation time required for an input speech signal can be approximately expressed as $\{a + b \cdot |\mathbf{V}|\}$, where a is the time to compute the output probabilities and b is a positive constant for a particular input speech signal. The proposed searching algorithm is a two-stage approach that tries to reduce $|\mathbf{V}|$ to a smaller number K in the first stage. The standard tree search procedure is then performed over this small set of vocabulary in the second stage.

In the first stage, the input speech signal **A** is segmented into l syllable segments each of which contains an initial sub-segment and a final sub-segment. In this procedure, the segmentation is carried out using the Viterbi algorithm based on the generic initial and final models shown in Figure 1. Note that silence in the first stage is represented by the generic initial model. Therefore, the input speech signal in this stage can be summarized by the following equation:

$$\mathbf{A} = ([s_1^I] s_1^F)([s_2^I] s_2^F)\dots([s_{lk}^I] s_{lk}^F) \qquad (2)$$

wherein s_i^I ($i = 1, 2, \dots, l$) is the i-th initial sub-segment and s_i^F ($i = 1, 2, \dots, l$) is the i-th final sub-segment. The brackets enclosing the initial syllable sub-segments indicating that the initial sub-segment may be missing.

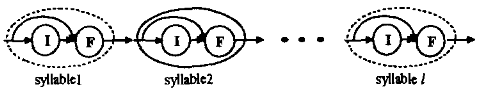

Figure 1 Concatenation of the generic initial and final models.

The next step is to compute $\log P(s_i^I \mid m_j^I)$ and $\log P(s_i^F \mid m_k^F)$ for all the sub-segments and for all the initial models m_j^I and final models m_k^F, respectively. Since each word W_n in **W** with length l is represented by an acoustic model $M_n = ([m_{j1}^I] m_{k1}^F)([m_{j2}^I] m_{k2}^F)\dots([m_{jl}^I] m_{kl}^F)$, a preliminary logarithmic probability $\log P(W_n)$ for acoustic model M_n is calculated according to the following formula:

$$\log P(W_n) = \sum_{r=1}^{l} (\log P(s_r^I \mid m_{nr}^I) + \log P(s_r^F \mid m_{nr}^F)). \qquad (3)$$

However, it is possible for Mandarin language that the input speech signal **A** may be segmented into other l's. If this is the case, then a different l is selected and the above preliminary screening procedure is repeated until all possible l's (usually l =2, 3, or 4 only) are exhausted. After all the $\log P(W_n)$'s are calculated, the words within the top Kth $\log P(W_n)$'s are selected . The K words with these largest values of $\log P(W_n)$ are then subjected to the second stage process. In the second stage, the conventional recognition procedure is applied to these K words. The search procedure is then performed in this K-words vocabulary rather than the original $|\mathbf{V}|$ words. If the classifier defined in the first stage is relatively effective, the K value can be selected to be much smaller than $|\mathbf{V}|$. In such a way, the computation time in this stage can be significantly reduced.

2.2 Linguistic Decoding

To decode the syllables into corresponding words, the language score $\log P(w_1^N)$ is modeled as the word-class bigram:

$$\log P(w_1^N) \approx \sum_{i=1}^{N} \{\log P(C_i|C_{i-1}) + \log P(w_i|C_i)\}, \qquad (4)$$

where C_i $(1 \le i \le N)$ denotes the word class of w_i. In this paper, the word classes are defined in statistical sense and are trained automatically.

Automatic Clustering of Words:

The most important issue for developing an efficient class-based language lies in the classification of words. While parts-of-speech, based on syntactic categories, are considered valuable word classes, they are not adopted in the present system because no manually tagged large corpus, counterpart of the Brown corpus or LOB corpus in Chinese, is available. Instead, the word classes adopted in the present system are automatically trained from a very large corpus via a simulated annealing procedure [6, 7]. The corpus used is the 1991 United Daily (UD) newspaper corpus [6, 7], which comprises 579,123 sentences (or 4,761,120 words) extracted from 19 days of newspapers. Since the corpus has wide coverage in different topics, the class-based bigram model trained with these materials have been shown to be quite robust [6,7]. At the end of the clustering procedure, a total of 500 word classes are used in our system. Readers who are interested in the clustering procedure are referred to [6, 7] for details.

A Discriminative Approach for Integration:

In general, the decoding of the output word sequence for a conventional dictation system is expressed as the following formula:

$$\arg\max_{w_1^N} \sum_{i=1}^{N} \left\{\log P\left(C_i|C_{i-n+1}^{i-1}\right) + \log P(w_i|C_i) + \log P(\mathbf{a}_i|w_i)\right\}. \qquad (5)$$

However, equation (5) assumes that the speech and the language scores contribute equally to the overall discrimination power of the system. This assumption is inappropriate for most systems because uncertainty and ambiguity exist in different extents at speech and language processing levels. In addition, the log likelihood values $\log P(C_i|C_{i-1})$, $\log P(w_i|C_i)$ and $\log P(\mathbf{a}_i|w_i)$ usually have different dynamic ranges, the scoring function defined as in equation (5) would over-emphasize the term that has largest dynamic range and de-emphasize the others.

Therefore, to resolve the above-mentioned problems, the discrimination function of the dictation system must take into account not only the discrimination power but also the dynamic ranges of all likelihood values. To do that, the discrimination function for the word sequence $w_1 w_2 \ldots w_N$ with respect to $\mathbf{a}_1 \mathbf{a}_2 \ldots \mathbf{a}_N$ is defined as the following equation in the present system:

$$g\left(w_1^N|\mathbf{a}_1^N\right) = \sum_{i=1}^{N} \left\{ \begin{array}{l} w_c \cdot \log P\left(C_i|C_{i-n+1}^{i-1}\right) \\ +w_w \cdot \log P(w_i|C_i) \\ +w_a \cdot \log P(\mathbf{a}_i|w_i) \end{array} \right\}, \qquad (6)$$

where w_c, w_w, w_a are the weights reflecting the discrimination capabilities of the values $\log P\left(C_i|C_{i-n+1}^{i-1}\right)$, $\log P(w_i|C_i)$ and $\log P(\mathbf{a}_i|w_i)$, respectively. Thus, the criterion for finding the best word sequence is expressed as follows:

$$\hat{w}_1^N = \arg\max_{w_1^N} g\left(w_1^N|\mathbf{a}_1^N\right). \qquad (7)$$

At present, the values of the weights are determined automatically by using the joint learning algorithm proposed by Chiang et al. [8, 9]. Since the contributions of the speech and the language modules to the dictation system are jointly considered in the learning procedure, the parameters of these two modules and their corresponding weights can be thus adjusted simultaneously [9]. The advantage of the joint leaning procedure is to consider both the speech and the language modules at the same time to avoid overtuning either of these two modules. With this learning procedure, not only the discrimination power of the system can be ensured, the dynamic range variations of the different likelihood values can be also well compensated.

3. EVALUATIONS

Evaluation on the Fast Searching Algorithm

First, experiments are performed to evaluate the proposed fast searching algorithm. In the first experiment, utterances of the 1,000 words in the vocabulary recorded by one male speaker is used for test. The recognition performance versus different number of candidates retained in the first stage, i.e., K value, is shown in Table 1. It is found that the accuracy rate with the fast searching algorithm is almost the same as the conventional tree search scheme if 30 candidates are retained in the first stage. Therefore, the experiments conducted in the rest of this paper are all referred to the condition of K=30.

	Error Rate (%)
Tree Search	8.7

Fast Search	**Error Rate (%)**
K=10	9.3
K=20	9.0
K=30	8.9

Table 1 The performance of the two searching algorithms

In addition, the processing time, running on SUN SPARC 10 workstation, for the two search methods under different vocabulary sizes is plotted in Figure 2. The superiority of the proposed fast search algorithm is clearly demonstrated since the CPU time required for this approach is almost independent of the size of the vocabulary and is shorted than that required for the tree search approach. The advantage for adopting the fast search algorithm is particularly significant in a large vocabulary case.

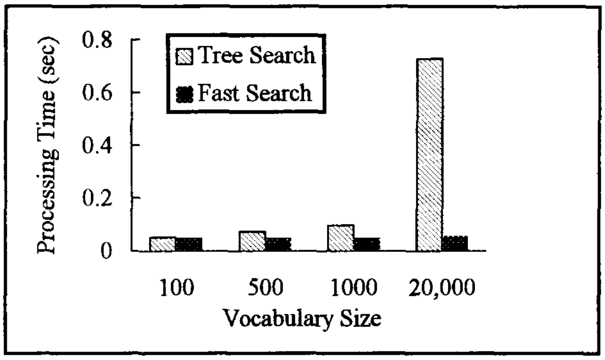

Figure 2 The processing time versus vocabulary size

Evaluation on the Discriminative Scoring Function

To evaluate the effect of the discriminative scoring function, utterances of 250 sentences (1,200 words or 2,736 characters) from newspapers read by 5 male speakers are tested. When tested on a lexicon of 1,000 words, the speech recognizer, as shown in Table 2, has syllable accuracy rate of 82.1% and the coverage rate of 98.9% and 99.7% for top 10 and top 20 candidates, respectively. Therefore, for each input utterance of word, 20 best candidates generated by the speech recognizer are used for lexicon access and word-lattice construction in the current implementation.

TopN	Syllable Accuracy Rate (%)
N=1	82.1
N=5	96.8
N=10	98.9
N=20	99.7

Table 2 The performance of the speech recognizer

Finally, the conventional integration approach, i.e., Eq.(5), and the proposed discrimination-based integration approach are evaluated; the results are shown in Table 3. Compared to the result shown in Table 2, it is found that the syllable accuracy rates are improved both for the conventional approach and the discriminative approach. In addition, the word accuracy rate for the discriminative approach is 5% better than the conventional approach, which corresponds to 36.5% error reduction. This sound result has shown the superiority of the proposed discriminative scoring function for speech and language integration.

	Conventional Approach	Discriminative Approach
syllable accuracy	87.7	92.0
word accuracy	86.3	91.3

Table 3 The performance of different integration approaches

4. SUMMARY

This paper presents the first known results for the speaker-independent large-vocabulary Mandarin dictation system, **CCLMDS'96**, developed by CCL/ITRI. In this paper, a fast searching algorithm is proposed to speed up the searching process. With this searching algorithm, the CCLMDS'96 can operate in real time running on a low-end computer. In addition, a discriminative scoring function is proposed to integrate the speech recognizer and the word-class-based bigram language model in a more effective manner. With the discriminative scoring function, the system achieves word accuracy rate of 91.3%, significantly outperforming the conventional approach.

To make our dictation system practical to use, our future work on improving the prototype system includes incorporating a tone recognizer; enhancing sub-syllable modeling for base-syllable recognition, and increasing the vocabulary size, e.g., 20,000 or 40,000 words.

ACKNOWLEDGEMENT

This paper is a partial result of the project no. 37H2100 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C.

REFERENCES

[1] H. W. Hon and K. F. Lee et al., "Towards Large Vocabulary Mandarin Chinese Speech Recognition," Proceedings of ICASSP'94, pp. 545-548, 1994.

[2] Y. Gao, H. W. Hon, Z. Lin, G. Loudon, "TANGERINE: A Large Vocabulary Mandarin Dictionary System," Proceedings of ICASSP'95, pp. 77-80, 1995.

[3] L. S. Lee et al., "Golden Mandarin (II) – An Improved Single-Chip Real-Time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary," Proceedings of ICASSP'93, pp. 503-506, 1993.

[4] R. Y. Lyu and L. S. Lee et al., "Golden Mandarin (III) – A User-Adaptive Prosodic-Segment-Based Dictation Machine for Chinese Language with Very Large Vocabulary," Proceedings of ICASSP'95, pp. 57-60, 1995.

[5] C. M. Pengwu, "A Fast Recognition Method for Isolated Words in CV Structure Language," to appear in Proceedings of the 1996 International Symposium Multi-Technology Information Processing, ISMIP'96, 1996.

[6] C. H. Chang, "Word Class Discovery for Postprocessing Chinese Handwriting Recognition," Proceedings of 1994 International Conference on Computational Linguistics, pp. 1221-1225, 1994.

[7] C. H. Chang, "Simulated Annealing Clustering of Chinese Words for Contextual Text Recognition," Pattern Recognition Letters, Vol 17, pp. 57-66, 1996.

[8] T. H. Chiang, Y. C. Lin, and K. Y Su, "A Study of Applying Adaptive Learning to a Multi-module System" Proceedings of 1994 ICSLP'94, pp. 463-466, Yokohama, Japan, 1994.

[9] T. H. Chiang, Y. C Lin and K. Y. Su, "On Jointly Learning the Parameters in a Character-Synchronous Integrated Speech and Language Model, " IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 3, pp. 167-189, 1996.